Emotion Editing in Head Reenactment Videos using Latent Space Manipulation

Valeriya Strizhkova, Yaohui Wang, David Anghelone, Di Yang, Antitza Dantcheva, François Brémond

Emotion Editing in Head Reenactment Videos using Latent Space Manipulation.

Valeriya Strizhkova[1,2], Yaohui Wang[1,2], David Anghelone[1,2], Di Yang[1,2],
Antitza Dantcheva[1,2], François Brémond[1,2]
[1] Inria [2] Université Côte d'Azur

***Abstract*— Video generation greatly benefits from integrating facial expressions, as they are highly pertinent in social interaction and hence increase realism in generated talking head videos. Motivated by this, we propose a method for editing emotions in head reenactment videos that is streamlined to modify the latent space of a pre-trained neural head reenactment system. Specifically, o ur m ethod s eeks t o disentangle emotions from the latent pose and identity representation. The proposed learning process is based on cycle consistency and image reconstruction losses. Our results suggest that despite its simplicity, such learning successfully decomposes emotion from pose and identity. Our method reproduces facial mimics of a person from a driving video, as well as allows for emotion editing in the reenactment video. We compare our method to the state-of-art for altering emotions in reenactment videos, producing more realistic results that the state-of-art.**

I. Introduction

Generative adversarial networks (GANs) [10] have become a leading paradigm in high fidelity face synthesis [17], [18], [36], [34], [33] and talking head video generation [3], [32], [35]. Despite remarkably appealing generated video quality, one remaining challenge has to do with *exploring the latent representation* and related control of the generation process. Specifically, a novel question that we explore in this work involves whether given a latent code of a pre-trained generated talking head, we can manipulate/edit emotion in associated talking head video (without affecting face identity or speech). We note that incorporating emotions into generated facial videos is essential, as humans are highly sensitive to subtle artefacts, rendering face generation highly challenging [21].

Further, *editing emotions* in reenactment videos is of particular interest, as emotions are highly correlated with facial expressions related to speech. For example, a facial expression can substantially alter the perception of a speech; a speech delivered, while smiling can be perceived differently than the same speech delivered, while frowning. Thus, disentangling emotion from speech related facial dynamics is pertinent.

Motivated by the above, we here jointly reenact faces, as well as edit emotions (*e.g.*, anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise) in a single model. We focus particularly on editing emotions rather than on changing attributes such as hair color, gender and age, as emotions are correlated with both, appearance and speech, contributing highly to realistic generated videos.

Approaches for editing facial attributes in reenactment videos include the following. Firstly, approaches, which use images with target emotions as input of a head reenactment system. However, finding input images with required emotions highly restricts the use cases. Secondly, approaches, which edit an emotion in a reenactment video by editing each image of a synthesized reenactment video. Such approaches involve two models trained on two different tasks, head reenactment, as well as emotion editing. We note that using two models trained for different generative tasks is time-consuming and might result in large identity gap and less realistic results.

Deviating from the above, we here propose a novel architecture that allows for direct modification of emotion in a reenactment system by alteration of the latent space in a pre-trained and fixed head reenactment GAN. This approach has the benefits that it does not require selecting images with the required emotion as input, nor does it necessitate the use of additional facial attributes editing models. To the best of our knowledge, we are the first to propose such an architecture for emotion editing in reenactment videos.

In particular, we adopt the state-of-the-art head reenactment system Latent Pose Descriptors (LPD) [3] and propose to disentangle emotion, identity and pose in the latent space.

Our contributions are summarized as follows.

- We propose an approach aimed at emotion editing in reenactment videos based on modification of the latent space of a pretrained GAN.
- We compare a set of techniques for emotion editing in head reenactment videos and show that our proposed method produces the most realistic results.

II. Related Work

A detailed review of GAN architectures, as well as discussion of loss functions can be found in recent overview articles [23], [37].

A. Conditional GANs

A dynamic area of research has to do with designing GANs that incorporate conditions (*e.g.*, attribute labels) into the generation process. Such conditional GANs (CGANs) generate images with desired properties under the constraint of additional conditional discrete or continuous variables [4], [24], conditional images [3], [5] or conditions of different modalities [32]. *Editing facial attributes* allows the change of attributes such as gender and age [4], [15], [20], [24].

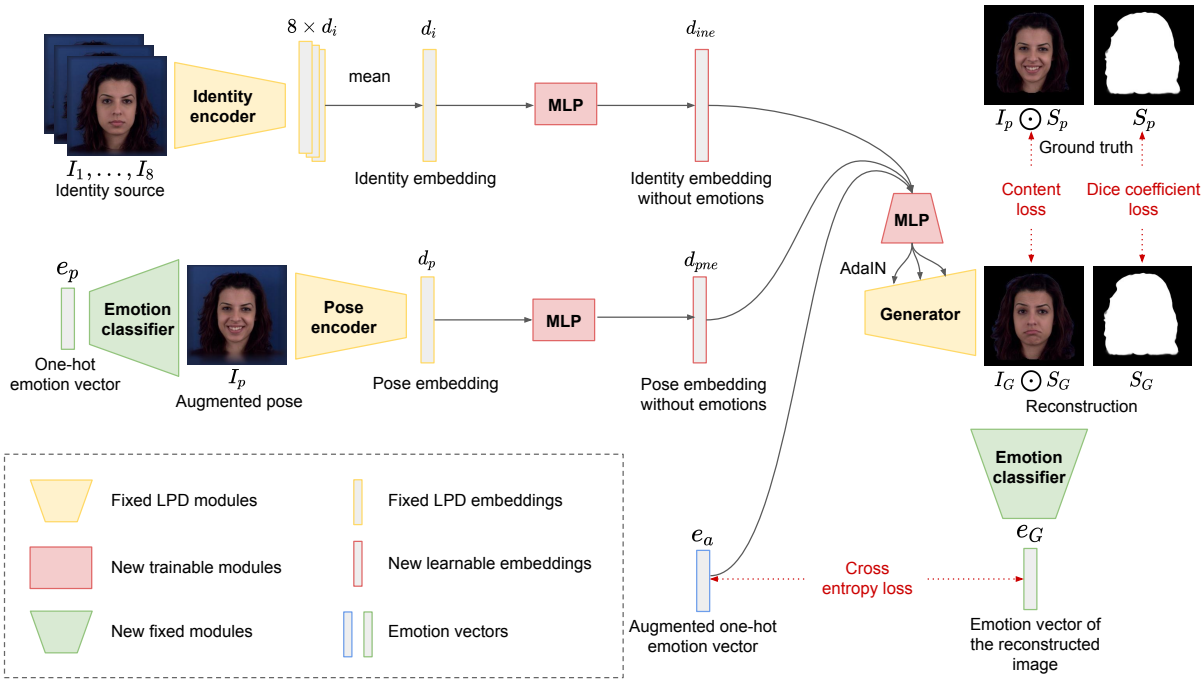


Fig. 1: Architecture of the proposed emotion editing system. The Latent Pose Descriptors (LPD) head reenactment system is extended by three MLPs. Identity encoder, Pose encoder and Generator stem from the original LPD, and are being fixed during training of our proposed emotion editing system. The Emotion classifier, a new module trained on the MUG or MEAD datasets, is also fixed during training and is used to classify emotions of the input driving pose image and the reconstructed image.

CGANs can be trained even with unpaired training data using cycle-consistency loss [4], [5], [40]. Our approach is related to these works exploiting cycle consistency to preserve key attributes between input and mapped image. While conditional GANs provide a level of (semantic) attribute control, they fail to reach image-quality produced by unconditional GANs, *i.e.,* the synthesized images can be blurry and encompass large face identity gap. We here perform attribute-based editing by conditionally exploring the latent space in a pre-trained fixed GAN, rather than conditional generation, requiring attribute-based retraining.

B. Facial Expression Manipulation

Facial expression manipulation can be treated as a CGAN that aims at editing facial expression with a given condition as an image or a variable. StarGAN [4] constitutes a method that considers discrete emotion categories (*e.g.*, happy, neutral and sad) and exploits cycle consistency. Similar works [20], [24] edit an input image under the guidance of facial Action Units (AU) to generate the desired expression. The above-mentioned facial expression manipulation methods change emotions in *images*. These facial expression manipulation methods can be applied for each frame of a head reenactment video, however the resulting quality is not guaranteed to be temporarily consistent and speech related facial expression might not be preserved. Differently, our method is designed to *edit emotions in talking head videos*, while preserving speech related facial expression.

Emotional Video Portraits (EVP) [15] is capable of generating emotion-controllable talking portraits and change smoothly their emotion by interpolating the latent space. However, EVP requires audio as input. Out method is the first to alter emotions in head reenactment videos using as input only the visual modality.

C. Head Reenactment

Motivated by face animation in computer-generated movies and digital games, head reenactment entails generating a video sequence, where a head from an identity *source image* is *animated based on* facial expressions and head movements in a *driving video*. Hence, facial dynamics and facial expressions can be transferred from a driving *video* to a source image [3], [7], [28]. Siarohin *et al.* [28] used a set of self-learned keypoints jointly with local affine transformations to model complex motions. Burkov *et al.* [3] proposed a neural head reenactment system, driven by a latent pose representation that is capable of predicting the foreground segmentation alongside the RGB image. Proposed system generated realistic reenactments of arbitrary talking heads using arbitrary driving videos to drive pose by firstly decomposing pose and identity. Gafni *et al.* [7] synthesized novel head poses as well as changes in facial expression by reconstructing a dynamic neural radiance field representing a 4D facial avatar. In addition, faces can be animated *based on audio* [15], [32]. Our method adopts as pre-trained GAN the LPD [3] head reenactment approach.

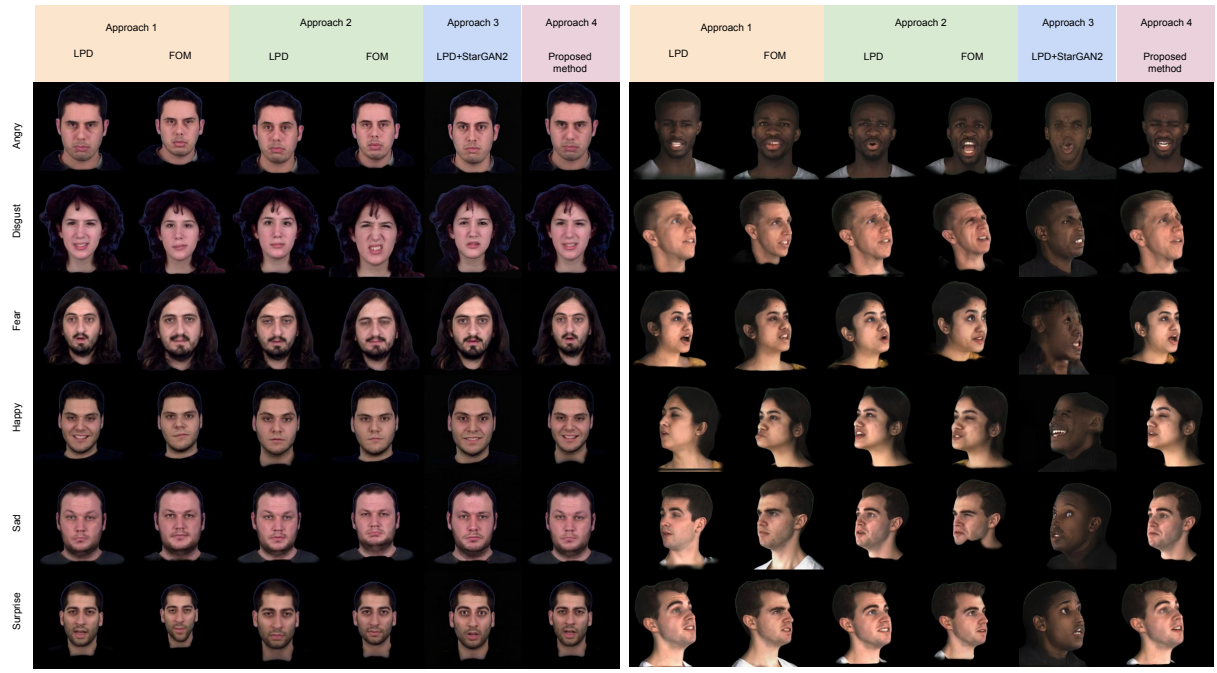


Fig. 2: Emotion editing in reenactment videos on the MUG (left) and MEAD (right) datasets. Each row represents a set of frames generated by a method seeking to reenact the emotion indicated on the left.

Among state-of-the-art head reenactment systems, LPD [3] produces good quality reenactment videos, as well as it entails an encoder-decoder architecture with a latent representation of pose and identity, that allows for modification of the latent space of the pre-trained LPD model, in order to change the desired emotion. In spite of high quality of synthesized videos, the functionality of LPD is limited with head reenactment generation. We extend LPD with a new functionality of emotion editing.

D. Study on Latent Space of GANs

The latent space of GAN models incorporates semantically meaningful directions. Moving in these directions corresponds to human interpretable image transformations, such as changing facial attributes. Associated works [13], [19], [25], [27] proposed linear manipulations of the latent space. Some works [19], [25] studied the vector arithmetic property in the GAN latent space. InterFaceGAN [27] interpreted the face semantics emerging in the latent space of GANs with the help of off-the-shelf classifiers. GANSpace [13] unsupervisedly discovered the latent semantics learned by GANs using PCA.

Goetschalckx *et al.* [8] navigated the manifold in the latent space, rendering images more or less memorable, *i.e.* aiming at affecting the human memory performance. Jahanian *et al.* [14] shifted the data distribution by steering the latent code to fit camera movements and color changes. Yang *et al.* [38] explored the emergent semantic hierarchy in scene synthesis models. Voynov and Babenko [31] interpreted meaningful directions in the GAN latent space by unsupervisedly training a direction reconstructor. Our technique falls into the category of methods that do not design a separate architecture, but manipulate latent codes of a pretrained GAN.

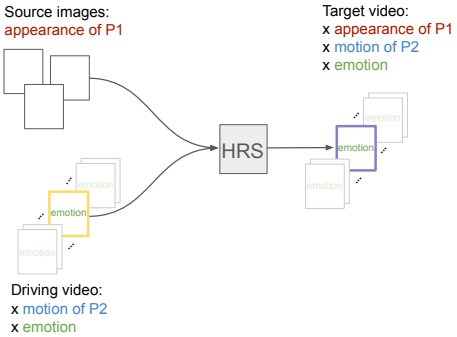
The abovementioned methods control the latent space of StyleGAN [17], StyleGAN2 [18], PGGAN [16], BigGAN [2], which synthesize images from noise. Differently, we focus on modifying the latent space of LPD, which takes as inputs source and target images. Synthesizing talking heads using pre-trained fixed image generators such as StyleGAN remains a challenging problem, as it is hard to find a meaningful trajectory in the image generator's latent space that renders temporally consistent images [29]. Therefore, it is not clear, how to apply existing methods designed to control the latent space of image generators such as StyleGAN to edit emotions in reenactment videos. Further, it is not fair to compare our model to the above mentioned methods, as the hyperparameters proposed to analyze the latent space of StyleGAN [17], StyleGAN2 [18], *etc.* are not guaranteed to provide meaningful results for analyzing the latent space of LPD.

III. Method

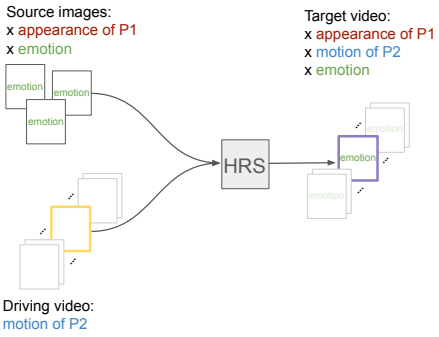
In this section, we describe our proposed framework and its training objective functions.

A. Architecture

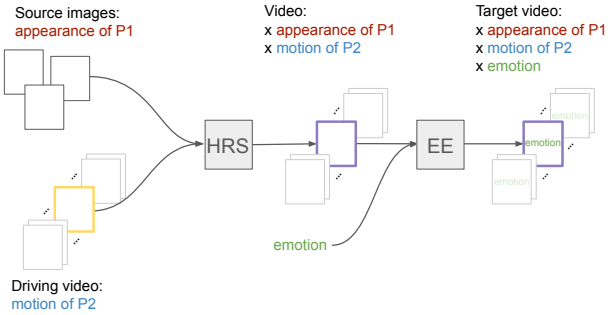
We select the LPD head reenactment system as the base-framework, which we extend by the emotion editing module, as LPD generates high-quality head reenactment videos, as well as the system entails latent representations of identity and pose that can be used for facial attributes editing via latent space manipulation. We note that LPD performs head reenactment based on latent pose vectors by decomposing



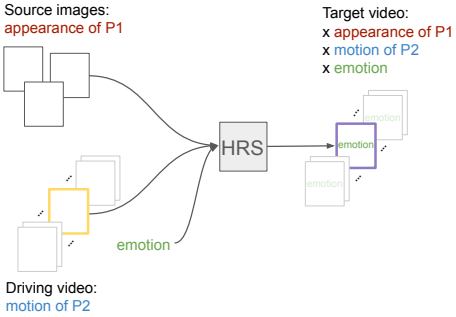
(a) Approach 1. Feeding a driving video with the target emotion into HRS.



(b) Approach 2. Feeding identity source images with the target emotion into HRS.



(c) Approach 3. Feeding the output of HRS into the emotion editing (EE) model which changes the emotion.



(d) Approach 4. Feeding the target emotion into HRS as a variable.

Fig. 3: Approaches to modify emotions in reenactment videos. In each of the four testing pipelines a frame of a driving video (yellow) and identity source images are fed into a head reenactment system (HRS).

pose and identity, preserving the identity of the reenacted person.

We fix LPD during training of our proposed emotion editing module and train only new modules designed to manipulate the latent space of pose and identity. In particular, we extend LPD with following three modules (see Figure 1).

- $MLP_{identity}$ takes as input the identity embedding from LPD and removes the information pertaining to emotions, producing a new identity embedding.
- MLP_{pose} is similar to $MLP_{identity}$, as it removes the emotion information from the LPD pose embedding.
- MLP_{adain} takes as input the target emotion, identity and pose embeddings, which are outputs from $MLP_{identity}$ and MLP_{pose}, respectively, and produces AdaIn [12] parameters which are fed into the generator.

Each MLP is a three linear layer followed by a ReLU activation function.

Emotion classification network is a ResNet-50 trained on the MUG and MEAD datasets to predict the emotion label. This module is fixed during the training of MLPs.

B. Training pipeline

The training pipeline consists of two steps (see Figure 4). During the first step nine images $I_1, \ldots, I_9$ and corresponding nine segmentations $S_1, \ldots, S_9$ are selected from the videos of one person performing facial expressions with different emotions. Eight images $I_1, \ldots, I_8$ are used as identity source and one image I_9 associated to emotion e_9 is used as the driving pose. Each of the eight identity source images is fed into the Identity encoder, which outputs eight identity embeddings $8 \times d_i$ and the mean of the identity embedding d_i is fed into $MLP_{identity}$. Pose embedding d_p is obtained after feeding the augmented driving pose image I_9 into the Pose encoder and is fed into the MLP_{pose} module. We use the same augmentation of the driving pose image as in the original LPD framework. Augmented emotion e_a, *i.e.*, the emotion different from the emotion e_9, and output vectors from the $MLP_{identity}$ and MLP_{pose} modules are fed into MLP_{adain} that generates AdaIn parameters which are fed into the generator. The output of the generator includes the image I_1^G and its segmentation mask S_1^G, *i.e.*, black background. See details in Figure 1.

During step 1 we employ *emotion classification* and *dice coefficient* [22] loss functions.

Emotion classification loss. For a given input identity source images $I_1, \ldots I_8$, the driving pose image I_9 and the target augmented emotion label e_a, our goal is to translate $I_1, \ldots I_9$ into an output at step 1 image I_1^G with the identity of $I_1, \ldots I_8$ and the pose of I_9, which is properly classified to the target augmented emotion e_a. To achieve this condition, we use an emotion classifier Cls and impose the emotion classification loss:

$$L_{step1-cls} = _{I_9, e_a} \left[-\log \mathrm{Cls}(e_a | I_1^G) \right]. \quad (1)$$

TABLE I: Quantitative evaluation for the MUG (left) and MEAD (right) datasets. Emotion classification score (ECS), Fréchet inception distance (FID) and Average Content Distance (ACD) metrics are calculated. The proposed method outperforms three approaches in editing emotions in reenactment videos.

(a) MUG dataset

Method	Approach	ECS ↑	FID ↓	ACD ↓
LPD [28]	A1	0.40	22.7	0.12
FOMM [28]	A1	0.21	28.42	0.13
LPD	A2	0.21	41.3	0.12
FOMM	A2	0.60	20.17	0.10
LDP & StarGAN2 [5]	A3	0.80	27.6	0.14
Ours	A4	0.88	19.7	0.10

(b) MEAD dataset

Method	Approach	ECS ↑	FID ↓	ACD ↓
LPD [28]	A1	0.51	47.4	0.18
FOMM [28]	A1	0.21	40.41	0.16
LPD	A2	0.18	46.0	0.18
FOMM	A2	0.54	24.03	0.13
LDP & StarGAN2 [5]	A3	0.16	127	0.83
Ours	A4	0.58	25.5	0.13

Dice coefficient loss. Segmentation maps S_9 and S_1^G are matched with the following loss:

$$L_{step1-dice} = \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2}, \tag{2}$$

where p_i and g_i represent pairs of corresponding pixel values of the predicted at step 1 segmentation S_1^G and the ground truth segmentation S_9, respectively.

Step 2 is similar to Step 1, except the augmented reconstructed image from Step 1 is fed into the Pose encoder as a driving pose image and the original emotion e_9 is fed into MLP_{adain} instead of the emotion e_t.

We expect image I_2^G and segmentation S_2^G synthesized by the generator at step 2 to be as close as possible to the original driving pose image I_9 and its segmentation S_9, respectively. We achieve this with the help of several loss functions.

Dice coefficient and *emotion classification* at step 2 are similar to the corresponding loss functions at step 1:

$$L_{step2-cls} = _{I_9,e_1}\ [-\log \mathrm{Cls}(e_1|I_2^G)]. \tag{3}$$

$$L_{step2-dice} = \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2}, \tag{4}$$

where p_i and g_i represent pairs of corresponding pixel values of the predicted at step 2 segmentation S_2^G and the ground truth segmentation S_9, respectively.

L1 loss is a per-pixel loss function:

$$L_1 = \frac{1}{N}\sum |p_i - g_i|, \tag{5}$$

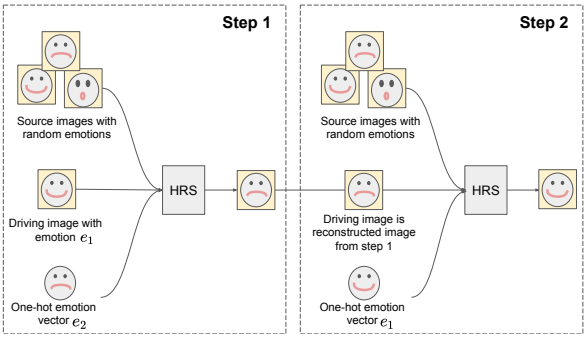


Fig. 4: The training pipeline consists of two steps. During step 1 identity source images, the driving image with the emotion e_1 and the augmented emotion e_2 are fed into a head reenactment system (HRS). During step 2 the inputs into the HRS are identity source images, the reconstructed image from the step 1, the same as in driving image at step 1 emotion e_1 .

where p_i and g_i represent pairs of corresponding pixel values of the predicted segmented image $I_2^G \bigodot S_2^G$ and the ground truth segmented image $I_9 \bigodot S_9$, respectively.

VGG and VGGFace feature losses are L1 losses computed over ReLU activation layers of the VGG-19 model trained for ImageNet classification and the VGGFace model trained for face recognition are used, respectively.

Finally, the objective function to optimize our model is as follows.

$$\begin{aligned} L = \lambda_{step1-cls} L_{step1-cls} + \lambda_{step1-dice} L_{step1-dice} + \\ \lambda_{step2-cls} L_{step2-cls} + \lambda_{step2-dice} L_{step2-dice} + \\ \lambda_{VGG} L_{VGG} + \lambda_{VGGFace} L_{VGGFace} + \lambda_{L_1} L_1 \end{aligned} \tag{6}$$

IV. Experiments

A. Datasets and preprocessing

We evaluate the performance of our proposed method on following two datasets.

- *The MUG Facial Expression Database* [1] consists of image sequences of 50 subjects performing facial expressions including anger, disgust, fear, happiness, neutral, sadness, and surprise.
- *MEAD* [32] is a talking-face video corpus featuring 60 actors and actresses talking with eight different emotions (anger, contempt, disgust, fear, happiness, neutral,

sadness, and surprise) at three different intensity levels. We use the publicly available subset of MEAD that contains 30 subjects and we use front, left and right viewpoints.

In each image, we re-crop the annotated face by capturing its bounding box with the S3FD detector [39] and making that box square by enlarging the smaller side, increasing the bounding box' sides by 80%, while keeping the center, and finally resizing the cropped image to 256×256. Segmentation is obtained by the Graphonomy model [9].

B. Experimental Setup

We split the MUG and MEAD datasets into train and test subsets. MUG train and test sets include 10 and 40 subjects, respectively, MEAD train and test sets include 6 and 24 subjects, respectively. 100 random triples of source subject, target subject and emotion were randomly selected from the test subset for each dataset.

We use the publicly available LPD model, pre-trained on the Voxceleb2 dataset [6]. For each source subject from the list of test triples we fine-tune LPD on the images of this source subject. Next, we fix the parameters of the generator, identity encoder and pose encoder and train $MLP_{identity}$, MLP_{pose} and MLP_{adain}.

Fig. 5: FOMM head reenactment system produces unrealistic results on the MEAD dataset. The chin is shifted on some images with viewpoint from left.

C. Comparison with state-of-the-art

We consider four different approaches aimed at modifying emotions in reenactment videos (see Figure 3). For each approach, we propose state-of-the-art methods to compare our method to.

1) Approach 1 and Approach 2 to modify emotions in reenactment videos: The first approach to change emotion in a reenactment video is to feed images with the desired emotion into the head reenactment system. There are two ways to feed images with the target emotion into the head reenactment system. Figure 3a illustrates the first approach, where a driving pose video with a target emotion is taken as input. This case might be applied, in case it is easy to find a driving video with a target emotion. For example, given that a user seeks to animate a face-image with a video, it is feasible to render a driving video with the desired emotion, based on a single image.

Figure 3b shows the second approach, where identity source images with the target emotion are fed into the head reenactment system. It is reasonable to apply this method, in case that there are limited number of target videos and additionally identity source images with the desired emotion are available. For example, if the use case is to animate an image with a driving video of a famous actor and change the emotion in the reenactment video, it might be challenging to find a video pertained to the actor performing desired facial expressions with the desired emotion. However, it is easier to take photos of oneself with the desired emotion.

We use this approach to change emotions in videos produced by LPD [3] and FOMM [28] head reenactment systems. For the MUG and MEAD datasets we pre-train FOMM on the corresponding train subset.

We show that Approach 1 works for LPD but not for FOMM, and, on the other hand, Approach 2 produces good results with a desired emotion for LPD but not for FOMM (see Figure 2 and Table I). Moreover, Figure 5 shows that FOMM sometimes fail to synthesize images with viewpoint from left and right.

2) Approach 3 to modify emotions in reenactment videos: Figure 3c demonstrates the third approach to modify emotions in reenactment videos. In this setting, a video is synthesized by a head reenactment system and a method for emotion editing is applied to each frame of the synthesized video to change the emotion in each frame. In order to synthesize reenactment videos, we use FOMM [28] and LPD [3]. As the current state-of-the-art method in facial attributes editing, StarGAN2 [5] is taken as our baseline model. For a fair comparison, we use the code released by the authors and train the model on the MUG and MEAD datasets with default hyperparameters.

Figure 2 depicts that this approach shows good results on the MUG dataset, however, it fails on the MEAD dataset.

3) Approach 4 to modify emotions in reenactment videos: The fourth approach aims at feeding the target emotion as a variable into the head reenactment system during image synthesis (see Figure 3d). Our proposed method falls into this class, as the latent space of LPD is changed in accordance to the target emotion vector.

It is possible to apply methods to modify latent space of the pre-trained fixed GAN, *e.g.,* InterfaceGAN, GANSpace, *etc.,* however, these methods are designed to modify the latent space of StyleGAN, BigGAN, other image generators, however not the latent space of head reenactment generators, hence it is not fair to use the hyperparameters, which were proposed for analyzing the latent space of image generators.

D. Quantitative evaluation

Evaluating a GAN *w.r.t.* one criterion does not reliably reveal its overall performance. Therefore, in this work we conduct model evaluation using following three metrics.

- *Emotion classification score*. To consistently evaluate the ability of our model in expression editing, we use a classifier trained to predict emotion and calculate the probability of the target emotion (the higher, the better).
- *Fréchet inception distance (FID)* FID [11] is a metric that calculates the distance between feature vectors calculated for real and generated images.

TABLE II: Ablation study *w.r.t.* MLP modules after identity and pose embeddings. All metrics significantly benefit from the proposed MLP modules.

(a) MUG dataset

Method	ECS ↑	FID ↓	ACD ↓
Ours w/o identity and pose MLPs	0.43	25.1	0.13
Ours w/o identity MLP	0.60	23.0	0.12
Ours w/o pose MLP	0.63	22.6	0.11
Ours with identity and pose MLPs	0.88	19.7	0.10

(b) MEAD dataset

Method	ECS ↑	FID ↓	ACD ↓
Ours w/o identity and pose MLPs	0.48	44.3	0.18
Ours w/o identity MLP	0.52	35.6	0.15
Ours w/o pose MLP	0.54	33.4	0.15
Ours with identity and pose MLPs	0.58	25.5	0.13

- *Average Content Distance (ACD).* ACD [30] measures the L1-distance between embedded features of the input and generated images. A lower value indicates better identity similarity between images before and after editing. We employ the prominent facial recognition network DeepFace [26] to extract face code for each individual and calculate the distance for each expression editing.

We provide comparison results in Tables Ia and Ib for the MUG and MEAD datasets, respectively. We observe that our approach significantly outperforms approaches 1, 2, 3 for the ECS and FID metrics. We have that ACD of our method is the same as approach 2 of FOMM method.

E. Ablation study

To quantify the need of MLPs after identity and pose embeddings, we conduct ablation experiments by removing both these modules and by removing MLP after identity embedding, while using MLP after pose embedding and vice versa. The measured degradation in quality is shown in Tables IIa and IIb for the MUG and MEAD datasets respectively.

V. Conclusions

We present a simple and powerful way to create reenactment videos with desired emotion using a pre-trained fixed GAN. Rather than training a new model, we take existing identity and pose representations and discover techniques for controlling them. In particular, we interpret the face representation learned by Latent Pose Descriptors (LPD) head reenactment system and conduct a study on disentanglement of pose, identity and emotion. By leveraging the semantic knowledge encoded in the latent space, we are able to realistically edit emotions in reenactment videos. We compare presented method with the state-of-art methods for editing emotions in reenactment videos, and show that our method provides the most realistic results, as well as preserves face identity best, while generating videos with the desired emotion.

VI. Acknowledgement

The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

References

[1] N. Aifanti, C. Papachristou, and A. Delopoulos. The mug facial expression database. *11th International Workshop on Image 19 Analysis for Multimedia Interactive Services (WIAMIS)*, 2010.

[2] A. Brock, J. Donahue, and K. Simonyan. Largescale gan training for high fidelity natural image synthesis. *International Conference on Learning Representations (ICLR)*, 2019.

[3] E. Burkov, I. Pasechnik, A. Grigorev, and V. Lempitsky. Neural head reenactment with latent pose descriptors. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha. Stargan v2: Diverse image synthesis for multiple domains. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[6] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. *INTERSPEECH*, 2018.

[7] G. Gafni, J. Thies, M. Zollhofer, and M. Niessner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[8] L. Goetschalckx, A. Andonian, A. Oliva, and P. Isola. Ganalyze: Toward visual definitions of cognitive image properties. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[9] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang, and L. Lin. Graphonomy: Universal human parsing via graph transfer learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[11] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[12] X. Huang and S. Belongie. Arbitrary style transfer in realtime with adaptive instance normalization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[13] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[14] A. Jahanian, L. Chai, and P. Isola. On the "steerability" of generative adversarial networks. *International Conference on Learning Representations (ICLR)*, 2020.

[15] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, X. Cao, and F. Xu. Audio-driven emotional video portraits. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[16] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability,and variation. *International Conference on Learning Representations (ICLR)*, 2018.

[17] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improvingthe image quality of stylegan. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[19] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *International Conference on Machine Learning (ICML)*, 2016.

[20] J. Ling, H. Xue, L. Song, S. Yang, R. Xie, and X. Gu. Toward fine-grained facial expression manipulation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[21] F. Meissen. Emotion-aware facial animation. *Master Thesis*, 2019.

[22] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *International Conference on 3D Vision (3DV)*, 2016.

[23] Y. Mirsky and W. Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 2020.

[24] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[25] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *International Conference on Learning Representations (ICLR)*, 2016.

[26] S. I. Serengil and A. Ozpinar. Lightface: A hybrid deep face recognition framework. *Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2020.

[27] Y. Shen, C. Yang, X. Tang, and B. Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[28] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[29] Y. Tian, J. Ren, M. Chai, K. Olszewski, X. Peng, D. N. Metaxas, and S. Tulyakov. A good image generator is what you need for high-resolution video synthesis. *International Conference on Learning Representations (ICLR)*, 2021.

[30] S. Tulyakov, M. Liu, X. Yang, and J. Kautz. Mocogan: Decomposing motion and content for video generation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[31] A. Voynov and A. Babenko. Unsupervised discovery of interpretable directions in the gan latent space. *International Conference on Machine Learning (ICML)*, 2020.

[32] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[33] Y. Wang, P. Bilinski, F. Bremond, and A. Dantcheva. G3AN: This video does not exist. Disentangling motion and appearance for video generation. *arXiv preprint arXiv:1912.05523*, 2019.

[34] Y. Wang, P. Bilinski, F. F. Bremond, and A. Dantcheva. ImaGINator: Conditional Spatio-Temporal GAN for Video Generation. In *WACV*, 2020.

[35] Y. Wang, F. Bremond, and A. Dantcheva. Inmodegan: Interpretable motion decomposition generative adversarial network for video generation. *arXiv preprint arXiv:2101.03049*, 2021.

[36] Y. Wang, A. Dantcheva, and F. Bremond. From attribute-labels to faces: face generation using a conditional generative adversarial network. In *ECCVW*, 2018.

[37] Z. Wang, Q. She, and T. E. Ward. Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Computing Surveys (CSUR)*, 2020.

[38] C. Yang, Y. Shen, and B. Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis. *International Journal of Computer Vision (ICCV)*, 2020.

[39] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, , and S. Z. Li. S3fd: Single shot scale-invariant face detector. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[40] J. Zhu, T. Park, P. Isola, and A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.