International Journal of Computer Vision (2020) 128:2629–2650
https://doi.org/10.1007/s11263-020-01338-7

SliderGAN: Synthesizing Expressive Face Images by Sliding 3D Blendshape Parameters

Evangelos Ververas[1] · **Stefanos Zafeiriou**[1]

Received: 15 May 2019 / Accepted: 10 May 2020 / Published online: 11 June 2020

Abstract
Image-to-image (i2i) translation is the dense regression problem of learning how to transform an input image into an output using aligned image pairs. Remarkable progress has been made in i2i translation with the advent of deep convolutional neural networks and particular using the learning paradigm of generative adversarial networks (GANs). In the absence of paired images, i2i translation is tackled with one or multiple domain transformations (i.e., CycleGAN, StarGAN etc.). In this paper, we study the problem of image-to-image translation, under a set of continuous parameters that correspond to a model describing a physical process. In particular, we propose the SliderGAN which transforms an input face image into a new one according to the continuous values of a statistical blendshape model of facial motion. We show that it is possible to edit a facial image according to expression and speech blendshapes, using sliders that control the continuous values of the blendshape model. This provides much more flexibility in various tasks, including but not limited to face editing, expression transfer and face neutralisation, comparing to models based on discrete expressions or action units.

Keywords GAN · Image translation · Facial expression synthesis · Speech synthesis · Blendshape models · Action units · 3DMM fitting · Relativistic discriminator · Emotionet · 4DFAB · LRW

1 Introduction

Interactive editing of the expression of a face in an image has countless applications including but not limited to movies post-production, computational photography, face recognition (i.e. expression neutralisation) etc. In computer graphics facial motion editing is a popular field, nevertheless mainly revolves around constructing person-specific models having a lot of training samples (Suwajanakorn et al. 2017). Recently, the advent of machine learning, and especially Deep Convolutional Neural Networks (DCNNs) provide very exciting tools making the community to re-think the problem. In particular, recent advances in Generative Adversarial Networks (GANs) provide very exciting solutions for image-to-image (i2i) translation.

i2i translation, i.e. the problem of learning how to transform aligned image pairs, has attracted a lot of attention during the last few years (Isola et al. 2017; Zhu et al. 2017; Choi et al. 2018). The so-called pix2pix model and alternatives demonstrated excellent results in image completion etc. (Isola et al. 2017). In order to perform i2i translation in absence of image pairs the so-called CycleGAN was proposed, which introduced a cycle-consistency loss (Zhu et al. 2017). CycleGAN could perform i2i translation between two domains only (i.e. in the presence of two discrete labels). The more recent StarGAN (Choi et al. 2018) extended this idea further to accommodate multiple domains (i.e. multiple discrete labels).

StarGAN can be used to transfer an expression to a given facial image by providing the discrete label of the target expression. Hence, it has quite small capabilities in expression editing and arbitrary expression transfer. Over the last few years, quite some deep learning related methodologies have been proposed for transforming facial images (Choi et al. 2018; Wiles et al. 2018; Pumarola et al. 2018). The most closely related work to us is the recent work Pumarola et al.

Communicated by Jun-Yan Zhu, Hongsheng Li, Eli Shechtman, Ming-Yu Liu, Jan Kautz, Antonio Torralba.

✉ Evangelos Ververas
e.ververas16@imperial.ac.uk

Stefanos Zafeiriou
s.zafeiriou@imperial.ac.uk

[1] Department of Computing, Imperial College London, Queens Gate, London SW7 2AZ, UK

(2018) that proposed the GANimation model. GANimation follows the same line of research as StarGAN to translate facial images according to the activation of certain facial Action Units (AUs)[1] and their intensities. Even though AU coding is a quite comprehensive model for describing facial motion, detecting AUs is currently an open problem both in controlled, as well as in unconstrained recording conditions[2] (Benitez-Quiroz et al. 2018, 2017). In particular, in unconstrained conditions the detection accuracy for certain AUs is not high-enough yet (Benitez-Quiroz et al. 2018, 2017), which affects the generation accuracy of GANimation[3]. One of the reasons of the low accuracy of automatic annotation of AUs, is the lack of annotated data and the high cost of annotation which has to be performed by highly trained experts. Finally, even though AUs 10-28 model mouth and lip motion, only 10 of them can be automatically recognized i.e. AUs 10, 12, 14, 15, 17, 20, 23, 25, 26, 28. To make matters worse, the 10 AUs can only be recognized with low accuracy, thus they cannot describe all possible lip motion patterns produced during speech. Hence, GANimation cannot be used in straightforward manner for transferring speech.

In this paper, we are motivated by the recent successes in 3D face reconstruction methodologies from in-the-wild images (Richardson et al. 2017; Tewari et al. 2017; b; Booth et al. 2018, 2017), which make use of a statistical model of 3D facial motion by means of a set of linear blendshapes, and propose a methodology for facial image translation using GANs driven by the continuous parameters of the linear blendshapes. The linear blendshapes can describe both the motion that is produced by expression (Cheng et al. 2018) and/or motion that is produced by speech (Tzirakis et al. 2019). On the contrary, neither discrete emotions nor facial action units can be used to describe the motion produced by speech or the combination of motion from speech and expression. We demonstrate that it is possible to transform a facial image along the continuous axis of individual expression and speech blendshapes (Fig. 1).

Moreover, contrary to StarGAN, which uses discrete labels regarding expression, and GANimation, which utilizes annotations with regards to action units, our methodology does not need any human annotations, as we operate using pseudo-annotations provided by fitting a 3D Morphable Model (3DMM) to images (Booth et al. 2018) (for expression deformations) or by aligning audio signals (Tzirakis et al. 2019) (for speech deformations). Building on the automatic annotation process exploited by SliderGAN, a by-product of our training process is a very robust regression DCNN that estimates the blendshape parameters directly from images. This DCNN is extremely useful for expression and/or speech transfer as it can automatically estimate the blendshape parameters of target images.

A recent approach of efficient GAN optimization which has been used to produce higher quality textures (Wang et al. 2018), is the Relativistic GAN (RGAN) (Jolicoeur-Martineau 2019). RGAN was suggested in order to train the discriminator to simultaneously decrease the probability that real images are real, while increasing the probability that the generated images are real. In our work, we incorporate RGAN in the training process of SliderGAN and demonstrate that it can improve the generator which produces more detailed results in the task of i2i translation for expression and speech synthesis, when compared to training with WGAN-GP. In particular, we employ the Relativistic average GAN (RaGAN) which decides whether an image is relatively more realistic than the others on average, rather than whether it is real or fake. More details, as well as the benefits from this mechanism are presented in Sect. 4.1.

To summurize, the proposed method includes quite a few novelties. First of all, we showcase that SliderGAN is able to synthesize smooth deformations of expression and speech in images by utilizing 3D blendshape models of expression and speech respectively. Moreover, it is the first time to the best of our knowledge that a direct comparison of blendshape and AU coding is presented for the task of expression and speech synthesis. In addition, our approach is annotation-free but offers much better accuracy than AUs-based methods. Furthermore, it is the first time that Relativistic GAN was employed for the task of expression and speech synthesis. We demonstrate in our results that SliderGAN trained with the RaGAN framework (SliderGAN-RaD) benefits towards producing more detailed textures, than when trained with the standard WGAN-GP framework (SliderGAN-WGP). Finally, we enhance the training of our model with synthesized data, leveraging the reconstruction capabilities of statistical shape models.

2 Related Work

2.1 Facial Attribute Editing and Reenactment in Images

Over the past few years, quite some models have been proposed for the task of transforming images and especially facial attributes in images of faces, e.g. expression, pose, hair color, age, gender etc. A rough categorization of them can be made depending on whether they are targeted to single

[1] AUs is a system to taxonomize motion of the human facial muscles (Ekman et al. 2002).

[2] The state-of-the-art AU detection techniques achieve around 50% F1 in EmotioNet challenge and from our experiments OpenFace (Amos et al. 2016) achieves lower than 20-25%

[3] The accuracy of the GANimation model is highly related to both the AU detection, as well as the estimation of their intensity, since the generator is jointly trained and influenced by a network that performs detection and intensity estimation.

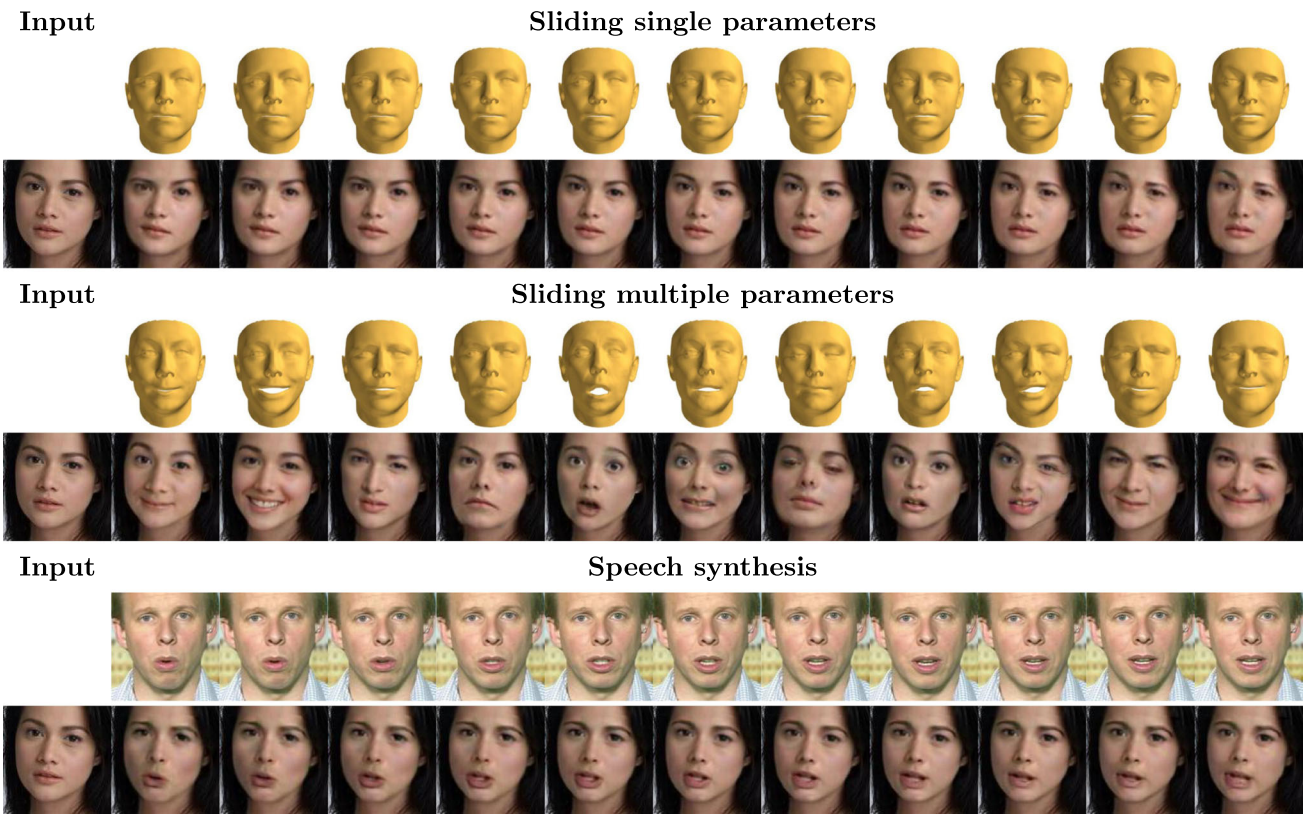

Fig. 1 Expressive faces generated by sliding a single or multiple blendshape parameters in the normalized range $[-1, 1]$. Rows 1 and 3 depict 3D expressive faces generated by a linear blendshape model of natural face motion and a set of expression parameters. The corresponding edited images generated by SliderGAN using the same set of parameters are depicted in rows 2 and 4. As it is observed, the generated images accurately replicate the 3D faces' motion. The robustness of blendshape coding of facial motion allows SliderGAN to perform speech synthesis, as demonstrated in rows 5 (target speech) and 6 (synthesized speech), for which a 3D blendshape model of human speech was utilized

image manipulation or to face reenactment in a sequence of frames.

Single Image Manipulation Targeted to single image manipulation, DIAT (Li et al. 2016) uses an adversarial loss to learn a one directional mapping between images of two domains. CycleGAN (Zhu et al. 2017), also learns a two direction mapping between images of two domains, using a cycle consistency loss for regularization and separate generator and discriminator for each direction of the mapping. IcGAN (Perarnau et al. 2016) is a conditional GAN for image attribute editing, which can handle multiple attributes with one generator. IcGAN learns an inverse mapping from input images to latent vectors and manipulates attributes by changing the condition for fixed latent vectors. Furthermore, SatarGAN (Choi et al. 2018) employs a single generator and a cycle consistency loss to learn one mapping for multiple attributes from multiple databases.

The methods described above are designed to handle single or multiple discrete attributes of images. Instead, Ganimation (Pumarola et al. 2018) builds upon StarGAN and performs expression editing in single images, utilizing facial action unit activations as continuous condition vectors. However, as we demonstrate throughout this paper, blendshape coding is a more robust, efficient and intuitive alternative for conditioning continuous expression editing in images. Moreover, PuppetGAN (Usman et al. 2019) introduced a new approach to training image manipulation systems. In particular, PuppetGAN transforms attributes in images based on examples of how the desired attribute affects the output of a crude simulation (e.g. a 3D model of facial expression). Also, PuppetGAN uses synthetic data to train attribute disentanglement eliminating the need for annotations for the real data, as the disentanglement is extended to the real domain, too. Instead of learning a generator, X2Face (Wiles et al. 2018) changes expression and pose from driving images, pose or audio codes, utilizing an embedding network and a driving network. It is trained with videos requiring no annotations apart from identity, but can be tested on single source and target frames.

Finally, we acknowledge (Geng et al. 2019) which is a concurrent work, very closely related to ours. In this work the authors similarly to us employ blendshape parameters for

expression editing but follow a different approach in image editing, handling 3D texture (UV maps) and shape separately and composing them in a final output image by rendering. This method produces realistic results in expression manipulation, but involves 3DMM fitting and rendering during testing which can be computationally demanding. Nevertheless, it demonstrates the usefulness of blendshapes in the task of automatic face manipulation.

Sequence Manipulation Face reenactment is the process of animating a target face using the face, audio, text, or other codes from a source video to drive the animation. Differently to most of i2i translation methods, face reenactment methods most often require thousands if not millions of frames of the same person for training, testing or both. Targeted to sequence manipulation, Face2Face (Thies et al. 2016) animates facial expression of the target video, based on rendering a face with the requested expression, warping the texture from the available frames of the target video and then blending. Face2Face, also, does not require training. Deep Video Portraits (Kim et al. 2018), produces similar results to Face2Face but animates the whole head and is trained for specific source and target videos, meaning that training has to be repeated when the source or target changes. Other methods drive the animation using audio or text as driving codes (Suwajanakorn et al. 2017; Fried et al. 2019). Finally, Deferred Neural Rendering (DNR) (Thies et al. 2019) is based on learning neural textures, feature maps associated with the scene capturing process, employed by a neural renderer to produce the outputs. DNR is trained for specific source and target videos, too.

2.2 GAN Optimization in i2i Translation Methods

i2i translation models have achieved photo-realistic results by utilizing different GAN optimization methods in literature. pix2pix employed the original GAN optimization technique proposed in Goodfellow et al. (2014). However, the loss function of GAN may lead to the vanishing gradients problem during the learning process. Hence, more effective GAN frameworks emerged that were employed by i2i translation methods. CycleGAN uses LSGAN, which builds upon GAN adopting a least squares loss function for the discriminator. StarGAN and GANimation use WGAN-GP (Gulrajani et al. 201), which enforces gradient clipping as a measure to regularize the discriminator. WGAN-GP, builds upon WGAN (Arjovsky et al. 2017) which minimizes an approximation of the Wasserstein distance to stabilize training of GANs. In this work, we employ the Relativistic GAN (RGAN) (Jolicoeur-Martineau 2019), which decides whether an image is relatively more realistic than the others, rather than whether it is real or fake. RGAN has been proven to enhance the texture quality in i2i translation settings (Wang et al. 2018).

3 Face Deformation Modelling with Blendshapes

3.1 Expression Blendshape Models

Blendshape models are frequently used in computer vision tasks as they constitute an effective parametric approach for modelling facial motion. The localized blendshape model Neumann et al. (2013) proposed a method to localize sparse deformation modes with intuitive visual interpretation. The model was built by sequences of manually collected expressive 3D face meshes. In more detail, a variant of sparse Principal Component Analysis (PCA) was applied to a matrix $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_m] \in \mathbb{R}^{3n \times m}$, which includes m difference vectors $\mathbf{d}_i \in \mathbb{R}^{3n}$, produced by subtracting each expressive mesh from the neutral mesh of each corresponding sequence. Therefore, the sparse blendshape components $\mathbf{C} \in \mathbb{R}^{h \times 1}$ where recovered by the following minimization problem:

$$\text{argmin} \, \|\mathbf{D} - \mathbf{BC}\|_F^2 + \Omega(\mathbf{C}) \quad \text{s.t.} \quad \mathcal{V}(\mathbf{B}), \tag{1}$$

where the constraint $\mathcal{V}$ can either be $\max(|\mathbf{B}_k|) = 1, \ \forall k$ or $\max(\mathbf{B}_k) = 1, \ \mathbf{B} \geq 1, \ \forall k$, with $\mathbf{B}_k \in \mathbb{R}^{3n \times 1}$ denoting the k^{th} component of the sparse weight matrix $\mathbf{B} = [\mathbf{B}_1, \cdots, \mathbf{B}_h]$. According to Neumann et al. (2013), the selection of the constraints mainly controls whether face deformations will take place towards both negative and positive direction of the axes of the model's parameters or not, which is useful for describing shapes like muscle bulges. The regularization of sparse components $\mathbf{C}$ was performed with $\ell 1/\ell 2$ norm (Wright et al. 2009; Bach et al. 2012), while to compute optimal $\mathbf{C}$ and $\mathbf{B}$, an iterative alternating optimization was employed. The exact same approach was employed by Cheng et al. (2018), in the construction of the 4DFAB blendshape model exploited in this work. The 5 most significant deformation components of the 4DFAB expression model are depicted in Fig. 2.

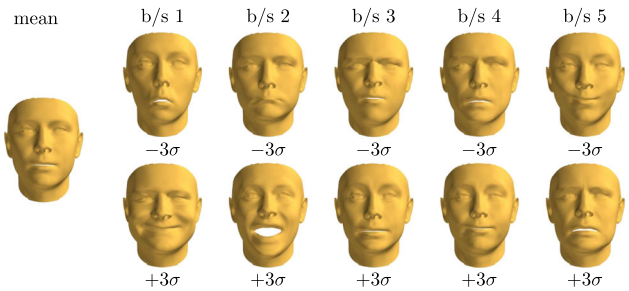

Fig. 2 Visualization of the 5 most significant components of the blendshape model $\mathcal{S}_{exp}$. The 3D faces of this figure have been generated by adding the multiplied components to a mean face

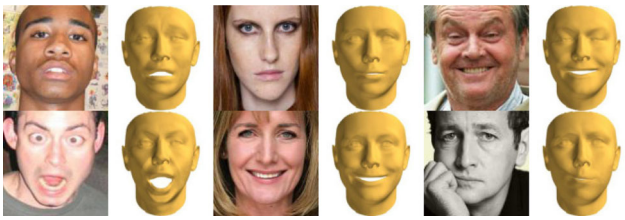

Fig. 3 Examples of the 3D representation of the expression of an image by the model $\mathcal{S}_{exp}$. The 3D faces of this figure have been generated by 3DMM fitting on the corresponding images

3.2 Extraction of Expression Parameters by 3DMM Fitting

3DMM fitting for 3D reconstruction of faces consists of optimizing three parametric models, the *shape*, *texture* and *camera* models, in order to render a 2D instance as close as possible to the input image. To extract the expression parameters from an image we employ 3DMM fitting and particularly the approach proposed in Booth et al. (2018).

In our pipeline we employ the identity variation of LSFM (Booth et al. 2016), which was learned from 10,000 face scans of unique identity, as the shape model to be optimized. To incorporate expression variation in the shape model, we combine LSFM with the 4DFAB blendshape model (Cheng et al. 2018), which was learned from 10,000 face scans of spontaneous and posed expression. The complete shape model can then be expressed as:

$$\begin{aligned}\mathcal{S}(\mathbf{p}_{id}, \mathbf{p}_{exp}) &= \bar{\mathbf{s}} + \mathbf{U}_{s,id}\mathbf{p}_{id} + \mathbf{U}_{s,exp}\mathbf{p}_{exp} \\ &= \bar{\mathbf{s}} + [\mathbf{U}_{s,id}, \mathbf{U}_{s,exp}][\mathbf{p}_{id}^{\top}, \mathbf{p}_{exp}^{\top}]^{\top},\end{aligned} \tag{2}$$

where $\bar{\mathbf{s}}$ is the mean component of 3D shape, $\mathbf{U}_{s,id}$ and $\mathbf{U}_{s,expr}$ are the identity and expression subspaces of LSFM and 4DFAB respectively, and $\mathbf{p}_{id}$ and $\mathbf{p}_{expr}$ are the identity and expression parameters which are used to determine 3D shape instances.

Therefore, by fitting the 3DMM of Booth et al. (2018) in an input image $\mathbf{I}$, we can extract identity and expression parameters $\mathbf{p}_{id}$ and $\mathbf{p}_{exp}$ that instantiate the recovered 3D face mesh $\mathcal{S}(\mathbf{p}_{id}, \mathbf{p}_{exp})$. Based on the independent shape parameters for identity and expression, we exploit parameters $\mathbf{p}_{exp}$ to compose an annotated dataset of images and their corresponding vector of expression parameters $\{\mathbf{I}^i, \mathbf{p}_{exp}^i\}_{i=1}^K$, with no manual annotation cost.

4 Proposed Methodology

In this section we develop the proposed methodology for continuous facial expression editing based on sliding the parameters of a 3D blendshape model.

4.1 Slider-Based Generative Adversarial Network for Continuous Facial Expression and Speech Editing

Problem Definition Let us here first formulate the problem under analysis and then describe our proposed approach to address it. We define an input image $\mathbf{I}_{org} \in \mathbb{R}^{H\times W\times 3}$ which depicts a human face of arbitrary expression. We further assume that any facial deformation or grimace evident in image $\mathbf{I}_{org}$, can be encoded by a parameter vector $\mathbf{p}_{org} = [p_{org,1}, p_{org,2}, ..., p_{org,N}]^{\top}$, of N continuous scalar values $p_{org,i}$, normalized in the range $[-1, 1]$. In addition, the same vector $\mathbf{p}_{org}$ constitutes the parameters of a linear 3D blendshape model $\mathbf{S}_{exp}$ that, as in Fig. 3, instantiate the 3D representation of the facial deformation of image $\mathbf{I}_{org}$ which is given by the expression:

$$\mathcal{S}_{exp}(\mathbf{p}_{org}) = \bar{\mathbf{s}} + \mathbf{U}_{exp}\mathbf{p}_{org}, \tag{3}$$

where $\bar{\mathbf{s}}$ is a mean 3D face component and $\mathbf{U}_{exp}$ the expression eigenbasis of the 3D blendshape model.

Our goal is to develop a generative model which given an input image $\mathbf{I}_{org}$ and a target expression parameter vector $\mathbf{p}_{trg}$, will be able to generate a new version $\mathbf{I}_{gen}$ of the input image with simulated expression given by the 3D expression instance $\mathcal{S}_{exp}(\mathbf{p}_{trg})$.

Attention-Based Generator To address the above challenging problem, we propose to employ a Generative Adversarial Network architecture in order to train a generator network $\mathcal{G}$ that performs translation of an input image $\mathbf{I}_{org}$, conditioned on a vector of 3D blendshape parameters $\mathbf{p}_{trg}$; thus, learning the generator mapping $\mathcal{G}(\mathbf{I}_{org}|\mathbf{p}_{trg}) \rightarrow \mathbf{I}_{gen}$. In addition, to better preserve the content and the colour of the original images we employ an attention mechanism at the output of the generator as in Alami Mejjati et al. (2018) and Pumarola et al. (2018). That is we employ a generator with two parallel output layers, one producing a smooth deformation mask $\mathcal{G}_m \in \mathbb{R}^{H\times W}$ and the other a deformation image $\mathcal{G}_i \in \mathbb{R}^{H\times W\times 3}$. The values of $\mathcal{G}_m$ are restricted in the region $[0, 1]$ by enforcing a sigmoid activation. Then, $\mathcal{G}_m$ and $\mathcal{G}_i$ are combined with the original image $\mathbf{I}_{org}$ to produce the target expression $\mathbf{I}_{gen}$ as:

$$\mathbf{I}_{gen} = \mathcal{G}_m\mathcal{G}_i + (1 - \mathcal{G}_m)\mathbf{I}_{org}. \tag{4}$$

Relativistic Discriminator We employ a discriminator network $\mathcal{D}$ that forces the generator $\mathcal{G}$ to produce realistic images of the desired deformation. Different from the standard discriminator in GANimation which estimates the probability of an image being real, we employ the Relativistic Discriminator (Jolicoeur-Martineau 2019) which estimates the probability of an image being relatively more realistic than a generated one. That is if $\mathcal{D}_{img} = \sigma(\mathcal{C}(\mathbf{I}_{org}))$ is the activation of the standard discriminator, then $\mathcal{D}_{RaD,img} =$

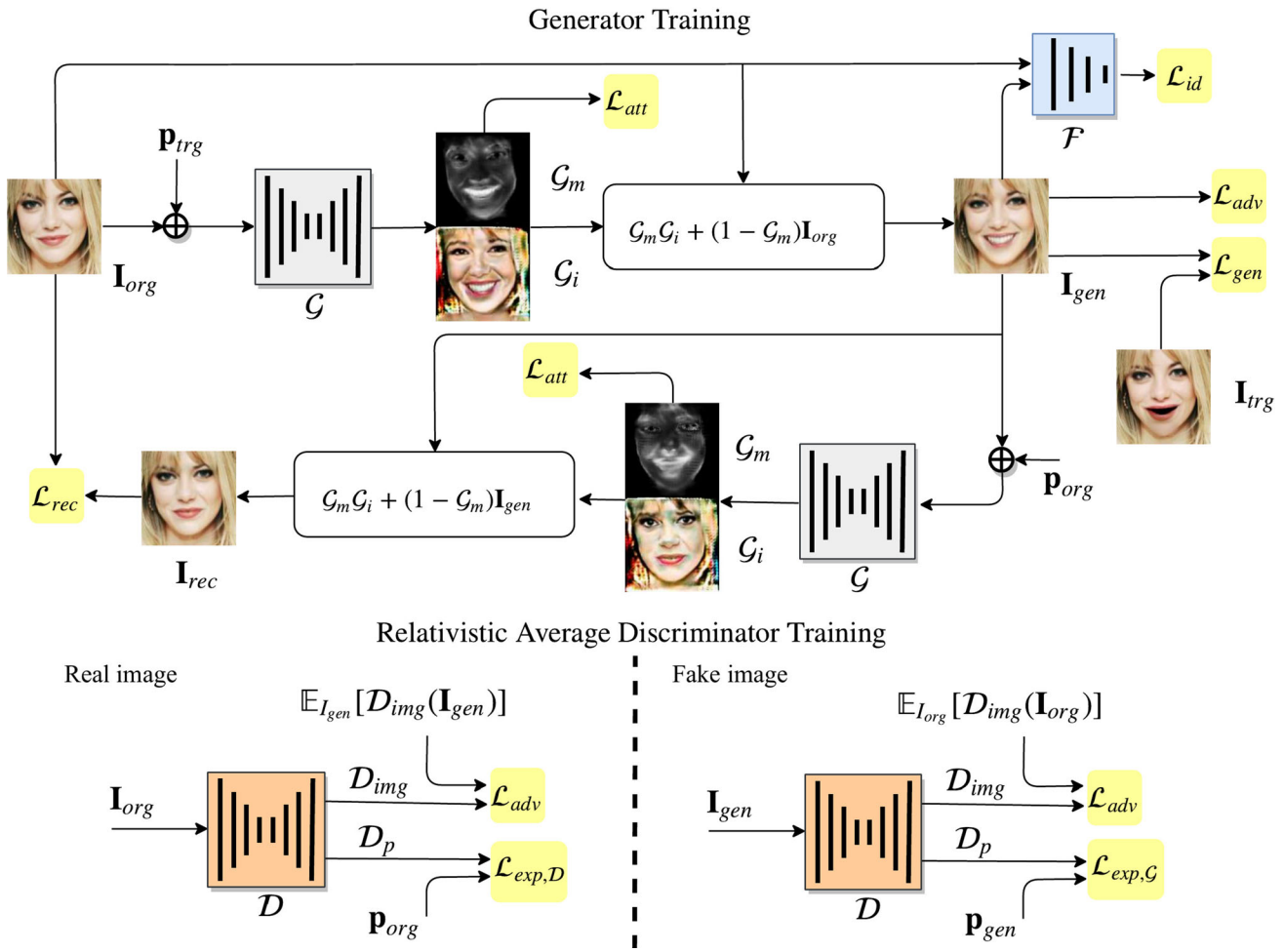

Fig. 4 Synopsis of the modules, losses and the training process of SliderGAN. A attention-based generator $\mathcal{G}$ is trained to generate realistic expressive faces from continuous parameters by employing a set of adversarial, generation, reconstruction, identity and attention losses. The performance of our model is significantly boosted by employing synthetic image pairs through the $\mathcal{L}_{gen}$ loss. Moreover, a relativistic discriminator $\mathcal{D}$ is trained to classify images as relatively more real or fake, as well as to regress expression parameters of the input images in order to increase the generation quality of $\mathcal{G}$

$\sigma(\mathcal{C}(\mathbf{I}_{org}) - \mathcal{C}(\mathbf{I}_{gen}))$ is the activation of the Relativistic Discriminator. Particularly, we employ the Relativistic average Discriminator (RaD) which accounts for all the real and generated data in a mini-batch. Then, the activation of the RaD is:

$$\mathcal{D}_{RaD,img} = \begin{cases} \sigma(\mathcal{C}(\mathbf{I}) - \mathbb{E}_{I_{gen}}[\mathcal{C}(\mathbf{I}_{gen})]), \text{ if } \mathbf{I} \text{ is a real image} \\ \sigma(\mathcal{C}(\mathbf{I}) - \mathbb{E}_{I_{org}}[\mathcal{C}(\mathbf{I}_{org})]), \text{ if } \mathbf{I} \text{ is a generated image} \end{cases} \quad (5)$$

where $\mathbb{E}_{I_{org}}$ and $\mathbb{E}_{I_{gen}}$ define the average activations of all real and generated images in a mini-batch respectively. We further extend $\mathcal{D}$ by adding a regression layer parallel to $\mathcal{D}_{img}$ that estimates a parameter vector $\mathbf{p}_{est}$, to encourage the generator to produce accurate facial expressions, $\mathcal{D}(\mathbf{I}) \rightarrow \mathcal{D}_p(\mathbf{I}) = \mathbf{p}_{est}$. Finally, we aim to boost the ability of $\mathcal{G}$ to maintain face identity between the original and the generated images by incorporating a face recognition module $\mathcal{F}$.

Semi-supervised Training We train our model in a semi-supervised manner with both data with no image pairs of the same person under different expressions $\{\mathbf{I}^i_{org}, \mathbf{p}^i_{org}, \mathbf{p}^i_{trg}\}^K_{i=1}$ and data with image pairs that we automatically generate as described in detail in Sect. 5.1, $\{\mathbf{I}^i_{org}, \mathbf{p}^i_{org}, \mathbf{I}^i_{trg}, \mathbf{p}^i_{trg}\}^L_{i=1}$. The supervised part of training essentially supports SliderGAN being robust on errors of expression parameters extracted from 3DMM fitting. Further discussion on the nature and effect of such errors is included in Sect. 5.6. The modules of our model, as well as the training process of SliderGAN are presented in Fig. 4.

Adversarial Loss To improve the photorealism of the synthesized images we utilize the Wasserstein GAN adversarial

objective with gradient penalty (WGAN-GP) (Gulrajani et al. 201). Therefore, the selected WGAN-GP adversarial objective with RaD is defined as:

$$\begin{aligned}\mathcal{L}_{adv} = &\mathbb{E}_{I_{org}}[\mathcal{D}_{RaD,img}(\mathbf{I}_{org})] \\ &- \mathbb{E}_{I_{org},p_{trg}}[\mathcal{D}_{RaD,img}(\mathcal{G}(\mathbf{I}_{org},\mathbf{p}_{trg}))] \\ &- \lambda_{gp}\mathbb{E}_{I_{gen}}[(\|\nabla_{I_{org}}\mathcal{D}_{img}(\mathbf{I}_{gen})\|_2 - 1)^2].\end{aligned} \tag{6}$$

Different from the standard discriminator, both real and generated images are included in the generator part of the objective of Eq. 6. This allows the generator to benefit from the gradients of both real and fake images, which as we show in the experimental section leads to generated images with sharper edges and more details. This contributes in better representing the distribution of the real data.

Based on the original GAN rational (Goodfellow et al. 2014) and the Relativistic GAN (Jolicoeur-Martineau 2019), our generator $\mathcal{G}$ and discriminator $\mathcal{D}$ are involved in a min-max game, where $\mathcal{G}$ tries to maximize the objective of Eq.(6) by generating realistic images to fool the discriminator, while $\mathcal{D}$ tries to minimize it by correctly classifying real images as more realistic than fake and generated images as less realistic than real.

Expression Loss To make $\mathcal{G}$ consistent in accurately transferring target deformations $\mathcal{S}_{exp}(\mathbf{p}_{trg})$ to the generated images, we consider the discriminator $\mathcal{D}$ to have the role of an inspector. To this end, we back-propagate a mean squared loss between the estimated vector $\mathbf{p}_{est}$ of the regression layer of $\mathcal{D}$ and the actual vector of expression parameters of an image.

We apply the expression loss both on original images and generated ones. Similarly to the classification loss of StarGAN Choi et al. (2018), we construct separate losses for the two cases. For real images $\mathbf{I}_{org}$ we define the loss:

$$\mathcal{L}_{exp,\mathcal{D}} = \frac{1}{N}\|\mathcal{D}(\mathbf{I}_{org}) - \mathbf{p}_{org})\|^2, \tag{7}$$

between the estimated and real expression parameters of $\mathbf{I}_{org}$, while for the generated images we define the loss:

$$\mathcal{L}_{exp,\mathcal{G}} = \frac{1}{N}\|\mathcal{D}(\mathcal{G}(\mathbf{I}_{org},\mathbf{p}_{trg})) - \mathbf{p}_{trg})\|^2, \tag{8}$$

between the estimated and target expression parameters of $\mathbf{I}_{gen} = \mathcal{G}(\mathbf{I}_{org},\mathbf{p}_{trg})$. Consequently, $\mathcal{D}$ minimizes $\mathcal{L}_{exp,\mathcal{D}}$ to accurately regress the expression parameters of real images, while $\mathcal{G}$ minimizes $\mathcal{L}_{exp,\mathcal{G}}$ to generate images with accurate expression according to $\mathcal{D}$.

Image Reconstruction Loss The adversarial and the expression losses of Eq.(6) and Eq.(7), Eq.(8) respectively, would be enough to generate random realistic expressive images which however, would not preserve the contents of the input image $\mathbf{I}_{org}$. To overcome this limitation we admit a cycle consistency loss (Zhu et al. 2017) for our generator $\mathcal{G}$:

$$\mathcal{L}_{rec} = \frac{1}{W \times H}\|\mathbf{I}_{org} - \mathbf{I}_{rec}\|_1, \tag{9}$$

over the vectorized forms of the original image $\mathbf{I}_{org}$ and the reconstructed image $\mathbf{I}_{rec} = \mathcal{G}(\mathcal{G}(\mathbf{I}_{org},\mathbf{p}_{trg}),\mathbf{p}_{org})$. Note that we obtain image $\mathbf{I}_{rec}$ by using the generator twice, first to generate image $\mathbf{I}_{gen} = \mathcal{G}(\mathbf{I}_{org},\mathbf{p}_{trg})$ and then to get the reconstructed $\mathbf{I}_{rec} = \mathcal{G}(\mathbf{I}_{gen},\mathbf{p}_{org})$, conditioning $\mathbf{I}_{gen}$ on the parameters $\mathbf{p}_{org}$ of the original image.

Image Generation Loss To further boost our generator towards accurately editing expression based on a vector of parameters, we introduce image pairs of the form $\{\mathbf{I}^i_{org},\mathbf{p}^i_{org},\mathbf{I}^i_{trg},\mathbf{p}^i_{trg}\}^L_{i=1}$ that we automatically generate from neutral images as described in detail in Sect. 5.1. We exploit the synthetic pairs of images of the same individuals under different expression by introducing an image generation loss:

$$\mathcal{L}_{gen} = \frac{1}{W \times H}\|\mathbf{I}_{trg} - \mathbf{I}_{gen}\|_1, \tag{10}$$

where $\mathbf{I}_{trg}$ and $\mathbf{I}_{gen}$ are images with either neutral or synthetic expression of the same individual. Here, we calculate the $L1$ loss between the synthetic ground truth image $\mathbf{I}_{trg}$ and the generated by $\mathcal{G}$, $\mathbf{I}_{gen}$, aiming to boost our generator to accurately transfer the 3D expression $\mathcal{S}_{exp}(\mathbf{p}_{trg})$ to the edited image.

Identity Loss Image reconstruction loss of Eq. (9), aids to maintain the surroundings between the original and generated images. However, the faces' identity is not always maintained by this loss, as also show by our ablation study in Sect. 5.9. To alleviate this issue, we introduce a face recognition loss adopted from ArcFace (Deng et al. 2018), which models face recognition confidence by an angular distance loss. Particularly, we introduce the loss:

$$\mathcal{L}_{id} = 1 - \cos(\mathbf{e}_{gen},\mathbf{e}_{org}) = 1 - \frac{\|\mathbf{e}_{gen}\|\|\mathbf{e}_{org}\|}{\mathbf{e}^\top_{gen}\mathbf{e}_{org}}, \tag{11}$$

where $\mathbf{e}_{gen} = \mathcal{F}(\mathbf{I}_{gen})$ and $\mathbf{e}_{org} = \mathcal{F}(\mathbf{I}_{org})$ are embeddings of $\mathbf{I}_{gen}$ and $\mathbf{I}_{org}$ respectively, extracted by the face recognition module $\mathcal{F}$. According to ArcFace, face verification confidence is higher as the cosine distance $\cos(\mathbf{e}_{gen},\mathbf{e}_{org})$ grows. During training, $\mathcal{G}$ is optimized to maintain face identity between $\mathbf{I}_{gen}$ and $\mathbf{I}_{org}$ which minimizes Eq.(11).

Attention Mask Loss To encourage the generator to produce sparse attention masks $\mathcal{G}_m$ that focus on the deformation regions and do not saturate to 1, we employ a sparsity loss $\mathcal{L}_{att}$. That is we calculate and minimize the $L1$-norm of the produced masks for both the generated and the reconstructed images, defining the loss as:

$$\mathcal{L}_{att} = \frac{1}{W \times H}\Big(\|\mathcal{G}_m(\mathbf{I}_{org}, \mathbf{p}_{trg})\|_1 + \|\mathcal{G}_m(\mathbf{I}_{gen}, \mathbf{p}_{org})\|_1\Big), \tag{12}$$

Total Training Loss We combine loss functions of Eqs. (6)–(12) to form loss functions $\mathcal{L}_{\mathcal{G}}$ and $\mathcal{L}_{\mathcal{D}}$ for separately training the generator $\mathcal{G}$ and the discriminator $\mathcal{D}$ of our model. We formulate the loss functions as:

$$\mathcal{L}_{\mathcal{G}} = \begin{cases} \mathcal{L}_{adv} + \lambda_{exp}\mathcal{L}_{exp,\mathcal{G}} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{id}\mathcal{L}_{id} + \lambda_{att}\mathcal{L}_{att}, \\ \text{for unpaired data } \{\mathbf{I}^i_{org}, \mathbf{p}^i_{org}, \mathbf{p}^i_{trg}\}_{i=1}^{K} \\ \mathcal{L}_{adv} + \lambda_{exp}\mathcal{L}_{exp,\mathcal{G}} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{gen}\mathcal{L}_{gen} + \lambda_{id}\mathcal{L}_{id}, \\ \quad + \lambda_{att}\mathcal{L}_{att}, \text{for paired data } \{\mathbf{I}^i_{org}, \mathbf{p}^i_{org}, \mathbf{I}^i_{trg}, \mathbf{p}^i_{trg}\}_{i=1}^{L} \end{cases} \tag{13}$$

$$\mathcal{L}_{\mathcal{D}} = -\mathcal{L}_{adv} + \lambda_{exp}\mathcal{L}_{exp,\mathcal{D}}, \tag{14}$$

where λ_{exp}, λ_{rec}, λ_{gen}, λ_{id} and λ_{att} are parameters that regularize the importance of each term in the total loss function. We discuss the choice of those parameters in Sect. 5.2.

As can be noticed in Eq.(13), we employ different loss functions $\mathcal{L}_{\mathcal{G}}$, depending on if the training data are the real data with no image pairs or the synthetic data which include pairs. The only difference is that in the case of paired data we use the additional supervised loss term $\mathcal{L}_{gen}$.

4.2 Implementation Details

Having presented the architecture of our model, here we report further implementation details. For the generator module $\mathcal{G}$ of SliderGAN, we adopted the architecture of CycleGAN (Zhu et al. 2017) as it is proved to generate remarkable results in image-to-iamge translation problems, as for example in StarGAN (Choi et al. 2018). We extended the generator by adding a parallel output layer to accomodate the attention mask mechanism. Moreover, for $\mathcal{D}$ we adopted the architecture of PatchGAN Isola et al. (2017) which produces probability distributions of the multiple image patches to be real or generated, $\mathcal{D}(\mathbf{I}) \rightarrow \mathcal{D}_{img}$. As described in Sect. 4.1, we extended this discriminator architecture by adding a parallel regression layer to estimate continuous expression parameters.

5 Experiments

In this section we present a series of experiments that we conducted in order to evaluate the performance of SliderGAN. First, we describe the datasets we utilized to train and test our model (Sect. 5.1) and provide details on the training setting

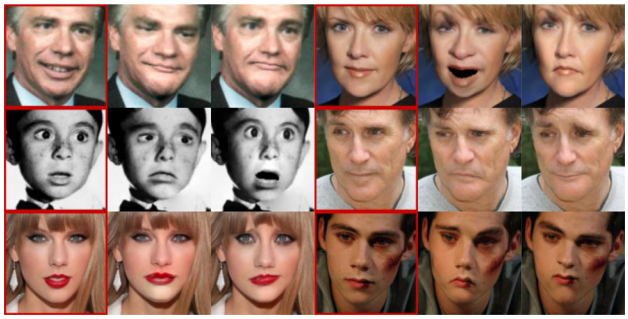

Fig. 5 Synthetic expressive faces, generated by fitting a 3DMM on the original images and rendering back with a randomly sampled expression. The images with a red frame are the original images

for each experiment (Sect. 5.2). Then, we test the ability of SliderGAN to manipulate the expression in images by adjusting a single or multiple parameters of a 3D blendshape model (Sect. 5.3). Moreover, we present our results in direct expression transfer between an input and a target image (Sect. 5.4), as well as in discrete expression synthesis (Sect. 5.5). Next, we test Ganimation on expression editing when trained with blendshape vectors instead of AUs (Sect. 5.6). We examine the ability of SliderGAN to handle face deformations due to speech (Sect. 5.7) and test the regression accuracy of our model's discriminator (Sect. 5.8). We close the experimental section of our work by presenting an ablation study on the contribution of the different loss functions (Sect. 5.9) and a discussion on limitations and failure cases of our technique (Sect. 5.10).

5.1 Datasets

Emotionet For the training and validation phases of our algorithm we utilized a subset of 250,000 images of the EmotioNet database (Benitez-Quiroz et al. 2016), which contains over 1 million images of expression and emotion, accompanied by annotations about facial Action Units. However, SliderGAN is trained with image - blendshape parameters pairs which are not available. Therefore, in order to extract the expression parameters we fit the 3DMM of Booth et al. (2018) on each image of the dataset in use. To ensure the high quality of 3D reconstruction, we employed the LSFM (Booth et al. 2016) identity model concatenated with the expression model of 4DFAB (Cheng et al. 2018. The 4DFAB expression model was built from a collection of over 10,000 expressive face 3D scans of spontaneous and posed expressions, collected from 180 individuals in 4 sessions over the period of 5 years. SliderGAN exploits the scale and representation power of 4DFAB to learn how to realistically edit facial expressions in images. The method described above constitutes a technique to automatically annotate the dataset and eliminates the need of costly manual annotation.

3D Warped Images One crucial problem of training with pseudo-annotations extracted by 3DMM fitting on images, is that the parameter values are not always consistent as small variations in expression can be mistakenly explained by the identity, texture or camera model of the 3DMM. To overcome this limitation, we augmented the training dataset with expressive images that we render and therefore know the exact blendshape parameter values. In more detail, we fit with the same 3DMM 10,000 images of EmotioNet in order to recover the identity and camera models for each image. A 3D texture can also be sampled by projecting the recovered mesh on the original image. Then, we combined the identity meshes with randomly generated expressions from the 4DFAB expression model and rendered back on the original images. Rendering 20 different expressions from each image, we augmented the dataset by 200,000 accurately annotated images. Some of the generated images are displayed in Fig. 5

4DFAB Images A common problem of developing generative models of facial expression is the difficulty in accurately measuring the quality of the generated images. This is mainly due to the lack of databases with images of people of the same identity with arbitrary expressions. To overcome this issue and quantitatively measure the quality of images generated by SliderGAN, as well as compare with the baseline, we created a database with rendered images from 3D meshes and textures of 4DFAB. In more detail, we rendered 100 to 500 images with arbitrary expression from each of the 180 identities and for each of the 4 sessions of 4DFAB, thus rendering 300,000 images in total. To obtain expression parameters for each rendered image, we projected the blendshape model $\mathcal{S}_{exp}$ on each corresponding 3D mesh $\mathbf{S}$ such that the obtained parameters are $\mathbf{p} = \mathbf{U}_{exp}^{\top}(\mathbf{S} - \bar{\mathbf{s}})$.

Lip Reading Words in 3D (LRW-3D) Lip Reading in the Wild (LRW) dataset (Chung and Zisserman 2016) consists of videos of hundreds of speakers including up to 1000 utterances of 500 different words. LRW-3D (Tzirakis et al. 2019) provides speech blendshapes parameters for the frames of LRW, which were recovered by mapping each frame of LRW that correspond to one of the 500 words to instances of a 3D blendshape model of speech, by aligning the audio segments of the LRW videos and those of a 4D speech database. Moreover, to extract expression parameters for each word segment of the videos we applied the 3DMM video fitting algorithm of Booth et al. (2018), which accounts for the temporal dependency between frames. In Sect. 5.7, we utilize the annotations of LRW-3D as well as the expression parameters to perform expression and speech transfer.

5.2 Training Details

In all experiments, we trained our models with images of size 128×128 pixels, aligned to a reference shape of 2D landmarks. As condition vectors we utilized the 30 most significant expression components of 4DFAB and the 10 most significant speech components of LRW-3D (Tzirakis et al. 2019). The later where only used for the combined expression and speech synthesis experiments. We set the batch size to 16 and trained our models for 60 epochs with Kingma and Ba (2014) ($\beta_1 = 0.5$, $\beta_2 = 0.999$). Moreover, we chose loss weights $\lambda_{adv} = 30$, $\lambda_{exp} = 1000$, $\lambda_{rec} = 10$, $\lambda_{gen} = 10$, $\lambda_{id} = 4$ and $\lambda_{att} = 0.3$. Larger values for λ_{id} significantly restrict $\mathcal{G}$, driving it to generate images very close to the original ones with no change in expression. Also, lower values for λ_{att}, lead to mask saturation.

In all our experiments training was performed in two phases over a total of 60 epochs. Particularly, we first trained our models for 20 epochs, utilizing only the generated image pairs of the "3D warped images" database presented in Sect. 5.1. This training phase makes our models robust to parameter errors as further discussed in Sect. 5.6. Then, we proceeded to unsupervised training for another 40 epochs with a dataset of unpaired real images, which we selected depending on the task. In this training phase our models learn to generate the realistic details related to expression and speech. For speech synthesis, we train the model from the beginning with an extended parameter vector of 40 elements, setting the speech parameters to zero for the first phase of training where we train only for expression.

In more detail, the datasets we employed for the second phase of training in our experiments are as follows. We employed:

- **EmotioNet** for our experiments on:
  - 3D model-based expression editing (Sect. 5.3),
  - expression transfer and interpolation on images of Emotionet (Sect. 5.4),
  - discrete expression synthesis (Sect. 5.5),
  - comparing with Ganimation conditioned on blendshape parameters (Sect. 5.6),
  - 3d expression reconstruction (Sect. 5.8),
  - the ablation study (Sect. 5.9),
  - limitations of our model (Sect. 5.10),
- **4DFAB Images** for the experiment on expression transfer and interpolation on images of 4DFAB (Sect. 5.4),
- **LRW-3D** for the combined expression and speech synthesis experiment (Sect. 5.7).

5.3 3D Model-Based Expression Editing

Sliding Single Expression Parameters In this experiment we demonstrate the capability of SliderGAN to edit the facial expression of images when single expression parameters are slid within the normalized range $[-1, 1]$. In Fig. 6 we provide results for 10 levels of activation of single parameters of the

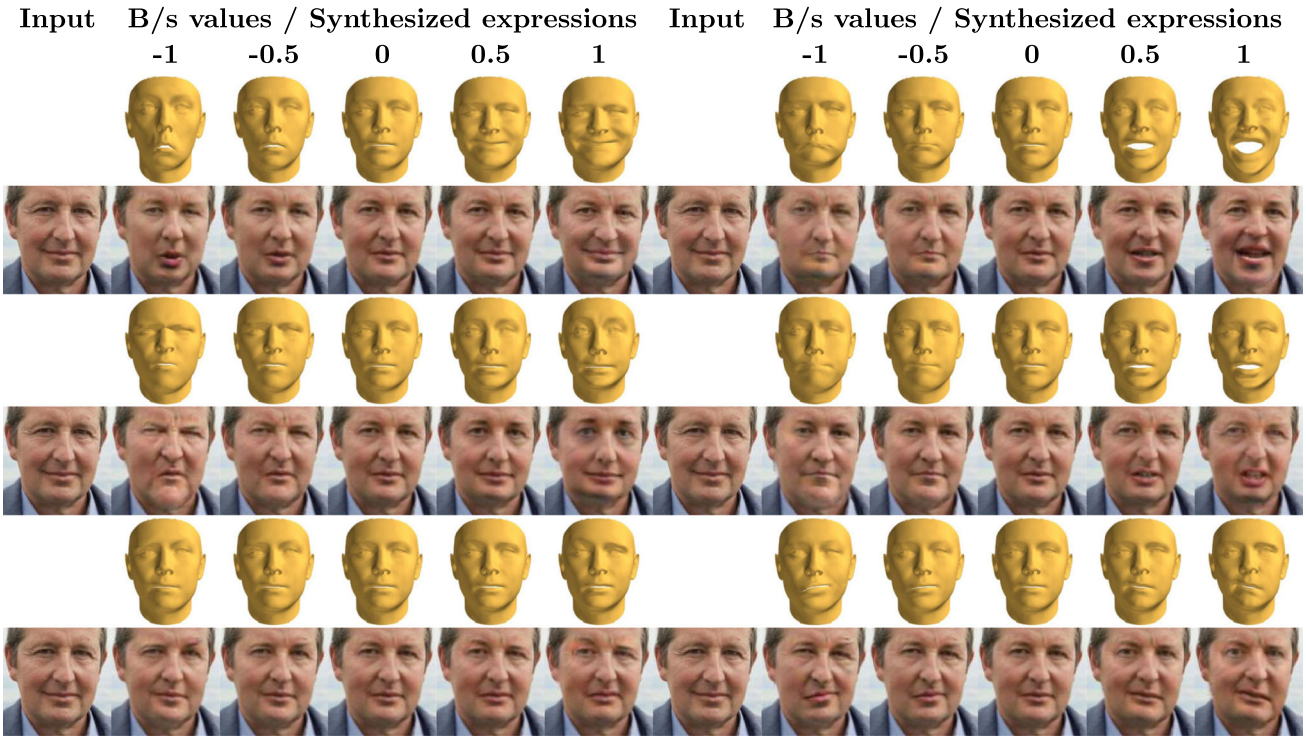

Fig. 6 Expressive faces generated by sliding single blendshape (b/s) parameters in the range $[-1, 1]$. As it is observed, the edited images accurately replicate the 3D faces' motion in the whole range of parameter values

model $(-1, -0.8, -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8, 1)$, while the rest parameters remain zero. As can be observed in Fig. 6, SliderGAN successfully learns to reproduce the behaviour of each blendshape separately, producing realistic facial expressions while adequately maintaining the identity of the input image. Also, the transition between the generated expressions is smooth for successive values of the same parameter and the intensity of the expressions dependent on the magnitude of the parameter value. Note that when the zero vector is applied, SliderGAN produces the neutral expression, whatever the expression of the original image.

Sliding Multiple Expression Parameters The main feature of SliderGAN is its ability to edit facial expressions in images by sliding multiple parameters of the model, similarly to sliding parameters in a blendshape model to generate new expressions of a 3D face mesh. To test this characteristic of our model, we synthesize random expressions by conditioning the generator input on parameter vectors with elements randomly drawn from the standard normal distribution. Note that the model was trained with expression parameters normalized by the square root of the eigenvalues e_i, $i = 1, ..., N$ of the PCA blendshape model. This means that all combinations of expression parameters within the range [-1, 1] correspond to feasible facial expressions.

As illustrated by Fig. 7, SliderGAN is able to synthesize face images with a great variability of expressions, while adequately maintaining identity. The generated expressions accurately resemble the 3D meshes' expressions when the same vector of parameters is used for the blendshape model. This fact makes our model ideal for facial expression editing in images. A target expression can first be chosen by utilizing the ease of perception of 3D visualization of a 3D blendshape model and then, the target parameters can be employed by the generator to edit a face image accordingly.

5.4 Expression Transfer and Interpolation

A by-product of SliderGAN is that the discriminator $\mathcal{D}$ learns to map images to expression parameters $\mathcal{D}_{\mathbf{p}}$ that represent their 3D expression through $\mathcal{S}_{exp}(\mathcal{D}_{\mathbf{p}})$. We capitalize on this fact to perform direct expression transfer and interpolation between images without any annotations about expression. Assuming a source image $\mathbf{I}_{src}$ with expression parameters $\mathbf{p}_{src} = \mathcal{D}_{\mathbf{p}}(\mathbf{I}_{src})$ and a target image $\mathbf{I}_{trg}$ with expression parameters $\mathbf{p}_{trg} = \mathcal{D}_{\mathbf{p}}(\mathbf{I}_{trg})$, we are able to transfer expression $\mathbf{p}_{trg}$ to image $\mathbf{I}_{src}$ by utilising the generator of SliderGAN, such that $\mathbf{I}_{src\rightarrow trg} = \mathcal{G}(\mathbf{I}_{src}|\mathbf{p}_{trg})$. Note that no 3DMM fitting or manual annotation is required to extract the expression parameters and transfer the expression, as this is performed by the trained discriminator.

Additionally, by interpolating the expression parameters of the source and target images, we are able to generate expressive faces that demonstrate a smooth transition from expression $\mathbf{p}_{src}$ to expression $\mathbf{p}_{trg}$. Interpolation of

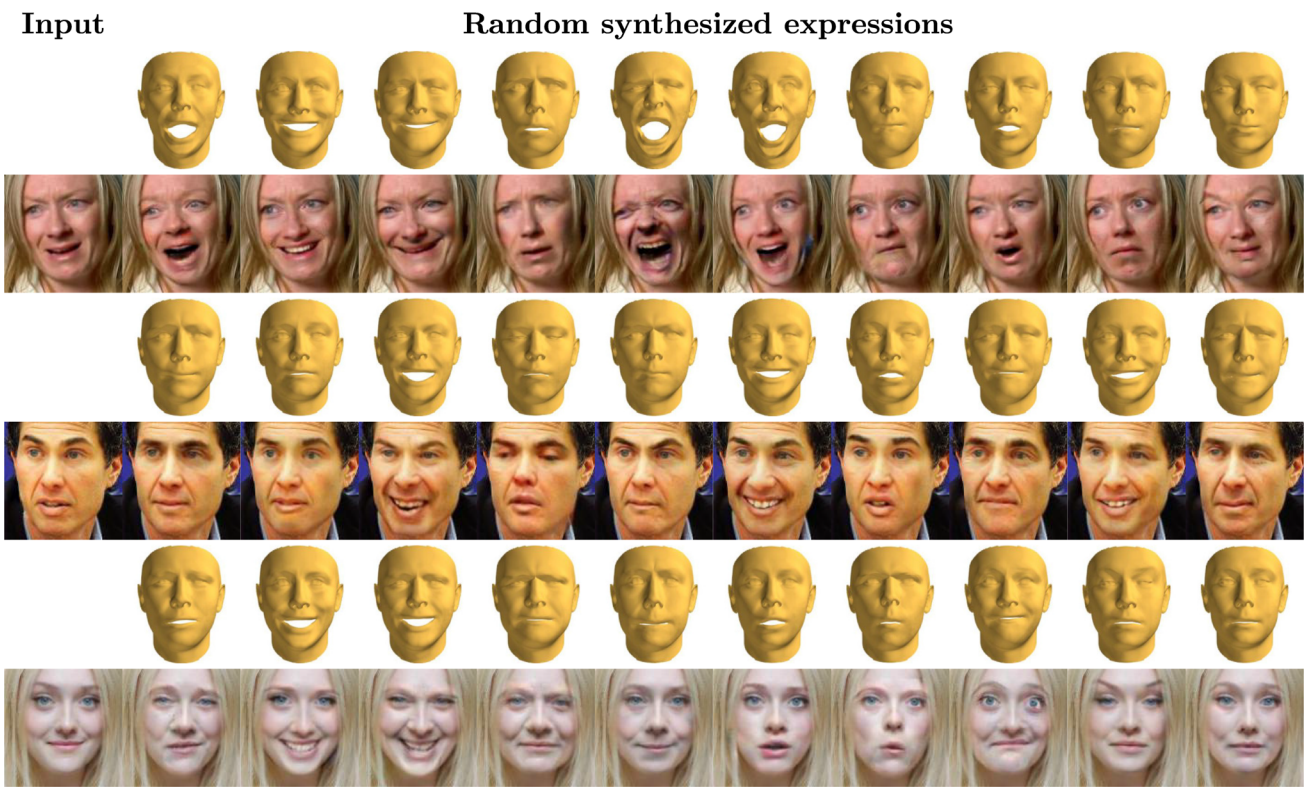

Fig. 7 Expressive faces generated by sliding multiple blendshape (b/s) parameters in the range [−1, 1]. As it is observed, the wide range of the edited images accurately replicate the 3D faces' motion

the expression parameters can be performed by sliding an interpolation factor a within the region [0,1] such that the requested parameters are $\mathbf{p}_{interp} = a\mathbf{p}_{src} + (1-a)\mathbf{p}_{trg}$.

Qualitative Evaluation Results of performing expression transfer and interpolation on images of the 4DFAB rendered database and Emotionet are displayed in Figs. 8 and 9 respectively, where it can be seen that the expressions of the generated images obviously reproduce the target expressions. The smooth transition between expressions $\mathbf{p}_{src}$ and $\mathbf{p}_{trg}$ indicates that SliderGAN successfully learns to map images to expressions across the whole expression parameter space. Also, it is evident that $\mathcal{D}$ accurately regresses the blendshape parameters from images $\mathbf{I}_{trg}$ by observing the recovered 3D faces. The accuracy of the regressed parameters is also examined in Sect. 5.8.

To further validate the quality of our results, we trained GANimation on the same dataset with AU annotations extracted with OpenFace (Amos et al. 2016) as suggested by the authors. We performed expression transfer between images and present results for SliderGAN-RaD, SliderGAN-WGP and GANimation. In Fig. 10, it is obvious that SliderGAN-RaD benefits from the Relativistic GAN training and produces higher quality textures than SliderGAN-WGP, while both SliderGAN implementations better simulate the expressions of the target images than GANimation.

Quantitative Evaluation In this section we provide quantitative evaluation on the performance of SliderGAN on arbitrary expression transfer. We employ the 4DFAB rendered images dataset which allows us to calculate the Image Euclidean Distance (Wang et al. 2005) between ground truth rendered images of 4DFAB and images generated by SliderGAN. Image Euclidea Distance is a robust alternative metric to the standard pixel loss for image distances, which is defined between two RGB images x and y each with $M \times N$ pipxels as:

$$\frac{1}{2\pi}\sum_{i=1}^{MN}\sum_{j=1}^{MN}\exp\{|P_i - P_j|^2/2\}(\|x_i - y_i\|^2)(\|x_j - y_j\|^2) \quad (15)$$

where P_i and P_j are the pixel locations on the 2D image plane and x_i, y_i, x_j, y_j the RGB values of images x and y at the vectorized locations i and j.

We trained SliderGAN with the rendered images from 150 identities of 4DFAB, leaving 30 identities for testing. To allow direct comparison between generated and real images, we randomly created 10,000 pairs of images of the same session and identity (this ensures that the images were rendered with the same camera conditions) from the testing set and performed expression transfer within each pair. To compare our model against the baseline model GANimation, we trained

Fig. 8 Expression interpolation between images of 4DFAB. First, we employ $\mathcal{D}$ to recover the expression parameters from an input and the target images. Then, we capitalize on these parameter vectors to animate the expression of the input image towards multiple targets

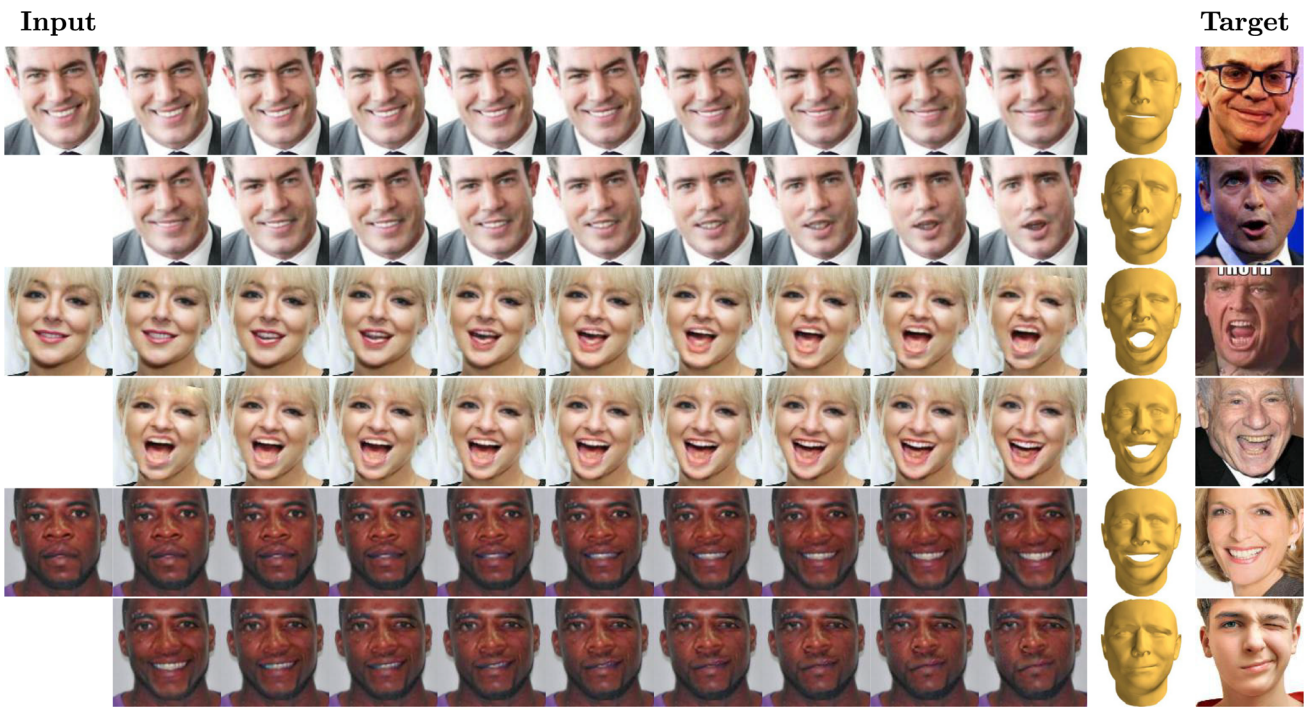

Fig. 9 Expression interpolation between images of Emotionet. First, we employ $\mathcal{D}$ to recover the expression parameters from an input and the target images. Then, we capitalize on these parameter vectors to animate the expression of the input image towards multiple targets

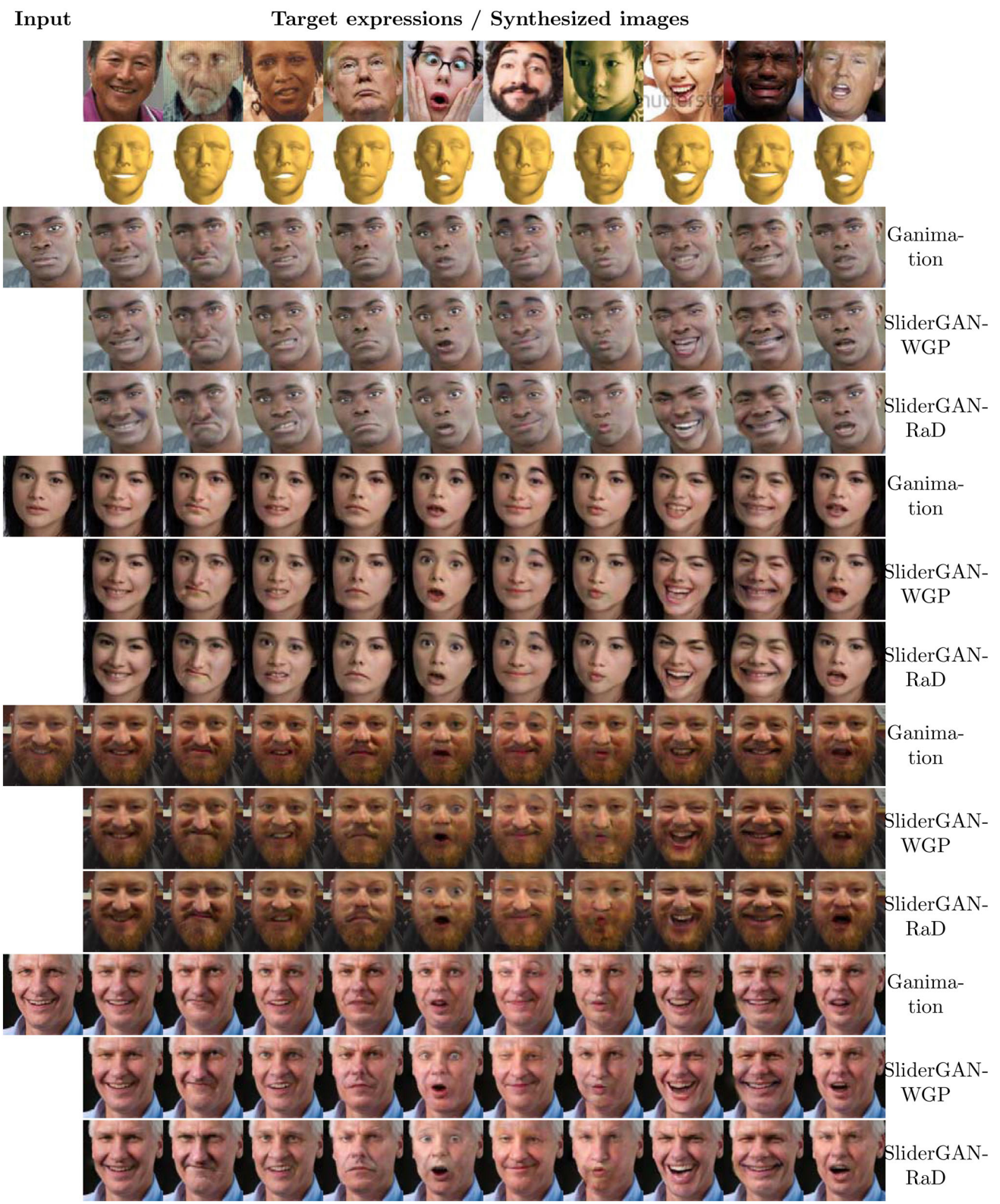

Fig. 10 Expression transfer between images of Emotionet. First, we employ $\mathcal{D}$ to recover expression parameters from the target images. Then, we utilize these parameter vectors to transfer the target expressions to the input images. From the results, SliderGAN-RaD produces higher quality textures than any of the other two methods (mostly evident in the mouth and eyes regions). Moreover, GANimation reproduces the target expressions with lower accuracy. (Please, zoom in the images to notice the differences in texture quality.)

Table 1 TImage Euclidean Distance (IED), calculated between ground truth images of 4DFAB and corresponding generated images by Ganimation (Pumarola et al. 2018), SliderGAN-WGP and SliderGAN-RaD

Method	IED
GANimation Pumarola et al. (2018)	1.04e−02
SliderGAN-WGP	7.932−03
SliderGAN-RaD	$6.84e-03$

Results from SliderGAN-RaD produce the lowest IED between the three methods

and performed the same experiment using GANimation on the same dataset with AUs activations that we obtained with OpenFace. Also, to showcase the benefits of the relativistic discriminator in image quality of the generated images, we repeated the experiment with SliderGAN-WGP. The results are presented in Table 1 where it can be seen that SliderGAN-RaD produces images with the lowest IED.

5.5 Synthesis of Discrete Expressions

Specific combinations of the 3D expression model parameters represent the discrete expressions anger, contempt, fear, disgust, happiness, sadness, surprise and neutral. To directly translate input images into these expressions, we need appropriate blendshape parameter vectors which reproduce the corresponding 3D model instances. Of course, as our condition vectors consist of real numbers, there do not exist unique 3D instances for each expression, but infinitely many with varying intensity.

To extract such parameter vectors we adopted the following approach. First, we manually picked 10 images for each category of the questioned expressions from EmotioNet. Then, we employed $\mathcal{D}$ to estimate parameter vectors for each image, similarly to the expression transfer of Sect. 5.4. We computed the mean vectors for each of the 7 expressions and manually adjusted the values trough visual inspection of the 3D model instances, to create 3D faces that depict the expressions in average intensity (removing any exaggeration or mistakes from the discriminator).

We employ these parameter vectors to synthesize expressive face images of the aforementioned discrete expressions and test our results both qualitatively and quantitatively.

Qualitative Evaluation To evaluate the performance of SliderGAN in this task, we visually compare our results against the results of five baseline models: DIAT (Li et al. 2016), CycleGAN Zhu et al. (2017), IcGAN Perarnau et al. (2016), StarGAN Choi et al. (2018) and GANimation Pumarola et al. (2018). In Fig. 11 it is evident that SliderGAN generates results that resemble the queried expressions while maintaining the original face's identity and resolution. The results are close to those of GANimation, however the Relativistic GAN training of SliderGAN allows for slightly higher quality of images.

The neutral expression can also be synthesized by SliderGAN when all the elements of the target parameter vector are set to 0. In fact, the neutral expression of the 3D blendshape model is also synthesized by the same vector. Results of image neutralization on in-the-wild images of arbitrary expression are presented in Fig. 12, where it can be observed that the neutral expression is generated without significant loss in faces' identity.

Quantitative Evaluation We further evaluate the quality of the generated expressions by performing expression recognition with the off-the-self recognition system (Li et al. 2017). In more detail, we randomly selected 10,000 images from the test set of Emotionet, translated them to each of the discrete expressions anger, disgust, fear, happiness, sadness, surprise, neutral and passed them to the expression recognition network. For comparison, we repeated the same experiment with SliderGAN-WGP and GANimation using the same image set. In Table 2 we report accuracy scores for each expression class separately, as well as the average accuracy score for the three methods. The classification results are similar for the three models, with both implementations of SliderGAN producing slightly higher scores, which denotes that GANimation's results include more failure cases.

5.6 Comparison with Ganimation Conditioned on Blendshape Parameters

It would be reasonable to be assumed that by just substituting AUs with blendshapes, Ganimation could be used to manipulate images based on blendshape conditions. However, this is not the case because Ganimation cannot handle errors of the expression parameters.

3DMM fitting, being an inverse graphics approach to 3D reconstruction, often produces errors related to mistakenly explaining identity and pose of faces as expression and the opposite. For example, a face with a long chin in a slightly side pose might be partially explained by a 3DMM fitting algorithm as a slightly open and shifted mouth or some other similar expression. This is the case for 3D mesh projection (as in the case of recovering parameters from the 4DFAB meshes), too, with which identity can be mistakenly reconstructed to an extend by the linear 3D expression model. This makes the extracted expression parameters to be associated with more attributes of images than only expression.

In the setting of Ganimation, these errors have a negative impact on the robustness and generalisation ability of the model. Particularly, the discriminator becomes dependent on more facial attributes than just expression in regressing the 3DMM parameters. This motivates the generator to reproduce the identity, pose and style of the training images

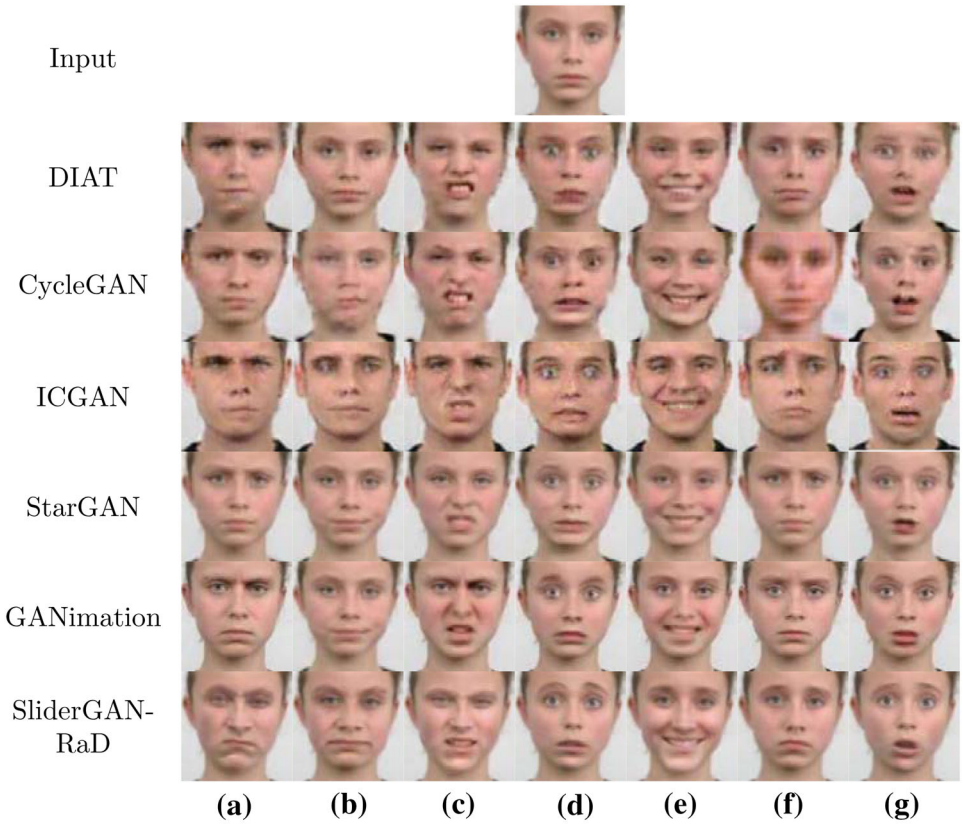

Fig. 11 Generation of the 7 discrete expressions **a** anger, **b** contempt, **c** disgust, **d** fear, **e** happiness, **f** sadness, **g** surprise. By comparing SliderGAN against DIAT (Li et al. 2016), CycleGAN Zhu et al. (2017), IcGAN Perarnau et al. (2016), StarGAN Choi et al. (2018) and GANimation Pumarola et al. (2018) we observe that our model generates results of high texture quality that resemble the queried expressions. The results of the rest of the methods where taken from Pumarola et al. (2018)

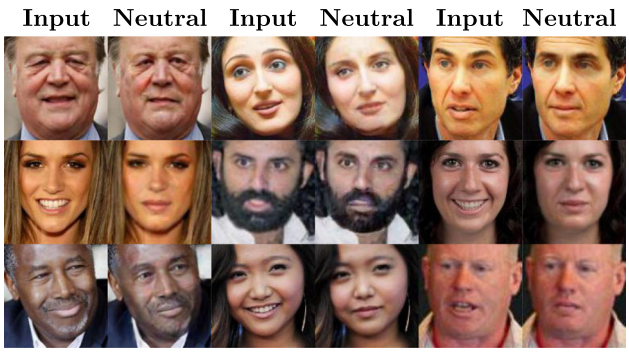

Fig. 12 Neutralization of in-the-wild images of arbitrary expression. The neutralization takes place by setting all blendshape parameter values to zero

rather than only the target expression, as the two modules compete in the min-max optimization problem of the GAN.

This problem is handled in SliderGAN by two of the main contributions of our work. First, the 3D warped images used for the 20 first epochs of the training, help the generator produce expressions consistent with the expression blendshapes, even though realistic texture deformations are missing at this stae (e.g wrinkles when smiling). Second, the face recognition error $\mathcal{L}_{id}$ substantially supports retaining the identity between input and generated images, making SliderGAN robust to the errors of 3DMM fitting. The contribution of both losses in training is further examined in Sect. 5.9. As it can be seen in Fig. 13, the results produced by Ganimation include significant artifacts which are directly related to the identity pose and style of the target images. Contrarily, images generated from SliderGAN do not present such artifacts in most cases and when such artifacts are visible they exist to a considerably lesser extent.

Table 2 Expression recognition results by applying the off-the-self expression recognition system (Li et al. 2017) of images generated by GANimation (Pumarola et al. 2018), SliderGAN-WGP and SliderGAN-RaD

Method	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Neutral	Average
GANimation Pumarola et al. (2018)	0.552	0.446	0.517	0.658	0.632	0.622	0.631	0.579
SliderGAN-WGP	0.550	0.463	0.514	0.762	0.633	0.678	0.702	0.614
SliderGAN-RaD	0.591	0.481	0.531	0.798	0.654	0.689	0.708	0.636

Accuracy scores from both SliderGAN models outperform those of GANimation, while SliderGAN-RaD achieves thehighest accuracy in all epressions

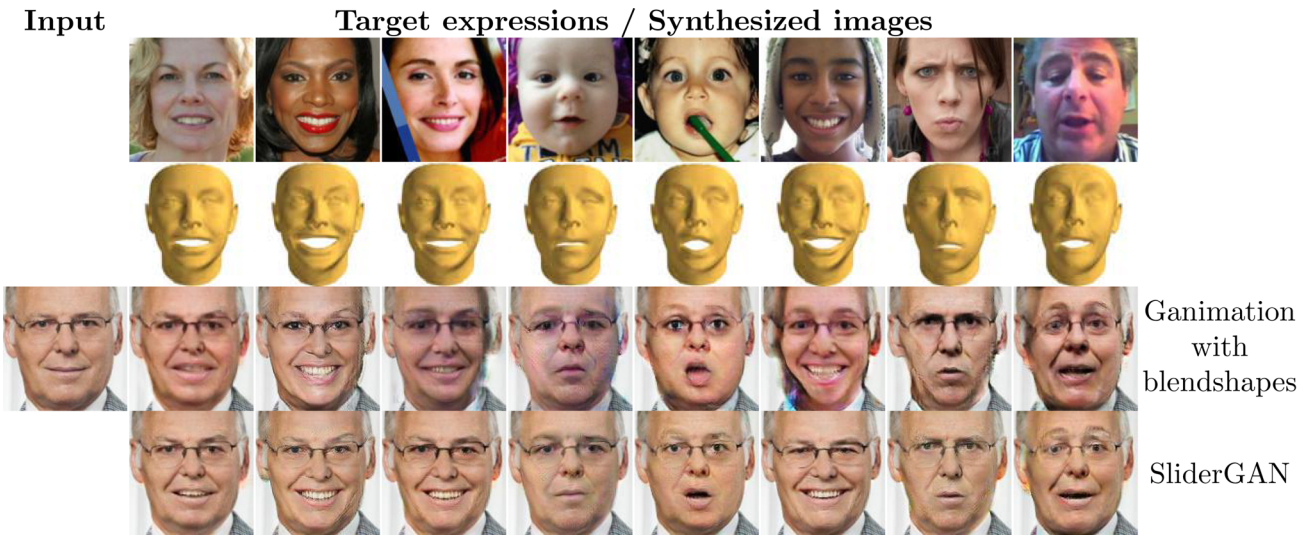

Fig. 13 Evaluation of Ganimation when switching AUs with blenddshapes. Ganimation is not able to handle the errors in expression parameters extracted from 3DMM fitting. The synthesized data, as well as the additional identity loss enables SliderGAN to better translate input images to target expressions

Fig. 14 Combined expression and speech animation from a single input image. We utilize as targets the expression and speech blendshape parameters of consecutive frames of videos of LRW, to synthesize sequences of expression and speech from a single input image

5.7 Combined Expression and Speech Synthesis and Transfer

Blendshape coding of facial deformations allows modelling arbitrary deformations (e.g. deformations due to identity, speech, non-human face morphing etc.) which are not limited to facial expressions, unlike AUs coding which is a system that taxonomizes the human facial muscles (Ekman et al. 2002). Even though AUs 10-28 model mouth and lip motion, not all the details of lip motion that takes place during speech can be captured by these AUs. Moreover, only 10 (AUs 10, 12, 14, 15, 17, 20, 23, 25, 26 and 28) out of these 18 AUs can automatically be recognized, which is achieved only with low accuracy. On the contrary, a blendshape model of the 3D motion of the human mouth and lips would better capture motion during speech, while it would allow the recovery

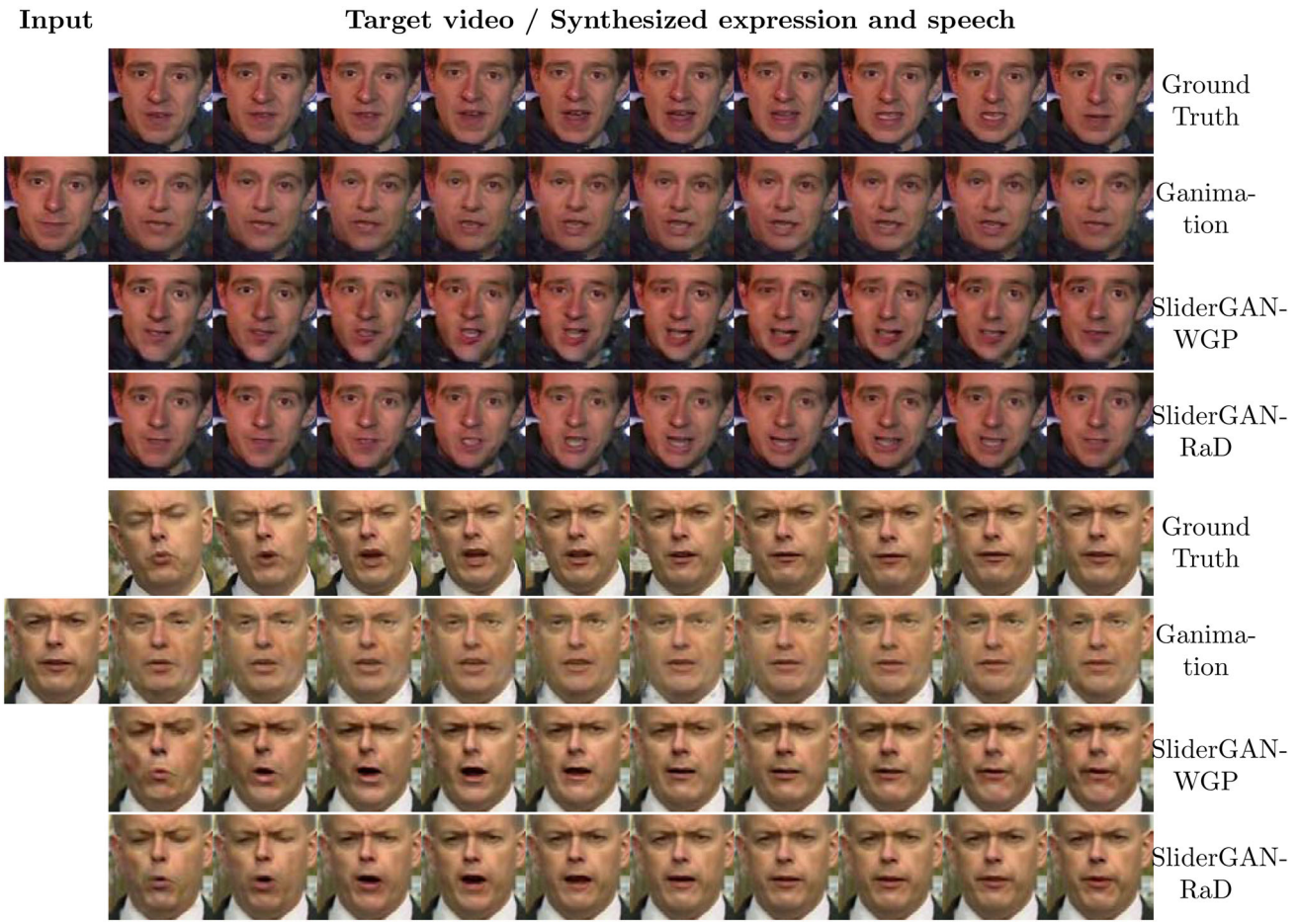

Fig. 15 Comparison of combined expression and speech animation from a single input image between GANimation (Pumarola et al. 2018), SliderGAN-WGP and SliderGAN-RaD. We utilize as targets the expression and speech blendshape parameters of consecutive frames of a video of LRW. Then we reconstruct the expression and speech from a single input frame of the same video. Both SliderGAN implementations reconstruct face motion more accurately than GANimation. Also, the texture quality of the results is higher in SLiderGAN-RaD than in SLiderGAN-WGP as expected. (Please, zoom in the images to notice the differences in texture quality.)

of robust representations from images and videos of human speech.

We capitalize on this fact and employ the mouth and lips blendshape model of Tzirakis et al. (2019), $\mathcal{S}_{speech}(\mathbf{q}) = \bar{\mathbf{s}} + \mathbf{U}_{speech}\mathbf{q}$, to perform speech synthesis from a single image with SliderGAN. Particularly, we employ the LRW-3D database which contains speech blendshape parameters annotations for the 500 words of LRW (Chung and Zisserman 2016), to perform combined expression and speech synthesis and transfer, which we evaluate both qualitatively and quantitatively.

LRW contains videos with both expression and speech. Thus, to completely capture the smooth face motion across frames we employed for each frame 30 expression parameters recovered by 3DMM fitting and 10 speech parameters of LRW-3D which correspond to the 10 most significant components of the 3D speech model $\mathcal{S}_{speech}$. That is we combined the parameters of two separate 3D blendshape models, $\mathcal{S}_{exp}$ and $\mathcal{S}_{speech}$, under our SliderGAN framework by stacking all 40 parameters in a single vector, to train a model which can generate frame sequences where both facial expression and lip/mouth motion varies. Simply stacking the parameters in one vector is a reasonable way to combine them in this case because $\mathcal{S}_{exp}$ and $\mathcal{S}_{speech}$ are linear models and have the same mean component (the LSFM mean face), which means that simple addition of instances of the two models yields possible 3D faces. Also, both include values in the interval $[-1, 1]$. We trained SliderGAN with 180,000 frames of LRW, after training with the warped images, without leveraging the temporal characteristics of the database, that is we shuffled the frames and trained our model with random target vectors to avoid learning person specific deformations.

Qualitative Evaluation Results of performing expression and speech synthesis from a video using a single image are

presented in Fig. 14 where the the parameters and the input frame belong to the same video (ground truth frames are available) and in Fig. 15 where the parameters and the input frame belong to different videos of LRW.

For comparison we trained GANimation on the same dataset with AU activations obtained by OpenFace. As can be seen by Fig. 15, GANimation is not able to accurately simulate the lip motion of the target video. On the contrary, SliderGAN-WGP simulates mouth and lip motion well, but produces textures that look less realistic. SliderGAN-RaD produces higher quality results that look realistic in terms of accurate deformation and texture.

Quantitative Evaluation To measure the performance of our model we employ Image Euclidean Distance (IED) (Wang et al. 2005) to evaluate the results of expression and speech synthesis when the input frame and target parameters belong to the same video sequence. Due to changes in pose in the target videos, we align all target frames with the corresponding output ones before calculating IED. The results are presented in Table 3, where it can be seen that SliderGAN-RaD achieves the lowest error.

Table 3 Image Euclidean Distance (IED), calculated between ground truth images of LRW and corresponding generated images by Ganimation (Pumarola et al. 2018), SliderGAN-WGP and SliderGAN-RaD

Method	**IED**
GANimation Pumarola et al. (2018)	$3.07e-02$
SliderGAN-WGP	$1.14e-02$
SliderGAN-RaD	$9.35e-03$

Results from SliderGAN-RaD produce the lowest IED between the three methods, which indicates the robustness of blendshape coding for speech utlized by SliderGAN

Table 4 Expression representation results on SLiderGAN-RaD (blendshape parameters coding) and Ganimation (AUs activations coding)

	SliderGAN	GANimation Pumarola et al. (2018)
$\frac{1}{N}\sum_{i=1}^{N}\frac{\|\mathbf{p}_{3DMM,i}-\mathbf{p}_{D,i}\|}{\|\mathbf{p}_{3DMM,i}\|}$	0.131	0.427
$\frac{1}{N}\sum_{i=1}^{N}\frac{\|\mathbf{p}_{trg,i}-\mathbf{p}_{D,i}\|}{\|\mathbf{p}_{trg,i}\|}$	0.258	0.513

SliderGAN is capable to accurately and robustly recover expression representations, while GANimation fails to detect AUs activations

5.8 3D Expression Reconstruction

As also described in Sect. 5.4, a by-product of SliderGAN is the discriminator's ability to map images to expression parameters $\mathcal{D}_{\mathbf{p}}$ that reconstruct the 3D expression as $\mathcal{S}_{exp}(\mathcal{D}_{\mathbf{p}})$. We test the accuracy of the regressed parameters on images of Emotionet in two scenarios: a) we calculate the error between parameters recovered by 3DMM fitting and those regressed by $\mathcal{D}$ on the same image (Table 4 row 1) and b) we test the consistency of our model and calculate the error between some target parameters $\mathbf{p}_{trg}$ and those regressed by $\mathcal{D}$ on a manipulated image which was translated to expression $\mathbf{p}_{trg}$ by SliderGAN-RaD (Table 4 row 2).

For comparison, we repeated the same experiment with GANimation for which we calculated the errors in AUs activations. For both experiments we employed 10000 images from our test set. The results demonstrate that the discriminator of SliderGAN-RaD extracts expression parameters from images with high accuracy compared to 3DMM fitting. On the contrary, GANimation's discriminator is less consistent in recovering AU annotations when compared to those of OpenFace. This, also, illustrates that the robustness of blendshape coding of expression over AUs, makes SliderGAN more suitable than GANimation for direct expression transfer.

Nevertheless, as it is reasonable to assume, 3DMM fitting is more stable and accurate in recovering expression parameters from images, than the trained discriminator. The superiority of 3DMM fitting is mostly evident in images with difficult faces and extreme expressions. As it can be seen in Fig. 16, $\mathcal{D}$ produces substantially close 3D reconstruction results to those of 3DMM fitting for the easier image cases, which result in almost identical translated images. Contrarily, the regressed 3D expression reconstructions of $\mathcal{D}$ are obviously less accurate for the harder cases, which affects the quality of expression transfer between input and target images.

Lastly, $\mathcal{D}$ does not achieve state-of-the art results in 3D reconstruction of expression but allows our model to be independent from additional 3DMM fitting during testing, which is clearly an advantage. Alternatives, for more stable expression transfer between images would be to employ different DCNN-based models dedicated to blendshape parameters regression, or 3DMM fitting but with a higher cost in required resources and execution time.

5.9 Ablation Study

In this section we investigate the effect of the different losses that constitute the total loss functions $\mathcal{L}_{\mathcal{G}}$ and $\mathcal{L}_{\mathcal{D}}$ of our algorithm. As discussed in Sect. 4.1, both training in a semi-supervised manner with loss $\mathcal{L}_{gen}$ and employing a face recognition loss $\mathcal{L}_{id}$ between the original and the generated images, contribute significantly in the training process of the generator $\mathcal{G}$. In fact, we only focus on these two terms as they are essential for making SliderGAN robust against errors in expression parameters used as ground truth during training. These errors, caused by limitations of 3DMM fitting, make parameters to be mistakenly associated to more attributes of images than just expression (e.g pose, identity), as further discussed in Sect. 5.6. The rest loss terms of SliderGAN are

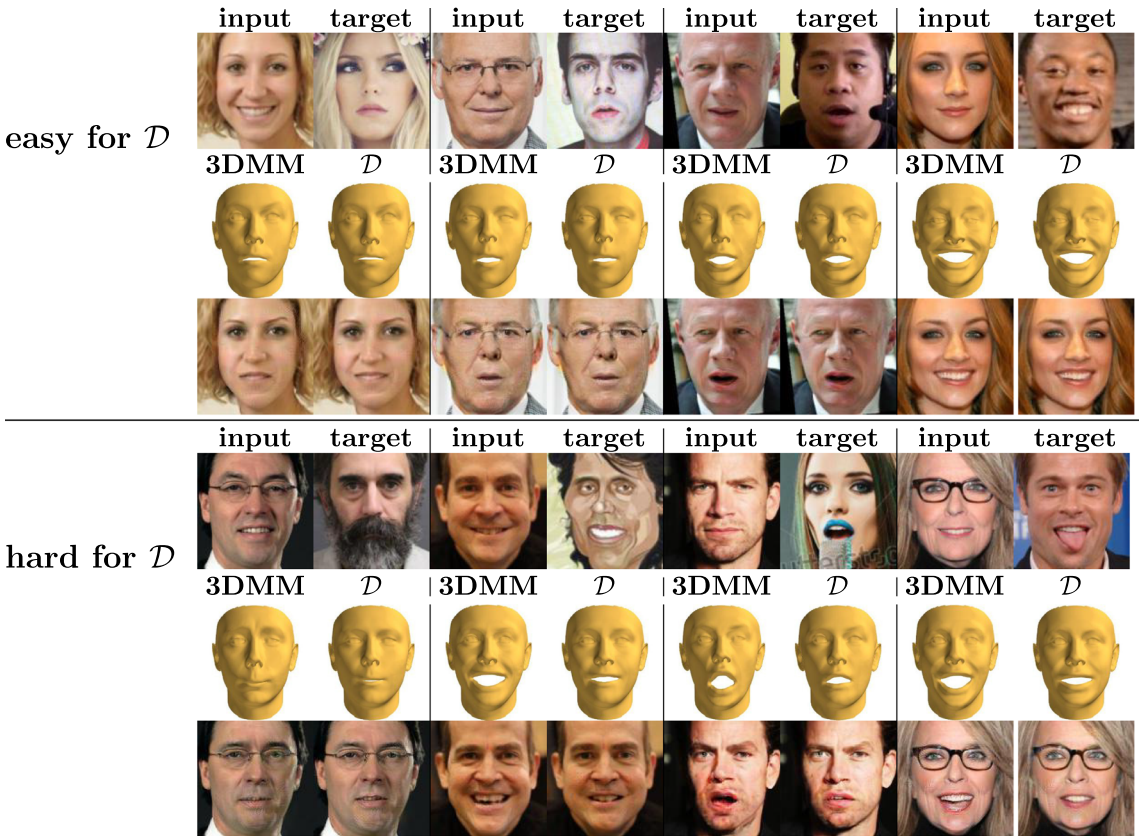

Fig. 16 Comparison of image translation with expression parameters recovered from 3DMM fitting and the discriminator of SliderGAN. $\mathcal{D}$ recovers expressions adequately close to those of 3DMM fitting for most images which are noted as "easy". Then, the image translation in the two cases is almost identical. However, on "hard" cases the accuracy of $\mathcal{D}$ drops, as also does the quality of expression editing

either essential in GAN training ($\mathcal{L}_{adv}$), or common in similar architectures such as the StarGAN and Ganimation ($\mathcal{L}_{rec}$, $\mathcal{L}_{exp,D}$, $\mathcal{L}_{exp,G}$, $\mathcal{L}_{att}$) and thus are not explicitly discussed.

To explore the extend at which these losses affect the performance of $\mathcal{G}$, we consider three different models trained with variations of the loss function of SliderGAN which are: a) $\mathcal{L}_{\mathcal{G}}$ does not include $\mathcal{L}_{id}$, b) $\mathcal{L}_{\mathcal{G}}$ does not include $\mathcal{L}_{gen}$ and c) $\mathcal{L}_{\mathcal{G}}$ does not include both $\mathcal{L}_{id}$ and $\mathcal{L}_{id}$. Fig. 17 depicts results for the same subject generated by the three models as well as SliderGAN. As it can be observed in row "without $\mathcal{L}_{id}$", the absence of $\mathcal{L}_{id}$ results in images that clearly reflect the target expressions, but with changed identity and artifacts. Thus, $\mathcal{L}_{id}$ substantially supports retaining the identity between input and generated images. As it is shown in row "without $\mathcal{L}_{gen}$", training our model utilizing $\mathcal{L}_{id}$ and not $\mathcal{L}_{gen}$ results in images with only slightly changed identity between input and output images, that however reflect other attributes of the target images along with expression such as pose, head shape and color.

When both $\mathcal{L}_{id}$ and $\mathcal{L}_{gen}$ are omitted as in row "without $\mathcal{L}_{id} + \mathcal{L}_{gen}$", both the identity preservation and the expression accuracy decrease drastically. Generally, the GAN loss

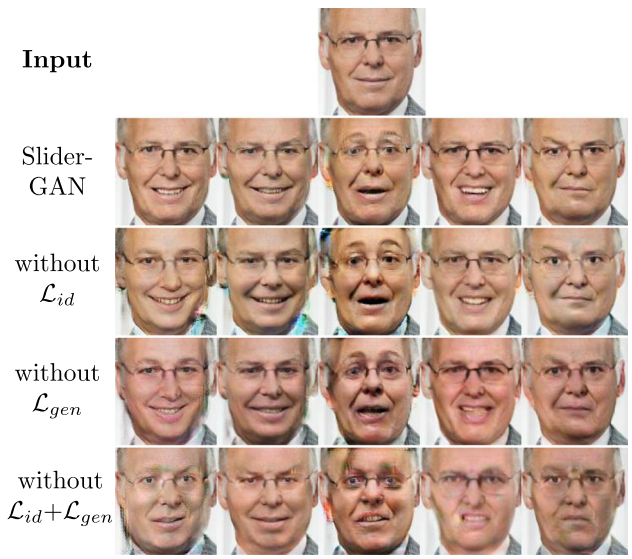

Fig. 17 Results from the ablation study on SliderGAN's loss function components. It is evident that both losses $\mathcal{L}_{id}$ and $\mathcal{L}_{gen}$ have significant impact on the training of the model

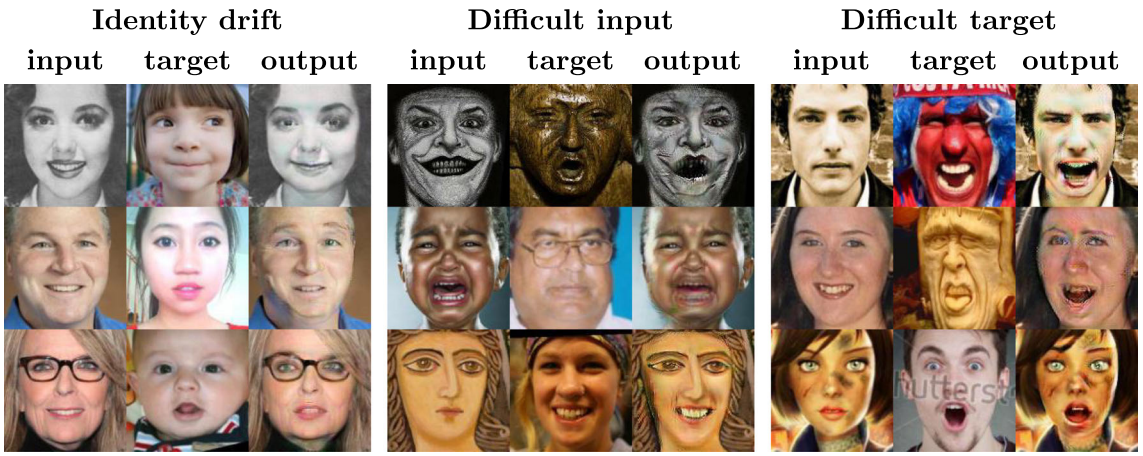

Fig. 18 Limitations of SliderGAN. The main limitations are the identity transfer from target images to the output, the unsuccessful manipulation of non-natural images and the compromised generation of extreme expressions

is responsible for generating realistic images with higher frequency details that an l_1 or l_2 reconstruction loss cannot produce. However, in this case the GAN loss is not enough, because of the inconsistency of expression parameters which makes image generation problematic.

Finally, including both loss functions in training, enables SliderGAN to produce images that preserve all attributes of the input images but expression, which is manipulated according to the target expressions.

5.10 Limitations of SliderGAN

In this section we discuss the main limitations of our proposed model to indicate possible directions for improvement.

One important limitation is that SliderGAN does not always maintain the identity of the input images completely unchanged as can be seen in Fig. 18. This happens mainly, in cases of extreme expressions or expressions with few close samples in the training set of real images. Thus, in those cases SliderGAN over-fits to specific images, reproducing the identity in the generator's output. This could probably be solved if a more balanced database in terms of expressions was employed. It is worth noting that the identities are perfectly maintained in the case of training with 4DFAB, which is a controlled database and includes lots of images for every expression.

Another limitation is generating extreme expressions or manipulating images with extreme expressions. In both cases, images often present a lot of artifacts as shown in Fig. 18. This is because extreme expressions are not well represented in the training dataset and of course, bigger parts of the image have to be edited which makes it a more difficult task for the generator.

Lastly, editing non real faces, such as sketches of faces, faces of character models, faces with makeup etc., most often produces artifacts as shown in Fig. 18, for the same reasons as editing extreme expressions.

6 Conclusion

In this paper, we presented SliderGAN, a new and very flexible way for manipulating the expression (i.e., expression transfer etc.) in facial images driven by a set of statistical blendshapes. To this end, a novel generator based on Deep Convolutional Neural Networks (DCNNs) is proposed, as well as a learning strategy that makes use of adversarial learning. A by-product of the learning process is a very powerful regression network that maps the image into a number of blendshape parameters, which can then be used for conditioning the inputs of the generator.

Acknowledgements E. Ververas was supported by the Teaching Fellowship of Imperial College London. S. Zafeiriou acknowledges funding from the EPSRC Fellowship DEFORM: Large Scale Shape Analysis of Deformable Models of Humans (EP/S010203/1). Additionally, we would like to thank the reviewers for their valuable comments that helped us to improve this paper.

References

Alami Mejjati, Y., Richardt, C., Tompkin, J., Cosker, D., & Kim, K.I. (2018). Unsupervised attention-guided image-to-image translation (pp. 3693–3703).

Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). Openface: A general-purpose face recognition library with mobile applications. Technical report CMU-CS-16-118, CMU School of Computer Science.

Arjovsky, M., Chintala, S., Bottou, L. (2017) Wasserstein generative adversarial networks. In *Proceedings of the 34th international conference on machine learning*, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, pp. 214–223.

Bach, F., Jenatton, R., Mairal, J., & Obozinski, G. (2012). Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, *4*(1), 1–106.

Benitez-Quiroz, C. F., Srinivasan, R., & Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 5562–5570).

Benitez-Quiroz, C. F., Wang, Y., & Martinez, A. M. (2017) Recognition of action units in the wild with deep nets and a new global-local loss. In *ICCV* (pp. 3990–3999).

Benitez-Quiroz, F., Srinivasan, R., & Martinez, A. M. (2018). Discriminant functional learning of color features for the recognition of facial action units and their intensities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*, 1683.

Booth, J., Antonakos, E., Ploumpis, S., Trigeorgis, G., Panagakis, Y., & Zafeiriou, S., et al. (2017). 3d face morphable models "in-the-wild". In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Booth, J., Roussos, A., Ververas, E., Antonakos, E., Poumpis, S., Panagakis, Y., et al. (2018). 3d reconstruction of "in-the-wild" faces in images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*, 2638.

Booth, J., Roussos, A., Zafeiriou, S., Ponniahy, A., & Dunaway, D.: A 3d morphable model learnt from 10,000 faces. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 5543–5552).

Cheng, S., Kotsia, I., Pantic, M., & Zafeiriou, S.: 4dfab: A large scale 4d database for facial expression analysis and biometric applications. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*. Salt Lake City, Utah, US.

Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Chung, J. S., & Zisserman, A.: Lip reading in the wild. In *Asian conference on computer vision*.

Deng, J., Guo, J., & Zafeiriou, S. (2018). Arcface: Additive angular margin loss for deep face recognition. arXiv:1801.07698

Ekman, P. (2002). Facial action coding system (facs). A human face.

Fried, O., Tewari, A., Zollhöfer, M., Finkelstein, A., Shechtman, E., Goldman, D. B., et al. (2019). Text-based editing of talking-head video. *ACM Transactions on Graphics*, *2019*(38), 1–14.

Geng, Z., Cao, C., & Tulyakov, S. (2019) 3d guided fine-grained face manipulation.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing system* (pp. 2672–2680). Red Hook: Curran Associates Inc.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 5767–5777). Red Hook: Curran Associates Inc.

Isola, P., Zhu, J. Y., Zhou, T., Efros, A. A. (204). Image-to-image translation with conditional adversarial networks. In *CVPR*.

Jolicoeur-Martineau, A. (2019). The relativistic discriminator: A key element missing from standard GAN. In *International conference on learning representations*.

Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, N., et al. (2018). Deep video portraits. In *ACM transactions on graphics 2018 (TOG)*.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. CoRR arXiv:1412.6980.

Li, M., Zuo, W., & Zhang, D. (2016). Deep identity-aware transfer of facial attributes. CoRR arXiv:1610.05586.

Li, S., Deng, W., & Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2584–2593). IEEE.

Neumann, T., Varanasi, K., Wenger, S., Wacker, M., Magnor, M., & Theobalt, C. (2013). Sparse localized deformation components. *ACM Transactions on Graphics (TOG)*, *32*(6), 179.

Perarnau, G., van de Weijer, J., Raducanu, B., & Álvarez, J. M.: Invertible conditional gans for image editing. In it CoRR arXiv:1611.06355.

Pumarola, A., Agudo, A., Martinez, A., Sanfeliu, A., & Moreno-Noguer, F. (2018). Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*.

Richardson, E., Sela, M., Or-El, R., & Kimmel, R. (2017). Learning detailed face reconstruction from a single image. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 5553–5562). IEEE.

Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: Learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, *36*(4), 95.

Tewari, A., Zollhöfer, M., Garrido, P., Bernard, F., Kim, H., Pérez, P., & Theobalt, C. (2017). Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz, vol. 2. arXiv preprint arXiv:1712.02859.

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2Face: Real-time face capture and reenactment of RGB videos. In *Proceedings computer vision and pattern recognition (CVPR)* IEEE.

Thies, J., Zollhhfer, M., & Niener, M. (2019). Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics*, *38*, 1–12.

Tran, L., & Liu, X. (2018). Nonlinear 3d face morphable model. arXiv preprint arXiv:1804.03786

Tzirakis, P., Papaioannou, A., Lattas, A., Tarasiou, M., Schuller, B., & Zafeiriou, S.: Synthesising 3d facial motion from in-the-wild speech. arXiv preprint arXiv:1904.07002.

Usman, B., Dufour, N., Saenko, K., & Bregler, C.: Puppetgan: Cross-domain image manipulation by demonstration. In *The IEEE international conference on computer vision (ICCV)*.

Wang, L., Zhang, Y., & Feng, J. (2005). On the euclidean distance of images. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1334–1339.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., & Dong, C., et al. (2018). Enhanced super-resolution generative adversarial networks. In *The European conference on computer vision workshops (ECCVW)*.

Wiles, O., Koepke, A., & Zisserman, A. (2018). X2face: A network for controlling face generation by using images, audio, and pose codes. In *European conference on computer vision*.

Wiles, O., Koepke, A. S., & Zisserman, A. (2018). X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the ECCV*.

Wright, S. J., Nowak, R. D., & Figueiredo, M. A. T. (2009). Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, *57*(7), 2479–2493.

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *2017 IEEE international conference on computer vision (ICCV)*.