Supplementary Material for Exploring Temporal Coherence for More General Video Face Forgery Detection

Yinglin Zheng [1] Jianmin Bao [2], Dong Chen [2], Ming Zeng [1]*, Fang Wen [2]
[1] School of Informatics, Xiamen University
[2] Microsoft Research Asia
{zhengyinglin@stu., zengming@}xmu.edu.cn, {jianbao, doch, fangwen}@microsoft.com

This supplementary material provides details that could not be included in the paper submission due to space limitations: Sec.1 provides details of our implementation. Sec.2 shows additional experiments which further indicates the effectiveness of our method. Sec.3 details our visualization method and shows some examples.

1. More Implementation Details

We apply RetinaFace[3] to detect and align the faces for each video. For the video clip, we jointly align all the faces to a mean face and crop the face region with the same area in the source video. Each clip comprises 32 frames, and the face is resized to 224×224. We employ random flip and the Cutout augmentation[4] during training. For Cutout, we randomly set n square regions to zero($n \in [1, 3]$), the area of the square regions ranges from 20% to 80%, the cutout regions are shared among all the frames in each clip.

All the experiments are conducted on 4 Nvidia Tesla v100 32 GB GPUs and Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz. Our method and other 3D R50 variants are implemented based on the PyTorch v1.4.0, build upon the opensource SlowFast[5] codebase.

2. Additional Experiments

2.1. Performance when trained on other datasets

To further validate the generalization capability of our method when trained on other datasets, we trained our model on Faceshifter(FSh), DeeperForensics(DFo) and test on Faceshifter(FSh), DeeperForensics(DFo), FaceSwap(FS), and Deepfake(DF). As shown in Table 1, our methods still achieve very high performance when trained on other datasets.

*Corresponding author.

2.2. Comparison with More Temporal-based Approaches

We further compare our method to more temporal-based methods [1, 6, 8]. For PE-LSTM [1] and R3D [6], we follow the setup in these papers. For Two-branch[8], we follow the experiment setup in LipForensics[7]. Table 2 reports the comparison results, where our method achieves better results, indicating the effectiveness of our method.

3. Visualization

3.1. Examples of Spatial-shuffled Clip

We show examples of spatial-shuffled clips in Figure 1. As the spatial-shuffled operation destroyed most spatial relationships, with such limited spatial information, our network can only rely on temporal information. As shown in Table 5 of the original paper, our 3D R50-FTCN-Shuffle trained upon such data still achieves impressive performance, which further justifies the motivation of our FTCN that mostly relies on temporal information rather than spatial information.

3.2. Localization of Temporal Incoherence

We show the video version of Figure 5 of the main text in the supplementary video. Window areas with higher fake probability are more temporal incoherent. In our implementation, the sliding window size is 32×32, and the sliding stride is 16. As our method takes a video clip as input, the resulted heatmap is clip level and keep the same for all the frames in a clip.

As shown in the supplementary video, temporal incoherence exists widely on various face forgery datasets, but their spatial-related artifacts are quite different. Our model generalizes well on different datasets and can robustly localize temporal incoherence.

We also show qualitative results explored by long-range incoherence in Figure 2. We can observe that Temporal Transformer helps to capture long-range inconsistencies.

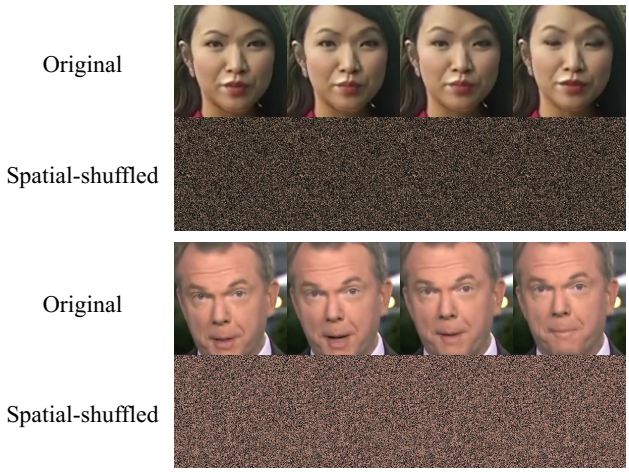

Figure 1. **Illustration of spatial-shuffled clip.** We show the first 4 frames of two clips. Spatial-shuffle pattern is shared among frames of the same clip.

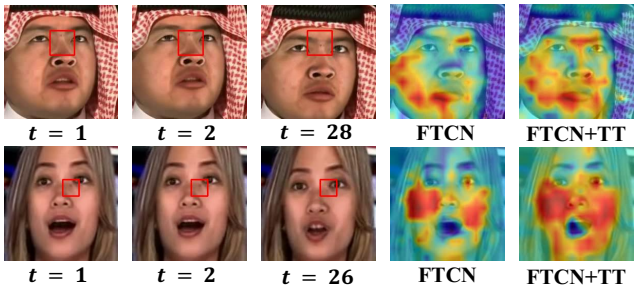

Figure 2. Visualization of long-range incoherence detection.

Train Set	Test set			
	FSh	DFo	FS	DF
FSh	**99.6**	95.4	95.1	99.1
Dfo	97.6	**100**	98.5	99.1

Table 1. Videl-level AUC(%) of our methods.

Methods	Train&Test Dataset	Accuracy	Videl-level AUC
PE-LSTM[1]	FF++(HQ)	85.3	—
Two-branch[8]		—	99.1
Ours		**99.1**	**99.8**
R3D[6]	FF++(HQ)&FSh	95.8	—
Ours		**99.7**	—
Two-branch[8]	FF++(LQ)	—	91.1
LipForensics[7]		—	98.1
Ours		—	**98.3**

Table 2. Comparison of our methods with existing works.

3.3. Frame-level Prediction

As our model takes video clips as input, we adopt the following strategy to obtain frame-level predictions. For an input video, RetinaFace is applied on every frame to detect faces. Then we do face tracking with SORT[2] to split the whole video into several segments that contain only one person. We applied a sliding window on temporal dimension to extract all video clips with 32 frames for every segment. We use reflective padding to enlarge the segment so that every frame is shared by 32 different clips. For a specific frame, its prediction is the average prediction of all clips that contain the frame. We present frame-level predictions of our method on challenging videos and the comparison with state-of-the-art methods in our supplementary video, our method outperforms state-of-the-art methods in terms of detection accuracy.

References

[1] Irene Amerini and Roberto Caldelli. Exploiting prediction error inconsistencies through lstm-based classifiers to detect deepfake videos. In *Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security*, pages 97–102, 2020.

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016.

[3] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[4] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019.

[6] Ipek Ganiyusufoglu, L Minh Ngô, Nedko Savov, Sezer Karaoglu, and Theo Gevers. Spatio-temporal features for generalized detection of deepfake videos. *arXiv preprint arXiv:2010.11844*, 2020.

[7] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. *arXiv preprint arXiv:2012.07657*, 2020.

[8] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. *arXiv preprint arXiv:2008.03412*, 2020.