Facial Reenactment from Sparse Landmarks using StyleGAN

Theresa Bruns Yoonha Choe
Technical University of Munich

Abstract

With powerful architectures now available in the field of visual computing, realistic images and videos can be generated in high-quality. Profiting from this advance is the task of face reenactment and manipulation that finds many relevant applications today. Building on the recently proposed version of StyleGAN3, this paper introduces a novel method of face reenactment combining expression transfer via facial landmarks and semantic manipulation of emotions. Results demonstrate high-quality outputs and on-par or better performance than the baseline methods.

1. Introduction

The task of face reenactment aims at taking the facial expression of a source person and transferring it to a (usually distinct) target identity [21, 9, 20, 18, 19]. As a result, output frames are generated depicting the target face with the expression of the source or driving face [18, 19]. The ability to realistically control images or videos of human faces forms the basis of many applications like the post-processing of movies, visual dubbing to different languages, virtual reality or video conferencing [21, 9, 17, 18]. In this paper, we propose a new method of facial reenactment that enables the control of an input identity in a two-fold way: On the one hand, the facial expressions and movements of a distinct source face can be transferred using sparse facial landmarks. At the same time, manipulation of the target identity's emotions is made possible in an intuitive and continuous manner. At the core of our method lies the StyleGAN3 architecture and its corresponding latent space [5]. Recently proposed in its third version [6, 7, 5], this style-based generator network has demonstrated powerful results in the generation of realistic human faces [5] and therefore presents the optimal basis for our approach.

2. Previous Work

The given problem as we tackle it in our approach can be roughly subdivided into three main parts. As a first step, a representation in the latent space of StyleGAN3 must be found for each input video frame in the form of an embedding vector in order to be able to perform image manipulation in this manifold. Based on the obtained latent vector the actual face reenactment of the target identity with the source expression will be performed in the second part. Lastly, the additional possibility of emotion manipulation as proposed by our method makes up the third part. For each of these subtasks, previous methods have been proposed that will be briefly presented in this section.

Generally, there exist two different approaches to finding proper embedding vectors: Learning an encoder that embeds the images or direct iterative optimization of the latent vector. Since the former often comes with drawbacks regarding generalization capabilities beyond the training dataset, current works mostly revert to the latter method for more stable reconstruction results [2, 1, 14]. On top of that, previous works have also introduced a hybrid method combining both the above approaches [22]. Having realized that the quality of the embedded reconstruction of an image is heavily reliant on a good initialization of the latent vector when doing direct optimization, Zhu et al. [22] first train an encoder network which predicts a suitable embedding that is then used as the initial vector for the subsequent optimization [22]. As this combined approach has shown better performance and a stronger resilience towards local minima [22], it is the method we adopt in this paper as well.

Regarding the task of face reenactment with source expressions, there are model-based approaches that base themselves on 3D morphable models encoding identity and expression [21, 19]. Two very prominent methods among these are Face2Face by Thies et al. [17] and Deep Video Portraits by Kim et al. [9]. A drawback of these model-based methods, however, is that the generation and tracking of such a 3D model can be complex and time-consuming. What is more, there are parts of the face that often have to be taken special care of like the mouth interior or eye gaze that would require separate rendering [17, 9, 4, 19, 18, 20].

Many works therefore make use of data-driven methods for the face reenactment task which do not explicitly construct a model for the face appearance. Instead, they learn patterns from large collections of data [21]. Earlier image-based approaches use 2D facial features and often perform an image matching step based on a database of expressions [8, 11, 10, 4]. Wiles et al. propose X2Face, a network

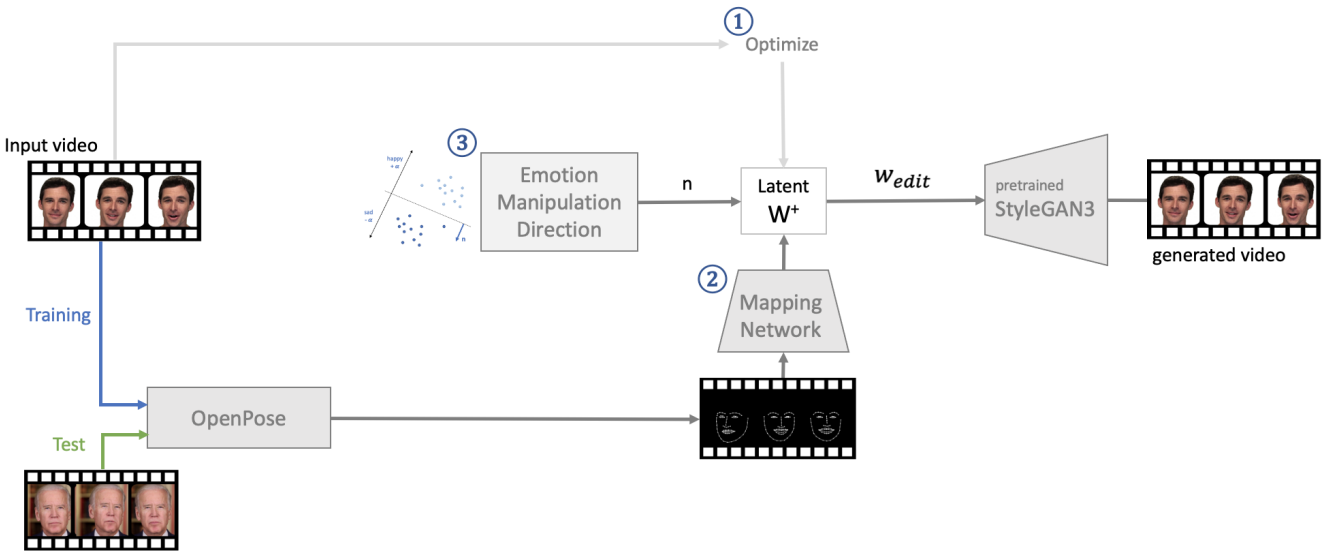

Figure 1: Overview of the entire proposed approach

that learns an embedded face representation from a source video and performs face frontalization in order to perform reenactment with another expression source [20]. More recently, Tripathy et al. present ICFace [18] and FACE-GAN [19], both GAN-based methods using so-called Action Units (AU) to represent facial attributes and movements [18, 19]. Another quite common descriptor of facial expressions are facial landmarks since they offer a very intuitive, effective and low-dimensional representation [16, 21]. Also, a landmark map is easily obtained from any face image or video using reliable algorithms available in the state-of-the-art [16, 3]. This is why, in this paper, facial landmarks are chosen as a representation in connection with the StyleGAN3 manifold that is conveniently disentangled [6]. The extraction of the landmarks is performed using OpenPose [3].

By additionally changing the latent vector in a certain way it is possible to change, e.g., the smile of a depicted person while not only preserving its identity but also all other attributes like their age or gender [19, 23, 15, 14]. There are different methods to determining the required change of latents that either work with a separate network optimization process [13, 22] or aim at identifying meaningful linear manipulation directions in the embedding space [15, 23]. In correspondence with the chosen dataset, our method is built upon the latter as inspired by InterFaceGAN [15].

3. Method

As introduced by the three subtasks in section 2 that can be distinguished in our approach, the description of our method will be structured in the following:

- Part 1: Hybrid Optimization of embeddings
- Part 2: Expression transfer with landmarks
- Part 3: Emotion Manipulation

3.1. Hybrid Optimization of embeddings

According to the architecture of the StyleGAN3 generator network [5], there are multiple latent spaces that could be candidates for the frame embedding approach. In current works, it is common to work with the extended latent space W+, which for StyleGAN3 is a concatenation of 16 512-dimensional w vectors where each corresponds to the input of a separate GAN layer, respectively [5, 2, 14]. We adopt a hybrid approach [22] for finding the optimal embedding. This involves first learning the optimal initialization and then do the embedding optimization itself using the learned initial latent vector. In our case the initialization is learnt using StyleGAN-generated images, which we know the exact representation in latent vector form of.

After having trained the encoder using MSE loss between predicted and known latents, the model parameters are saved for the subsequent optimization. Now taking the first frame of the actual input video, it is passed through the encoder to obtain a latent vector representation which is then iteratively updated. The objective used at this point is a combination of MSE and LPIPS loss between the original frames and the ones generated using the current latents. For subsequent input video frames, the previously obtained latent vector is taken as initialization instead so the trained encoder is only used once in the beginning for the very first frame. Generally, latents are adopted as the optimized one and saved for later use once the loss drops below a threshold of 0.03 or 1500 epochs are exceeded.

3.2. Expression transfer with landmarks

Once all necessary latent embeddings are available through the optimization step, this creates the possibility of manipulating the respective video in a desired manner. As mentioned already in section 2, for face reenactment this paper is based on landmarks that are extracted via OpenPose [3]. In order to be able to manipulate a given latent vector in such a way that only the facial expression is changed while the identity remains preserved, we aim at learning a concrete correspondence between the landmark map and a given embedding. This is realized by training a mapping network as shown in figure 2. Once the mapping from landmarks to the StyleGAN latent vector has been learned properly, the actual face reentactment can then be performed by simply taking the landmarks of a different source actor and feeding them to the mapping network that will then output an embedding which - when fed into the StyleGAN3 network - yields a frame of the target identity with the source expressions. This procedure can be seen in the overview of the entire method in figure 1.

3.3. Emotion Manipulation

For the additional semantic manipulation step of our method, we chose to focus on editing emotion attributes, i.e. changing the degree of happiness or sadness acted out by the identity in the video, for example. For this purpose and because we are working in the powerful latent space of StyleGAN3 [5], it is necessary to find the corresponding direction in latent space that leads to this desired face edit. To this end, a SVM classifier is adopted as shown in figure 3.

The classifier takes as input several labelled embeddings and finds a separating hyperplane of which a normal vector can be extracted. This normal vector corresponds to the

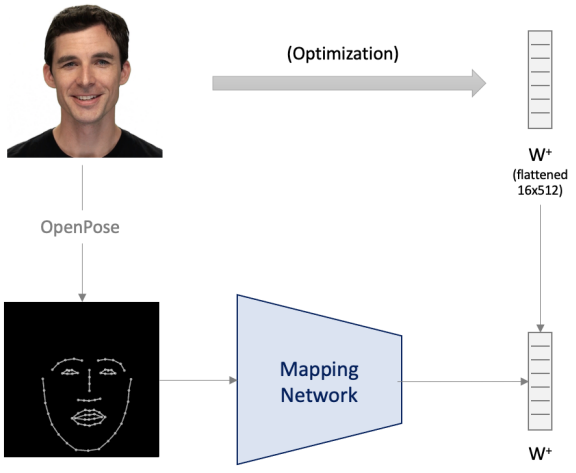

Figure 2: Overview of the mapping network training scheme

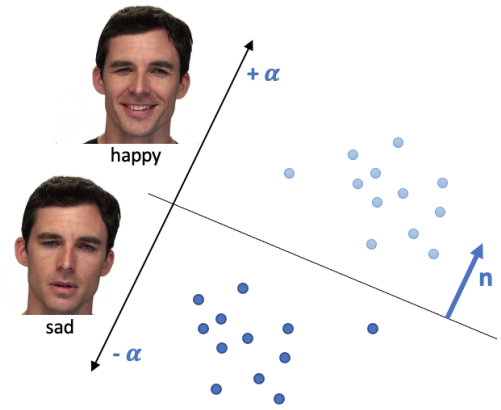

Figure 3: Principle of the SVM classifier approach for finding editing directions

desired editing direction. Finally, the latent vector for the semantically manipulated face is then obtained as follows:

$$w_{edit} = w + \alpha * n,$$

where α is the degree to which an edit is performed corresponding to the step size in latent space [15].

Due to this supervised SVM approach it is necessary to work with a dataset that is labelled accordingly. For this reason, we have decided to base all training procedures on the RAVDESS dataset [12].

4. Results

In this section we will show the results produced by our framework as well as draw comparisons to existing methods both qualitatively and quantitatively.

4.1. Qualitative Results

We first visualize some optimization results for the latent vectors in figure 4. Comparing the top and bottom images, it can be seen that the embedded images have high perceptual quality and reproduce almost identical image with the input frame.

For the face reenactment part, we evaluate our model against the baseline models ICFace [18] and X2Face [20]. The evaluation is done on the test set which consists of three speech videos (Obama1, Obama2, and Joe) containing 2735, 1803, and 2049 frames, respectively. Figure 5 shows a qualitative comparison of each method on the test set. It shows that our method preserves the identity of the target actor the most and generates the most realistic images. ICFace can capture the source actor's facial expressions and head movement, but the identity of the generated person is quite different to the target actor. X2Face can also capture the source actor's facial expressions, but if the source actor's face is not in front, then the generated image doesn't look like a real person and lacks a lot of details.

Figure 4: Qualitative results for optimization of latent vectors. (Top) Input frames. (Bottom) Visualization of optimized latent vector.

Regarding the third part of our framework, semantic manipulation of additional attributes, we extracted a latent direction that corresponds to a change in the emotion of sadness for negative alpha values and happiness for positive alpha values, respectively. The results can be seen in figure 6. One can clearly see the change of emotion in the actor's face. From the left-most face expressing complete sadness, corresponding to negative alpha values, there is a gradual and semantically meaningful change towards clear happiness in the right-most frame that is obtained with increasingly positive alpha values. The results turn out to be very intuitive and of high quality.

4.2. Quantitative Results

We compare the results of face reenactment quantitatively by using landmark difference (LMK) and Fréchet Inception Distance (FID) as evaluation metrics. LMK aims to evaluate the fidelity of the generated face images in terms of preserving the landmarks of the source video frame. We calculate the L2 landmark difference between the original source frame and our generated frame, where lower values indicate better performance. As shown in table 1, our method shows better performance than the baselines only for Joe's video, and X2Face achieves the lowest value for Obama's videos. FID aims to measure the realism and variation of generated frames. Table 2 shows that our method achieves the lowest value for FID scores for all source videos, which means our method generates the most realistic images.

In order to evaluate the emotion manipulation part of our method as well, we perform a separate evaluation with quantitative metrics in comparison with the chosen baseline of Zhuang et al. [23] based on the FFHQ dataset [6]. For the assessment of realism in the outputs FID is used as a measure here as well. Additionally, the Learned Perceptual Image Patch Similarity (LPIPS) is adopted, which has already been used as a loss in the embedding optimization part 3.

LMK ↓	Obama1	Obama2	Joe
ICFace	8.74	8.16	15.84
X2Face	**3.12**	**2.35**	13.25
Ours	5.22	4.24	**11.33**

Table 1: Quantitative results of face reenactment using landmark difference.

FID ↓	Obama1	Obama2	Joe
ICFace	76.19	76.95	75.01
X2Face	73.56	70.25	117.66
Ours	**39.01**	**39.56**	**36.84**

Table 2: Quantitative results of face reenactment using FID.

Method	LPIPS ↓	FID ↓
Zhuang et al.	**0.1627**	59.0868
Ours	0.5054	**47.778**

Table 3: Quantitative results of emotion manipulation using LPIPS and FID

The LPIPS is a perceptual similarity measure based on image patches and is used here in order to evaluate the perceptual quality between the semantically edited images and the original frame. Table 3 shows the results. While the baseline performs better with respect to the LPIPS measure, our approach surpasses them regarding FID which again points towards a high realism of our generated outputs. The performance of our method regarding LPIPS is actually well in the range of other state-of-the-art approaches as can be seen in the StyleFlow reference [13]. We hypothesize that the much lower value of Zhuang et al. is due to the explicit choice of images that their code has been made available on and based on which this evaluation has been performed.

5. Conclusion

All in all, our proposed approach provides yields some advantages compared to previous works. As a data-based method it does not require any re-rendering or 3D modeling but is instead based on a sparse set of facial landmarks for the expression tracking. Also, working in the StyleGAN latent space, it is able to profit from the strong image generation quality shown in recent advancements of StyleGAN3 [5]. Additionally, providing emotion manipulation capabilities in a simple but intuitive way, the proposed method represents a novel and powerful solution to the problem of face reenactment. Results show that our approach not only provides outputs of high visual quality but also that it surpasses current works regarding FID.

Figure 5: Qualitative results for face reenactment. The source images are shown in the first row, 5 frames from Obama and Joe's speech video, respectively. The results correspond to ICface (second row), X2Face (third row), and ours (last row).

Figure 6: Results for the semantic emotion manipulation with gradual emotion change from sad ($-\alpha$, left) to happy ($+\alpha$, right)

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? *CoRR*, abs/1911.11544, 2019.

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? *CoRR*, abs/1904.03189, 2019.

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1812.08008, 2018.

[4] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormählen, Patrick Pérez, and Christian Theobalt. Automatic face reenactment. *CoRR*, abs/1602.02651, 2016.

[5] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *CoRR*, abs/2106.12423, 2021.

[6] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, 2019.

[7] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *CoRR*, 2020.

[8] Ira Kemelmacher, Aditya Sankar, Eli Shechtman, and Steven Seitz. Being john malkovich. volume 6311, pages 341–353, 09 2010.

[9] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *CoRR*, abs/1805.11714, 2018.

[10] Kai Li, Qionghai Dai, Ruiping Wang, Yebin Liu, Feng Xu, and Jue Wang. A data-driven approach for facial expression retargeting in video. *IEEE Transactions on Multimedia*, 16(2):299–310, 2014.

[11] Kai Li, Feng Xu, Jue Wang, Qionghai Dai, and Yebin Liu. A data-driven approach for facial expression synthesis in video. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012.

[12] Stephen R. Livingstone and Frank A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. 2018.

[13] Niloy J. Mitra Rameen Abdal, Peihao Zhu and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics, Volume 40, Issue 3, Article No. 21, pp. 1–21*, 2021.

[14] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. *CoRR*, abs/2008.00951, 2020.

[15] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. *CoRR*, abs/1907.10786, 2019.

[16] Pu Sun, Yuezun Li, Honggang Qi, and Siwei Lyu. Landmarkgan: Synthesizing faces from landmarks. *CoRR*, abs/2011.00269, 2020.

[17] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of RGB videos. *CoRR*, abs/2007.14808, 2020.

[18] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Icface: Interpretable and controllable face reenactment using gans. *CoRR*, abs/1904.01909, 2019.

[19] Soumya Tripathy, Juho Kannala, and Esa Rahtu. FACEGAN: facial attribute controllable reenactment GAN. *CoRR*, abs/2011.04439, 2020.

[20] Olivia Wiles, A. Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation by using images, audio, and pose codes. *CoRR*, abs/1807.10550, 2018.

[21] Jiangning Zhang, Xianfang Zeng, Yusu Pan, Yong Liu, Yu Ding, and Changjie Fan. Faceswapnet: Landmark guided many-to-many face reenactment. *ArXiv*, abs/1905.11805, 2019.

[22] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. *CoRR*, abs/1609.03552, 2016.

[23] Peiye Zhuang, Oluwasanmi Koyejo, and Alexander G. Schwing. Enjoy your editing: Controllable gans for image editing via latent space navigation. *CoRR*, abs/2102.01187, 2021.