Adaptive Face Forgery Detection in Cross Domain (Supplementary Material)

Luchuan Song[1], Zheng Fang[2⋆], Xiaodan Li[3], Xiaoyi Dong[1], Zhenchao Jin[1], Yuefeng Chen[3], and Siwei Lyu[4]

[1] University of Science and Technology of China
{slc0826, dlight blwx96}@mail.ustc.edu.cn
[2] Shopee Inc.
fangzheng0827@gmail.com
[3] Alibaba Group
{fiona.lxd, yuefeng.chenyf}@alibaba-inc.com
[4] University at Buffalo
siweilyu@buffalo.edu

Abstract. This document supplementary material of our paper **Adaptive Face Forgery Detection in Cross Domain** by discussions on the technical fronts as well as computational performance analysis, additional experiments and visual performance.

1 Feature Extraction Flow

We use the algorithm chart to visually display the DICM feature extraction process in the main paper Fig.2. All symbols remain the same as those in Fig.2 in the main paper. The algorithm shown in Alg.1.

2 Hyper-Parameters on DICM

Throughout the main paper, we conducted the experiments on DICM with the fixed hyper-parameters where the number of frames in the feature sequence (n) is 3, the stride of selected frames (s) is 50 and the threshold of high-frequency mask filter for extracting high-frequency spectrum (t) is 0.25. Bigger n leads to the communal feature more generalized to the various frames while incorporating more parameters to extract attention for each frame. s affects the diversity of frames where smaller s makes the sampled frames seem similar to each other while bigger s improves the diversity of sampled frames. t controls the information in the high-frequency spectrum where bigger t leads to the masked spectrum consists of less information and focuses on high-frequency component, and vice versa. Here, we evaluate the proposed DICM on different hyper-parameters for face forgery detection by altering one type hyper-parameter, as shown in Tab. 2. There is a little performance difference on DICM with various hyper-parameters

⋆ Corresponding author.

Algorithm 1 DICM Process Procedure

```
Input:
  High-frequency: $\{\mathbf{F}_1^{\mathrm{H}}, \mathbf{F}_2^{\mathrm{H}}, \cdots, \mathbf{F}_n^{\mathrm{H}}\}$
  RGB: $\{\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_n\}$
Program:
  $\mathbf{S}^{\mathrm{H}} = \sum_i^n f_{\downarrow}^{\mathrm{H}}(\mathbf{F}_i^{\mathrm{H}})$;
  $\bar{\mathbf{A}}_i^{\mathrm{H}} = \mathrm{SoftAttention}(\mathbf{S}^{\mathrm{H}})[\ i\ ]$;
  $\bar{\mathbf{F}}_i^{\mathrm{H}} = f_{\uparrow}^{\mathrm{H}}(\mathbf{S}^{\mathrm{H}} \otimes \bar{\mathbf{A}}_i^{\mathrm{H}}) \oplus \mathbf{F}_i^{\mathrm{H}}$;
  $\mathbf{S} = \mathrm{Merge}(\sum_i^n f_{\downarrow}(\mathbf{F}_i), \mathbf{S}^{\mathrm{H}})$;
  $\bar{\mathbf{A}}_i = \mathrm{SoftAttention}(\mathbf{S})[\ i\ ]$;
  $\bar{\mathbf{F}}_i = f_{\uparrow}(\mathbf{S} \otimes \bar{\mathbf{A}}_i) \oplus \mathbf{F}_i$;
Output:
  Refined High-frequency:$\{\bar{\mathbf{F}}_1^{\mathrm{H}}, \bar{\mathbf{F}}_2^{\mathrm{H}}, \cdots, \bar{\mathbf{F}}_n^{\mathrm{H}}\}$
  Refined RGB: $\{\bar{\mathbf{F}}_1, \bar{\mathbf{F}}_2, \cdots, \bar{\mathbf{F}}_n\}$
```

Table 1. AUC of different multi-frame merging strategy on FF++. `Integration` indicates the group of frames share the same score. `Individual` indicates each frame in the group gets individual prediction.

Methods	`Integration`	`Individual`
Xception + DICM	0.944	0.944
CD-Net (Xception)	0.952	0.952

while this performance gap is limited and our DICM stably better than baseline (model 1). The DICM extracts communal representations of forgery patterns from multiple frames with less dependence on the number of frames or the stride of sampling on frames. High-frequency spectrum extracted with various threshold of high-frequency mask filter can be adaptively learned in DICM, helping the DICM less sensitive to hyper-parameters.

3 Details of Merging on Multiple Frames

In the CD-Net, information from multi-frames (3 frames in our experiments) are utilized and the features from each frame are merged together at the final convolution layer to get the prediction. This prediction can be either the score of all used frames, indicated as "`Integration`", or only the score of the middle frame, indicated as "`Individual`". The former strategy is more efficient and the running time is nearly the same with single-frame input approaches, which is adopted in our experiments for efficiency. The AUC score of each strategy on FF++ is listed in Tab. 1 and there is no performance difference on the AUC score, further demonstrating that our CD-Net is capable of learning communal feature from multiple frames.

Table 2. Experiments of DICM on FF++ c40 ((low quality)) to evaluate the effects of hyper-parameters. “n” means the number of frames in the feature sequence, “s” means the stride of frames and “t” means threshold of high-frequency mask filter.

ID	Methods	AUC
1	Xception	0.925
2	Xception + DICM (n=2)	0.940
3	Xception + DICM (n=3)	**0.944**
4	Xception + DICM (n=4)	0.944
5	Xception + DICM (s=40)	0.942
6	Xception + DICM (s=50)	**0.944**
7	Xception + DICM (s=60)	0.944
8	Xception + DICM (t=0.1)	0.940
9	Xception + DICM (t=0.25)	**0.944**
10	Xception + DICM (t=0.5)	0.941

4 Details on Dataset and Setting

We conduct experiments on widely-used datasets, *i*.e.FaceForensics++ (FF++) [5], DeeperForensics [2] and Deepfake Detection Challenge (DFDC) Dataset [1]. We follow previous settings used in their corresponding datasets and compare with other methods respectively. More details on these datasets are described below.

FaceForensics++ (FF++) is a face forgery detection video dataset containing 1,000 real videos, in which 720 videos are used for training, 140 videos are reserved for validation and 140 videos for testing. Most videos contain frontal faces without occlusions and were collected from Youtube with the consent of the subjects. Each video undergoes four manipulation methods to generate four fake videos, therefore there are 5,000 videos in total. When training and evaluating on FF++, we follow the sampling strategy mentioned in [5] that samples 270 frames per video for the training and 100 frames per video for validation and testing. We evaluated all no compression (raw), medium compression (c23) and high compression levels (c40) subsets.

DeeperForensics is a newly proposed face forgery dataset containing 1,000 real videos the same with FF++ c23 real videos and 1,000 fake videos generated using the Variational Auto-Encoder proposed in [2]. The training, validation and testing are separated different from FF++ with 703 videos for training, 96 videos for validation and 201 videos for testing. DeeperForensics performs different level distortion perturbations on data and level-5 is the hardest level for detection. Following the setting described in [2], we use the hardest setting that training on raw data without distortion perturbations and testing on both level-5 and random-level data to validate the generalization of our method to distortion perturbations.

DFDC Dataset is a preview dataset consisting 5,221 valid videos of digitally manipulated and real videos. This dataset are not separated train/test sets

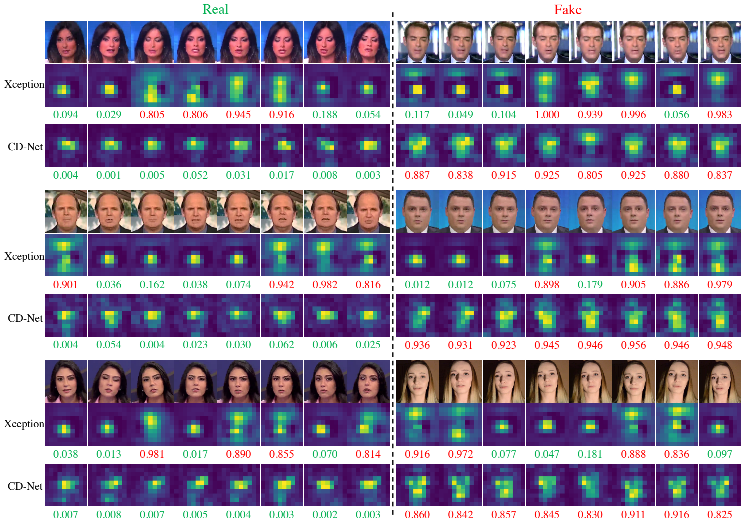

Fig. 1. The visualization of feature map extracted by Xception and the CD-Net respectively.

officially. We evaluate how well our model transfers to DFDC Preview Dataset given that it is trained on FF++ c40. The experiment is conducted following the setting described in [1]. Worth noting that, previous methods [1, 4] split the dataset to train/val for choosing better thresholds to determine the video-level prediction. To prove the robustness of our method, we simply use the average score of frames as the final video score, and the threshold to determine real/fake is 0.5 without any threshold searching.

Celeb-DF v2 contains 5, 639 fake videos and 590 real videos. Following the previous setting in [3, 4], we use the Celeb-DF v2 dataset to evaluate the generalization performance of our model on unseen data. We use the model trained on FF++ c40, FF++ c23 and FF++ raw to evaluated on Celeb-DF v2 test set with 518 videos.

5 Visible Results

More visible results of feature maps extracted by Xception (baseline) and our Xception-based CD-Net are shown in Fig.1. As presented in Fig.1(1) left, the predictions from Xception varies a lot with the head pose changes while our CD-Net achieves consistent representation on multiple frames from the same video. In the Fig.1(3) right, affected by the unrobust representations on forgery patterns in the baseline, frames which look nearly the same as each other get totally different prediction in Xception. Our CD-Net is capable of extracting the communal patterns existed in multiple frames and achieve robust predictions.

References

1. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C.C.: The deepfake detection challenge (dfdc) preview dataset. arXiv preprint arXiv:1910.08854 (2019)
2. Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C.: Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2886–2895. IEEE (2020)
3. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3207–3216 (2020)
4. Masi, I., Killekar, A., Mascarenhas, R.M., Gurudatt, S.P., AbdAlmageed, W.: Two-branch recurrent network for isolating deepfakes in videos. arXiv preprint arXiv:2008.03412 (2020)
5. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1–11 (2019)