DeepFake MNIST+: A DeepFake Facial Animation Dataset Supplementary Material

1. Data Generation And Collection Pipeline

The Figure 1 shows the pipeline to generate and collect our proposed DeepFake MNIST+ dataset. First, we collect the real videos from the VoxCeleb1 [5], extract the frames from these real videos as the source identity images. Then we shot driving videos with ten actions through the volunteers. In order to match the format of VoxCeleb1 videos, which are face-cropped and have the size of 256x256 resolution. We made a face-cropped version of driving videos with MTCNN modules [7] and also resize them into 256x256 resolution. We use Siarohin's framework [6] for animation video generation to produce face animation videos. It is a SOTA animation framework such that it even could capture the detail of eyeball moving. A single source image and a driving video were passed to the generator each time to produce single animation videos with a specific action. The generated videos were filtered with Liveness detector APIs to collect the challenging videos. Finally, 10,000 passed animation videos with ten actions (1000 videos for each action) and selected 10,000 real videos from VoxCeleb1 from our proposed DeepFake MNIST+ dataset.

2. More Examples of Data

2.1. Compressed Images

We demonstrate some raw and compressed (in both C23 and C40 compression rate under H.264 codec) video frames in Figure 2. The higher rate means heavier compression. As we can see, the c23 compression rate only leads to a minor impact on visual video quality. However, the frames suffer significant detail loss and blur effect under the c40 compression rate.

2.2. Driving Video samples

In this section, we present some driving video samples of different actions for animating the source images in Figure 3. The driving videos of embarrassment are picked from ADFES dataset [2].

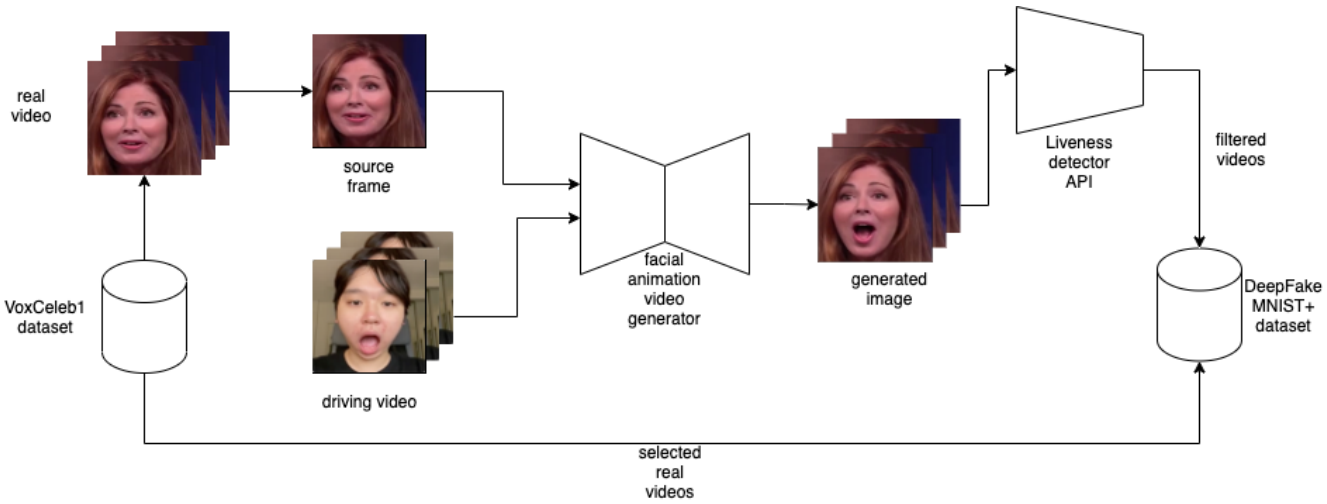

Figure 1: The pipeline to generate and collect our proposed DeepFake MNIST+ dataset.

3. More experiments

	DFDC[1]	DF-1.0[3]	Celeb-DF[4]
without finetune	61.2%	60.7%	57.2%
finetune	95.6%	96.1%	95.1%

Table 1: Accuracy of detecting DeepFake Mnist+ using the models trained with previous datasets. Fine-tuning means fine-tuned with proposed dataset.

Figure 2: The samples of raw and corresponding compressed (both c23 and c40) video frames.

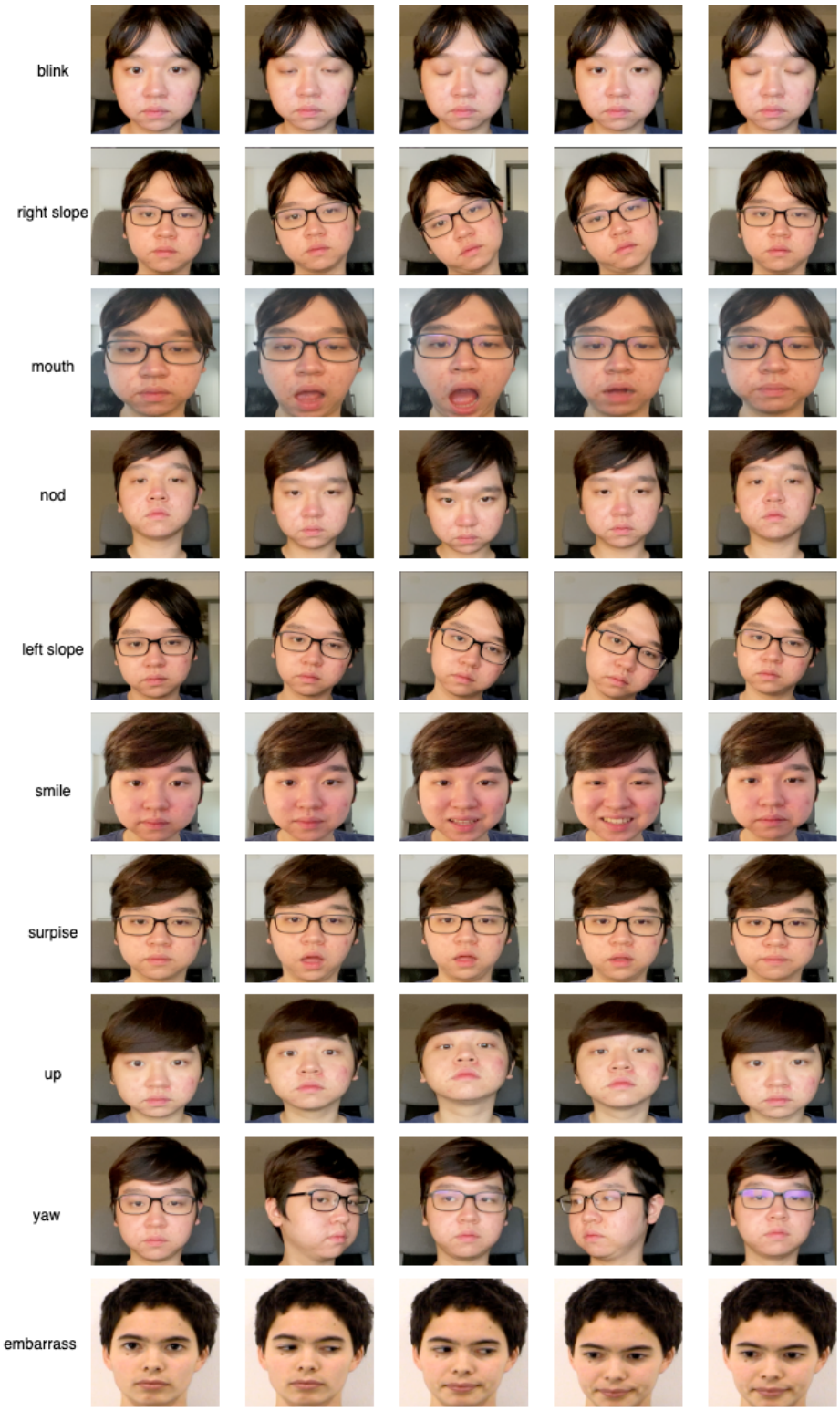

Figure 3: The driving video samples for animating the source images.

References

[1] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 1(2), 2020. 2

[2] ST Hawk, J Van der Schalk, and AH Fischer. Moving faces, looking places: The amsterdam dynamic facial expressions set (adfes). In *12th european conference on facial expressions, geneva, switzerland*, volume 4, 2008. 1

[3] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2889–2898, 2020. 2

[4] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020. 2

[5] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027, 2020. 1

[6] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *arXiv preprint arXiv:2003.00196*, 2020. 1

[7] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 1