Supplementary Material for:
AltFreezing for More General Video Face Forgery Detection

Zhendong Wang [1,*] Jianmin Bao [2,*] Wengang Zhou [1,3,†] Weilun Wang [1] Houqiang Li [1,3,†]
[1] CAS Key Laboratory of GIPAS, EEIS Department, University of Science and Technology of China
[2] Microsoft Research Asia
[3] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{zhendongwang,wwlustc}@mail.ustc.edu.cn
jianbao@microsoft.com, {zhwg,lihq}@ustc.edu.cn

1. More Implementation Details

Face detection and align. We use RetinaFace [3] to detect and align faces for each video. We crop the region of faces within the same range of the detected face area, *i.e.*, four times the detected face area, where the weight and height are equal to twice the weight and height of the detected face, respectively. Then during training, for each video clip that contains 32 frames, we align them to a mean face. After all, the images in each clip are resized to 224 × 224.

Network architecture. The backbone we used is the bottleneck design of 3D ResNet50 (R50) [2], in which the 3 × 3 convolution in the basic block is replaced with a consecutive 3 × 1 × 1 and 1 × 3 × 3 convolution. Our implementation is based on Pytorch 1.8.0 with Cuda 11.0 on 2 GeForce RTX 3090 GPUs.

2. Additional Experiments

More ablation study of the freezing ratio of AltFreezing. In our main paper, we have explained that adjusting the freezing ratio of $I_s : I_t$ can encourage the network to pay more attention to spatial or temporal artifacts. And in default, we set the freezing ratio larger than 1. Here, we conduct more experiments to identify the effect of freezing ratio smaller than 1. The AUC results are reported in Tab. 1. From the comparisons, we observe that AltFreezing's performance initially increases and then decreases as the freezing ratio varies from 1:1 to 1:20. The model achieves the best average AUC when the freezing ratio is 1:5, improving 0.4% AUC compared to the baseline (without AltFreezing) on average. Yet the performance is much lower than that when the freezing ratio $I_s : I_t$ is larger than 1. This is consistent with previous temporal-based methods [4, 6] that claim detecting temporal artifacts is more general than detecting spatial ones.

Freezing	Train on FF++			
ratio ($I_s : I_t$)	FF++	CDF	FSh	Avg
baseline	99.3	81.8	99.2	93.4
1:1	99.6	**82.4**	99.2	93.7
1:5	99.7	82.2	**99.4**	**93.8**
1:20	**99.8**	80.5	99.2	93.2

Table 1. **Ablation study of the ratio of freezing temporal kernels more than spatial ones of AltFreezing.** Video-level AUC(%) is reported. "baseline" means a 3D R50 with end-to-end training.

Aug.	Train on FF++				
	FF++	CDF	DFD	FSh	Avg
none	99.7	86.4	97.6	99.3	95.8
ours (w/o CB)	99.7	84.5	**98.8**	**99.4**	95.6
ours	99.7	**89.5**	98.5	99.3	**96.7**

Table 2. **Ablation study of the fake clip generation.** Video-level AUC(%) is reported. "Aug." means augmentation. "CB" denotes the clip-level blending in our fake clip generation.

Ablation study of the fake clip generation. To learn better video-level representation, we have proposed a set of fake video synthetic methods including temporal-level and spatial-level augmentations. We further conduct experiments to verify the effect of the components in the fake clip generation. The AUC results of the augmentations are reported in Tab. 2. We observe that enabled with the temporal augmentations (ours (w/o CB)), the model gets performance improvement on DFD [1] and FSh [5]. On CDF it gets a performance drop. In our experiments, we use temporal augmentations in default since they might benefit the generalization ability to more challenging scenes. Moreover, clip-level blending which introduces more general clip-level spatial artifacts without any temporal artifacts further boosts the performance, averaging AUC from 95.6% → 96.7%.

3. Evaluation on Real-world Scenarios

We further evaluate the performance of our model on more challenging scenes. The real-world DeepFake videos

[*] Equal contribution.
[†] Corresponding authors.

we used are downloaded from the YouTube channel "Ctrl Shift Face2"[1], which are carefully crafted so that humans cannot discriminate between real and fake videos easily. We compare our method with 3D R50 (baseline) without our AltFreezing and FTCN [6], as shown in the Youtube Url[2]. Our method has a more accurate judgment of real or fake. The comparison indicates that our method is much more robust than others in real-world scenarios.

4. Limitations

We are aware that our method cannot handle any type of face forgery. When facing some fake videos generated by artists using Adobe Photoshop or other realistic image editing applications, our method may not be able to detect them. Besides, our method is not fully robust to all perturbations. For example, when applied to heavily compressed videos, the performance of our method drops like other works.

References

[1] Contributing data to deepfake detection research. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html. Accessed: 2021-11-13. 1

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 1

[3] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, pages 5203–5212, 2020. 1

[4] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *CVPR*, pages 5039–5049, 2021. 1

[5] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *CVPR*, pages 5074–5083, 2020. 1

[6] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *ICCV*, pages 15044–15054, 2021. 1, 2

[1] https://www.youtube.com/channel/UCKpH0CKltc73e4wh0_pgL3g

[2] https://www.youtube.com/watch?v=q0m8r380P-A