Engineering Design Document
TalkingHead

Team Members:
Yuming Gu, 6677706145
Jiahui Ding, 9563949830
Hongli Chen, 1326663389
Guangyun Zhou, 1012722928
Chaofan Zhai, 8632446003
Mengyu Zhang, 4907442582

1. Goal of the Project

TalkingHead is a machine learning project that making realistic and personalized neural talking head models in a one-shot setting. The reasoning behind applying machine learning algorithms to our TalkingHead project is to efficiently reenact a talking facial model of different game characters for the movie-licensed video game from a few image view of a character, potentially even a single image, improve the game character model with detailed information and furthermore change the current longtime low-quality stereotype of the cinematic-inspired games.

2. Background and Overview of Movie-Licensed Video Game

Throughout the 2000s, game publishers worked alongside Hollywood to ship games based on the biggest blockbuster movies, day and date with the film's release in theaters. With rare exceptions, these tie-ins were not very good games: The tight schedules of film production meant that the games had to be produced very quickly and couldn't be delayed for further polish, and the fact that Hollywood was pulling the strings meant that the games' developers didn't have much, if any, creative freedom. And yet even though they were never any good, movie-games were seen as an inevitable fact of life for the gaming world, since consumers still bought them based solely on the strength of the license. Though a lack of quality had always been a problem with movie-games, ever since the days when Atari cranked out a horrendous 8-bit version of E.T. in six weeks, people still purchased them. When movie games were popular in the early 2000s, consumers were generally unaware of video game brands such as Medal of Honor or Call of Duty. A major movie license that they had heard of would be the thing most likely to catch their eyes, and their wallets.

But things changed a lot in recent years. Gamers are expecting higher quality of the game rather than only the famous movie name. Thus tons of movie licensed games with

low quality disappeared in the market. In 2018, the top ten highest grossing films worldwide only had one video game adaptation between them (the Lego version of the Incredibles 2). Compared to 2008, where there were six video game adaptations of the ten highest grossing films (Kung Fu Panda, Madagascar: Escape 2 Africa, Quantum of Solace, Iron Man, Wall-E and The Chronicles of Narnia: Prince Caspian). The change being that publishers have slowly begun to realise that video games that have their own development cycle and creative ideas outside of the film tend to be received far more positively and then subsequently perform far better. Batman: Arkham City released a few months prior to the Batman film: The Dark Knight Rises, had its own development separate to the film, so didn't have to be rushed out to match the film's release date and was an incredible success, selling over ten million units. Marvel's Spider-Man for PS4 sold an enormous nine million copies. These facts remind us that the high-quality movie licensed games could still be successful.

Our TalkingHead will help the game producers to improve the quality of the movie licensed game, especially in the game character modeling. And the machine learning model will efficiently reduce the time and cost to acquire a high-quality face model for game characters, which will help the game developers to fit in the intensive development cycle of the movie licensed game. TalkingHead would apply machine learning algorithms by taking the character's face from movies to create the face model in the related video game. It will also make the perfect face expression of the game character in motion. It is not a specific game but a technique which is able to put into use for the development of various movie licensed video games.

3. Methodology

Considering the fact that the shorter and cheaper development cycles of movie-licensed video games cannot be changed overnight. Our project mainly focused on in a few-shot setting, how to make personalized game character models and then overcome the severe degradation in the quality of the face reenactment, potential fitting in the time and budget sensitive tie-in games. Prior researches have proved the way of how realistic human or human-like images can be generated with the help of training convolutional neural networks. However, these works require training on a large dataset of images of a single individual. We need to learn from a few image views of a game character, potentially even a single image.

The basic design of our project is to transfer the head pose and facial expression from a source video of one person to a target image of a game character and make the target image moving. Based on the observation that 2D facial landmarks contain information of

pose/expression as well as person-specific identity features(e.g. size, shape, proportion, and layout), we used a Feature Dictionary-based Generative Adversarial Network that takes both target and synthetic landmarks as input and translates it to a new face or face-like image. Here, we present a system with such few-shot capability and divide it into three parts:

3.1 Preprocessing: extract landmarks

Firstly, we need to extract the landmarks from the human face using Openface 2.0, which is a facial behavior analysis toolkit presented by Baltrusaitis et al. OpenFace 2.0 is an extension of OpenFace toolkit, which enables more accurate facial landmark detection, head pose estimation, facial action unit recognition, and eye-gaze estimation. The core algorithms of OpenFace 2.0 shows most recent outcomes in all of the tasks above. Furthermore, Openface 2.0 tool is able to perform real-time and can run from the webcam without any specialist hardware. What's more, this toolkit's source code is freely open to the public for research purposes.

Openface will refine the facial information into 68 landmarks. Based on the facial landmarks, we could crop the video into a human-face size (See Figure 1). We believe these key points contain a person's identity information, expression and pose features. Also, since landmarks are discretized, we propose a way to rasterize the landmark points into a continuous image as shown in Figure 2.

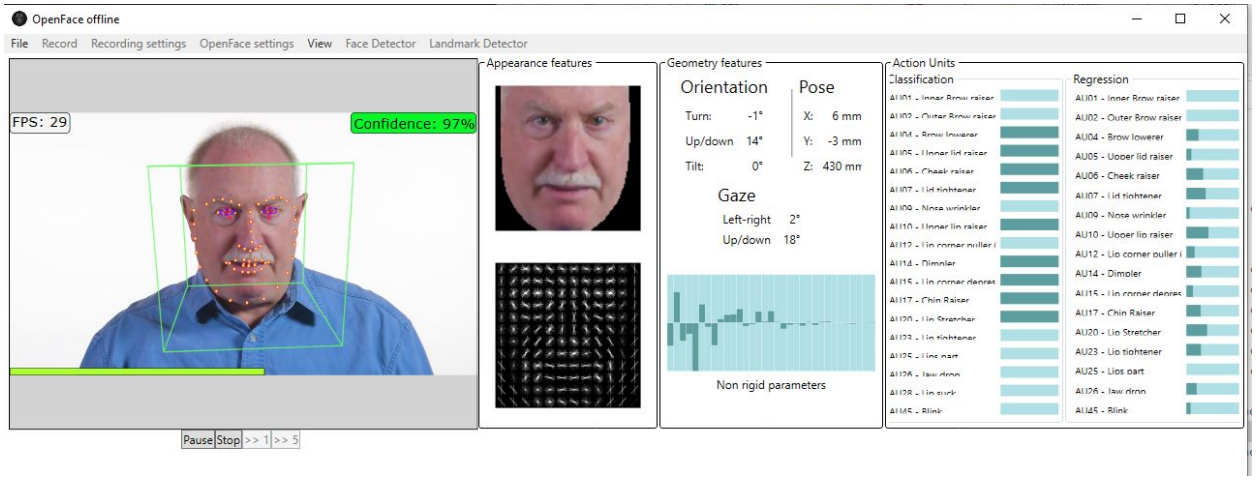

Figure 1 : Detecting the facial landmarks

Figure 2 : Cropping source videos into head size

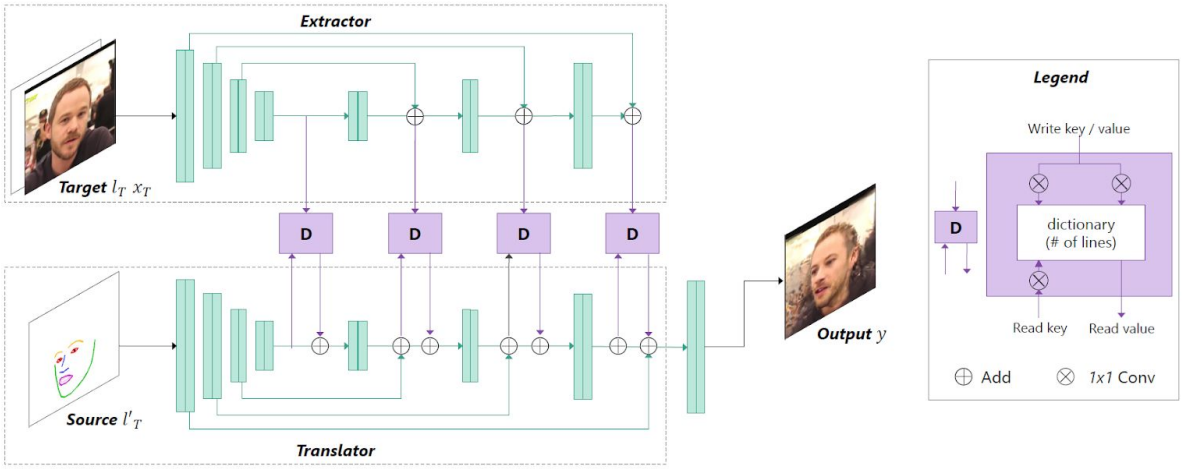

Figure 3:Dictionary-based generative adversarial networks

3.2 Generating: facial landmarks to a face image

With the predicted landmarks rasterized into a landmark image, our next goal is to translate it to a realistic face image. The translation procedures are as follows: a local patch around each location in the landmark image indicates “what facial part should be here”. And for each location, we want to translate it into “how should it look like”. Therefore, we are able to use a novel feature dictionary-based generative adversarial network(FD-GAN) to capture this intuition.

The architecture of our model is illustrated in Figure 3, which consists of an extractor and a translator. Given a target image and its corresponding landmark image, we will train an extractor that constructs a “feature dictionary” in the module D which is essentially a mapping from an annotation in the landmark image to its appearance in

the target image. Concurrently, given another landmark image and the feature dictionary, we train a translator that retrieves relevant facial features from the dictionary based on the landmarks and composes a face image.

3.3 Game Character Model: from only human to all game characters

Facial landmarks are important to driving one's face. Currently, we need to build a ready-made game character facial detector system and specific face detector in order to extract refined information from these characters. After finishing the above process, we need to transfer this generative network into game characters like pauline(mario's wife, see Figure 4) or other characters like Harry Potter and let them talk.

Figure 4: Example of our final result

Lots of previous face reenactment researches suffer from identity mismatch and produce inconsistent identities when a target and a driving subject are different, especially in one-shot setting. In our project, we aim to address identity preservation in cross-subject portrait reenactment from a single picture. Inspired by a work of Anonymous 2020 ECCV submission, we will utilize a novel technique that can disentangle identity from expressions and poses. In this way, even when the driver identity has great difference with the driver identity, we can still preserve its portrait reenactment. This effect can be achieved by novel landmark disentanglement network (LD-Net), which predict personalized facial landmarks that combine the identity of the target and expressions/pose of a different subject. We also need to deal with the portrait reenactment of unseen game character since it can be rebuilt over and over again. A feature which is dictionary-based generative adversarial network (FD-GAN) could locally translate 2D landmarks into a personalized portrait, enabling one-shot portrait reenactment under large pose and expression variations.

4. Related Researches

The traditional method of face reenactment are based on the use of explicit 3D modeling of human faces (Blanz and Vetter 1999) where the 3DMM parameters of the driver face and the target face are extracted from one-shot image, and combined to implement the motion (Thies et al. 2015; Thies et al. 2016). Another well-known approach is image warping, which uses the estimated flow extracted from 3D models (Cao et al. 2013) or sparse landmarks(Averbuch-Elor et al. 2017) to get the target image. The recent popular approach also exploring image-to-image translation architectures(Isola et al. 2017) on the success of neural networks, such as the works of Xu et al. (2017) and that of Wu et al.(2018), which included the cycle consistency loss (Zhu et al. 2017). Combined with these two methods, many studies proposed new approaches. Training an image translation network, Kim et al. (2018) maps reenacted render of a 3D face model into a photo-realistic output. Architectures, which are hybrid of the target's style information and driver's spatial information, have been studied recently. AdaIN (Huang and Belongie 2017; Huang et al. 2018; Liu et al. 2019) layer, attention mechanism (Zhu et al. 2019; Lathuiliere et al. 2019; Park and Lee 2019), deformation operation (Siarohin et al. 2018; Dong et al. 2018), and GAN-based method (Bao et al. 2018) are widely adopted as well. Other ideas such as the use of image-level (Wiles, Koepke, and Zisserman 2018) and feature-level (Siarohin et al. 2019) warping, and AdaIN layer in conjunction with a meta-learning (Zakharov et al. 2019) also make greatly progress. These approaches either require a dataset with image pairs that may be hard to acquire or an independent model per person.

5. References

** For some studies that have not been published in official conferences and journals, we didn't include them in the References. Thanks for those researches as well!*

Averbuch-Elor, H., Cohen-Or, D., Kopf, J., & Cohen, M. F. (2017). Bringing portraits to life. *ACM Transactions on Graphics (TOG)*, *36*(6), 196.

Blanz, V., & Vetter, T. (1999, July). A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques* (pp. 187-194).

Bao, J., Chen, D., Wen, F., Li, H., & Hua, G. (2018). Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6713-6722).

Cao, C., Weng, Y., Zhou, S., Tong, Y., & Zhou, K. (2013). Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, *20*(3), 413-425.

Dong, H., Liang, X., Gong, K., Lai, H., Zhu, J., & Yin, J. (2018). Soft-gated warping-gan for pose-guided person image synthesis. In *Advances in neural information processing systems* (pp. 474-484).

Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1501-1510).

Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 172-189).

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125-1134).

Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., ... & Theobalt, C. (2018). Deep video portraits. *ACM Transactions on Graphics (TOG)*, *37*(4), 1-14.

Liu, M. Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., & Kautz, J. (2019). Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 10551-10560).

Park, D. Y., & Lee, K. H. (2019). Arbitrary Style Transfer With Style-Attentional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5880-5888).

Siarohin, A., Sangineto, E., Lathuilière, S., & Sebe, N. (2018). Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3408-3416).

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019). Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2377-2386).

Siarohin, A., Lathuilière, S., Sangineto, E., & Sebe, N. (2019). Attention-based Fusion for Multi-source Human Image Generation.

Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., & Theobalt, C. (2015). Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, *34*(6), 183-1.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2387-2395).

Wiles, O., Sophia Koepke, A., & Zisserman, A. (2018). X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 670-686).

Wu, W., Zhang, Y., Li, C., Qian, C., & Change Loy, C. (2018). Reenactgan: Learning to reenact faces via boundary transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 603-619).

Xu, R., Zhou, Z., Zhang, W., & Yu, Y. (2017). Face transfer with generative adversarial network. *arXiv preprint arXiv:1710.06090*.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499-1503.

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).

Zakharov, E., Shysheya, A., Burkov, E., & Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 9459-9468).

Zhang, J., Zeng, X., Pan, Y., Liu, Y., Ding, Y., & Fan, C. (2019). Faceswapnet: Landmark guided many-to-many face reenactment. arXiv preprint arXiv:1905.11805.

Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., & Bai, X. (2019). Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2347-2356).