ReenactGAN: Learning to Reenact Faces via Boundary Transfer

Supplementary Material

Wayne Wu[1*], Yunxuan Zhang[1*], Cheng Li[1], Chen Qian[1], Chen Change Loy[2]

[1] SenseTime Research
[2] Nanyang Technological University
{wuwenyan,zhangyunxuan,chengli,qianchen}@sensetime.com ccloy@ntu.edu.sg

Abstract. In this document, more details of the network architecture and training of encoder are provided. In addition, the full comparison with Face2Face [5] in facial action consistency is reported.

1 Architecture

Here we describe the architectures for encoder, decoder, and transformer.
Encoder: The encoder is adapted from a two-stage stacked hourglass network [4]. For model acceleration, we also try to use only the first-stage result which turned to have little influence of the qualitative results. The input is a 256×256 RGB face image while the output is a 15-channel 64×64 boundary heatmaps, where each channel corresponds to one of the 15 predefined facial contours. More details are included in the supplementary material.
Decoder: To generate high-resolution faces, we firstly upsample the boundary maps to 15-channel 256×256 with two progressive modules of the form ReLU-Deconvolution-BatchNorm. Then, a U-Net architecture with skip connections from [1] is adapted to generate the final face image of resolution 256 × 256.
Transformer: We adapt the architecture of our transformer and discriminator from those in [6]. Specifically, the transformer network consists of two stride-2 convolutions, nine residual blocks, and two fractionally-strided convolutions. The discriminative network is implemented with 70×70 PatchGAN [6, 2], which has the advantage of fewer parameters and being able to be applied to images of arbitrary size.

2 Training Details of Encoder

To generate the ground truth of boundary heatmap map for supervision, we predefine 15 facial boundaries, *i.e.*, facial outer contour, upper side of left eyebrow, lower side of left eyebrow, upper side of right eyebrow, lower side of right eyebrow, nose bridge, nose boundary, left upper eyelid, left lower eyelid, right

*Equal contribution. This work was done during an internship at SenseTime Research.

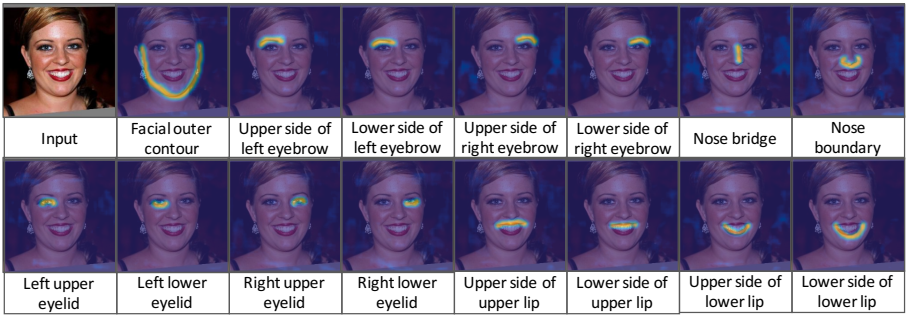


Fig. 1. The definition of boundary heatmaps.

upper eyelid, right lower eyelid, upper side of upper lip, lower side of upper lip, upper side of lower lip and lower side of lower lip. We demonstrate the heatmap of these 15 facial boundaries respectively in Figure. 1.

For the training of encoder, all training images are normalized to a mean shape with rigid transformation and cropped to 256×256 according to the landmarks predicted by state-of-the-arts. Then, in order to make the model more robust to data variations, standard data augmentation is performed including translation (± 10 pixels), rotation (± 5 degrees), scaling ($\pm 3\%$) and flip.

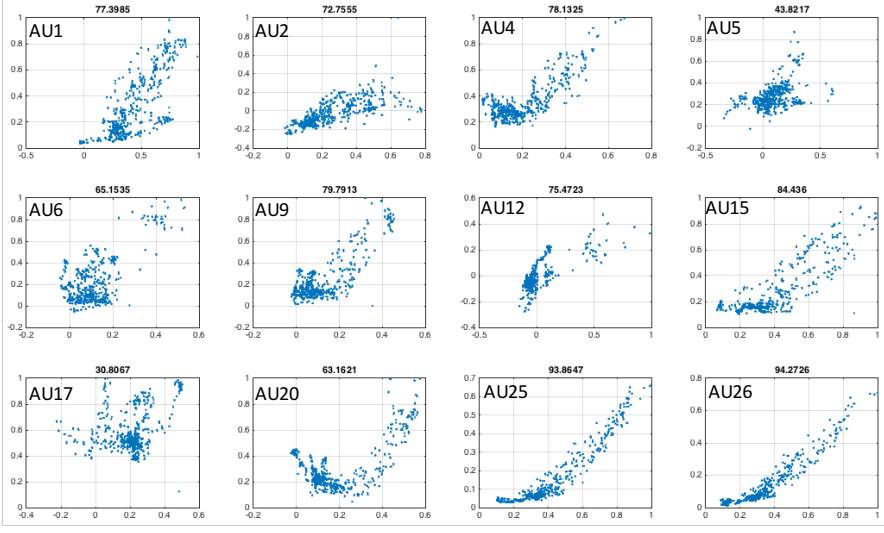
3 Facial Action Consistency

Following state-of-the-art methods of facial action units detection, we use 12 AUs defined in DISFA dataset [3] to evaluate the facial action consistency of our method. The definitions of the 12 AUs are Inner Brow Raiser (AU1), Outer Brow Raiser (AU2), Brow Lowerer (AU4), Upper Lid Raiser (AU5), Cheek Raiser (AU6), Nose Wrinkler (AU9), Lip Corner Puller (AU12), Lip Corner Depressor (AU15), Chin Raiser & Mentalis (AU17), Lip Stretcher (AU20), Lip Part (AU25) and Jaw Drop (AU26) respectively.

We report the full comparison with Face2Face of all 12 AUs correlation with the source in Figure. 2. It can be observed that in most of the AUs, our method records much higher correlations than Face2Face.

4 Video Demo

For the purpose of directly illustrating the face reenactment results of our method, a video is also included in the supplementary material.



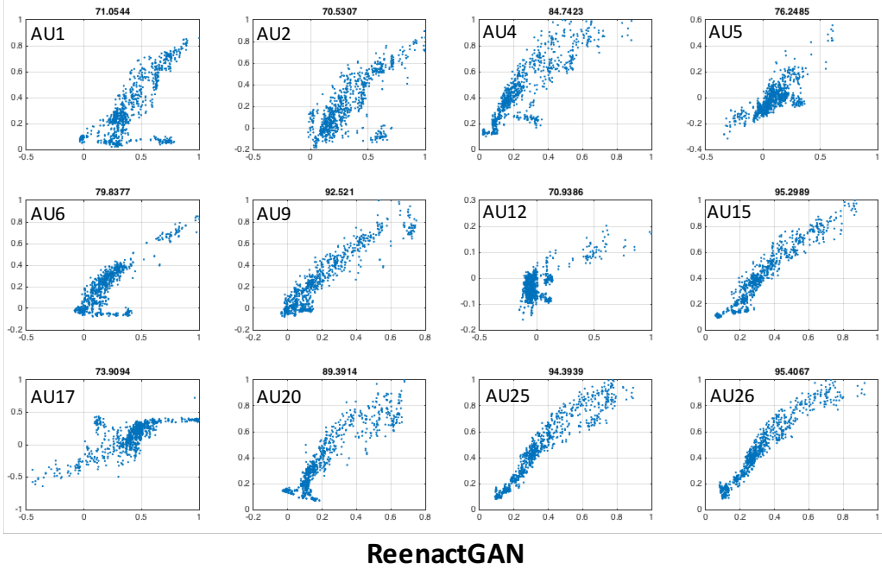


Fig. 2. Full comparison with Face2Face in facial action consistency.

References

1. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CVPR (2017)
2. Li, C., Wand, M.: Precomputed real-time texture synthesis with markovian generative adversarial networks. In: ECCV (2016)
3. Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., Cohn, J.F.: DISFA: A spontaneous facial action intensity database. IEEE Transactions on Affective Computing(TAC) **4**(2), 151–160 (2013)
4. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV (2016)
5. Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: CVPR (2016)
6. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networkss. In: ICCV (2017)