Vol.2020-CG-180 No.5
Vol.2020-DCC-26 No.5
Vol.2020-CVIM-223 No.5
2020/11/5

Video Face Swapping via Disentangled Representation Learning

Zhengyu Huang[1] Chunzhi Gu[2] Yi He[1] Chao Zhang[2] Xi Yang[3] Haoran Xie[1,a)]

Abstract: Face swapping is useful and popular technique to manipulate face from images and videos to be indistinguishable from authentic ones. However, it may cause mismatch with audio input in face appearance and shape by the conventional deep learning based approaches with facial masks. To solve this issue, we propose a novel face swapping approach using disentangled representation learning. First, the paired videos are used as training input after face alignment. We adopt GAN-based model to separate the features of the face appearance and mouth shapes explicitly by exchanging latent encodings. To maintain the continuity of video frames, we design the loss function between Inter-frame elaborately. With a given source image, the lip-sync speaking video can be generated by mimicking the target video. Our work is still in progress.

Keywords: Face Swapping, Disentanglement Learning, Face Appearance, Talking

1. Introduction

To swap faces from different images or videos is interesting and emerging technique due to the rapid development of deep learning approaches. Especially, deepfake can generate face images of a target person to a video of a source person and create a novel video of the target person doing or talking in the same styles of the original person in the video [3]. These approaches are usually implemented based on deep learning models, such as the open-source faceswap approach[6], which have been applied widely in computer vision research field [7].

Previous face swapping works usually depended on the facial mask or edge map extracted from images as the reference to enhance robustness of the result, meanwhile, the pre-trained facial parsing model or other prior knowledge is required in the learning processes [1]. ELEGANT proposed a disentanglement respresentation learning approach for face attribute transfer model [8]. Inspired from this work, we aim to explore a novel face swapping approach with no dependence of extra-information, such as facial parsing and prior facial information.

The main contribution of this work is exploring the disentanglement learning for face swapping in videos with the designed loss function to ensure frame continuity.

2. Framework

In order to achieve the face swapping of videos, our proposed model aims to swap the face to another person with the same lip-sync speech content by learning the features of the identity of face and the lip shape respectively.

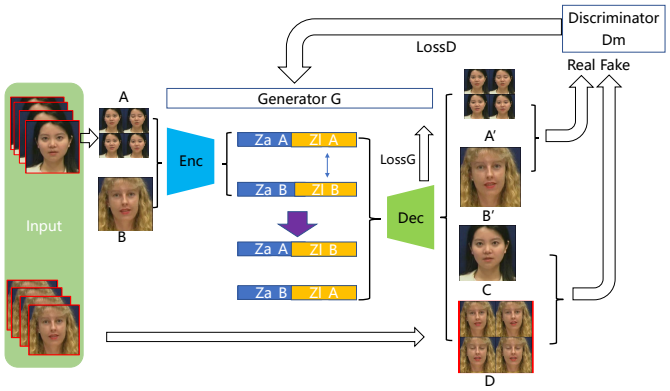


Fig. 1 The pipeline of our proposed face swapping model. The disentanglement embedding Z_a and Z_l are trained by the feature exchange respectively.

As shown in Figure 1, our proposed framework is based on Generative Adversarial Network (GAN) structure. We first train a swapping GAN networks consisting of a generator G (including encoder Enc and decoder Dec) and discriminator D_m. The system input of our proposed model includes the contents with a short speech video from one person and a source image of different person. In each step for training process, we choose two short videos of different persons with same speech contents from VidTIMIT Audio-Video Dataset [5]. Then, we resample the videos as two series of lip-sync frames based on the time stamps of words in the speech calculated by Aeneas [4]. We choose one as the input video frames A as shown in Figure 1, and the other video is used as the reference of the output video D for learning the correct feature embedding. C denotes the identity image of the person in video frames A. Meanwhile, the source image B can be resampled to the same size of the frames with video A as still video for data alignment.

[1] Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1205, Japan
[2] University of Fukui, Bunkyo, Fukui, 910-0017, Japan
[3] University of Tokyo, Bunkyo City, Tokyo 113-8654, Japan
[a)] xie@jaist.ac.jp

In this work, we divide the latent feature embedding z into two components, $Z = (Z_a, Z_l)$, and enforce the lip shape components Z_l of the two input separated from the identity component Z_a. Here, we adopt GAN loss including the $Loss_G$ and $Loss_D$ [2]. The latent feature embedding of input videos A and B is given as follows:

$$\begin{cases} Z_A = (Z_{aA}, Z_{lA}) = Enc(A) \\ Z_B = (Z_{aB}, Z_{lB}) = Enc(B) \end{cases} \quad (1)$$

We define C and D as the reconstruction results of face swapping by exchange the Z_l component in latent codes.

$$\begin{cases} C = Dec(Z_{aA}, Z_{lB}) \\ D = Dec(Z_{aB}, Z_{lA}) \end{cases} \quad (2)$$

The recovered videos A'and image B' is the reconstruction results from generator G in Figure 1. For discriminator D_m, A' and B' in outputs are thought as the real image while the C and D are regarded as fake image to improve the generator G to improve the generator G in training process.

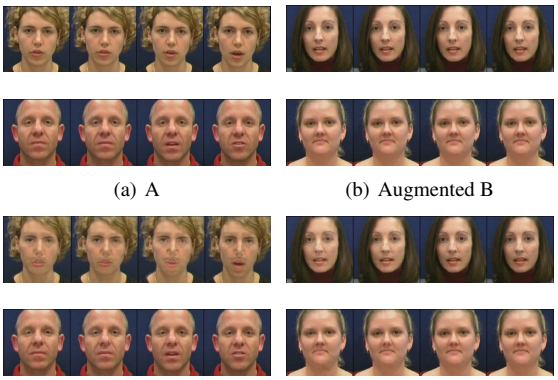
3. Results

In this paper, we would like to report the intermediate results of our proposed framework.

Fig. 2 Results of swapping attribute "similing" .From left to right: A, B, C, D; A and B are the input images in our model, C and D are swapping results.

To evaluate the effectiveness of our proposed model, we demonstrated the frames of images replaced by randomly sampled images on CelebA dataset. In this case study, our model adopted the similar structure as ElEGANT network [8]. Figure 2 shows that our proposed model can work well when swapping the smiling attributes. Our model can swap the facial features successfully between still images.

As shown in the Figure 3, we observed that the output frames D is closer to the input frames B while C is similar to A. However, C was blurred in 100th epoch because the gradient still existed at that time and disappeared in the training process later. Because the current dataset only contains short videos from 44 persons, we plan to augment the exist dataset or create novel one to improve the current results.



(a) A (b) Augmented B

(c) C (d) D

Fig. 3 Results of face swapping of videos. The upper rows of video frames are the results in 100th epoch while the lower ones are in 600th epoch.

References

[1] Burkov, E., Pasechnik, I., Grigorev, A. and Lempitsky, V.: Neural Head Reenactment with Latent Pose Descriptors (2020).

[2] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C. and Bengio, Y.: Generative Adversarial Nets, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada* (Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. and Weinberger, K. Q., eds.), pp. 2672–2680 (2014).

[3] Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C. S., RP, L., Jiang, J., Zhang, S., Wu, P., Zhou, B. and Zhang, W.: DeepFaceLab: A simple, flexible and extensible face swapping framework, *CoRR*, Vol. abs/2005.05535 (2020).

[4] Pettarin, A.: Aeneas, https://www.readbeyond.it/aeneas/ (2017). [Online; accessed 14-October-2020].

[5] Sanderson, C. and Lovell, B. C.: Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference, *Advances in Biometrics, Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings* (Tistarelli, M. and Nixon, M. S., eds.), Lecture Notes in Computer Science, Vol. 5558, Springer, pp. 199–208 (2009).

[6] shaoanlu: faceswap-GAN, https://github.com/shaoanlu/faceswap-GAN (2018). [Online; accessed 14-October-2020].

[7] Wang, T., Liu, M., Zhu, J., Yakovenko, N., Tao, A., Kautz, J. and Catanzaro, B.: Video-to-Video Synthesis, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada* (Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N. and Garnett, R., eds.), pp. 1152–1164 (2018).

[8] Xiao, T., Hong, J. and Ma, J.: ELEGANT: Exchanging Latent Encodings with GAN for Transferring Multiple Face Attributes, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X* (Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., eds.), Lecture Notes in Computer Science, Vol. 11214, Springer, pp. 172–187 (2018).