Protecting Celebrities from DeepFake with Identity Consistency Transformer

Xiaoyi Dong[1], Jianmin Bao[2], Dongdong Chen[3], Ting Zhang[2], Weiming Zhang[1],
Nenghai Yu[1], Dong Chen[2], Fang Wen[2], Baining Guo[2]
[1]University of Science and Technology of China
[2]Microsoft Research Asia [3]Microsoft Cloud + AI
{dlight@mail., zhangwm@, ynh@}.ustc.edu.cn cddlyf@gmail.com
{jianbao, Ting.Zhang, doch, fangwen, bainguo }@microsoft.com

Experiment Details

More Implementation detail. The identity consistency Transformer is trained from scratch at a resolution of 112×112 without any data argumentation. Our Transformer consists of 12 blocks and the number of head in the multi-head self-attention layer in each block is 12. Follow the setting in [1], we set the weight decay of the ICT as 0.3. The input image is divided to 14×14 patches so that each patch is 8×8. We then use linear projection to project the patch token to embedding features of dimension 384. Meanwhile, we randomly initialize the inner token and the outer token with the same dimension 384. The number of training epochs is 30, and the batch size is 1024. The initial learning rate is set to 0.0005 and divided by 10 after 12, 15, 18 epochs. The loss balancing weight parameter η is set as 4 at the first epoch, and increase by 0.5 after each epoch. To further ease the training, we set the margin m as 0 at the first epoch and increase to 0.3 after 10 epochs.

Saliency map generation. In Figure 7. of our main paper, we visualize the saliency map of the inner and outer token to see which part of the face contributes most to learning the inner identity and also the outer identity as well. Here the saliency maps are obtained by traversing all the pixel positions in the image with a $k \times k$ mask and computing the difference between the original identity and the identity extracted from the masked image. The shown saliency maps are averaged from $k = 10, 15, 20$.

Reference set building. As we shown in our main paper, the reference set could boost the ICT performance with a large margin. Here we introduce how to build a reference set in the real-world usage. When handling unknown videos in real-world usage, we would build the reference set with a large number of faces for celebrities. This is an important advantage of our method that we could leverage online data to improve our performance. For the case that there is no reference, our ICT w/o reference is still a strong detector, as shown in Fig.6 right.

Model	Training Data	CD2	Deeper
Two-branch	Video	76.7	–
LipForensics	Video	82.4	97.6
ICT	Image	87.6	95.3
ICT-Ref	Image	95.6	99.5

Table 1. Video-level AUC comparison

More Experiments

Comparing with video-based methods. In our main paper, we report frame-level AUC. Here we further present a comparison with video-based methods. Following LipForensics [2], the ICT score is calculated by averaging all frames. Even without temporal clues, our ICT still performs comparably to SOTA video-based methods.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[2] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5039–5049, 2021. 1