Dear Author,

Here are the proofs of your article.

- You can submit your corrections **online**, via **e-mail** or by **fax**.
- For **online** submission please insert your corrections in the online correction form. Always indicate the line number to which the correction refers.
- You can also insert your corrections in the proof PDF and **email** the annotated PDF.
- For fax submission, please ensure that your corrections are clearly legible. Use a fine black pen and write the correction in the margin, not too close to the edge of the page.
- Remember to note the **journal title**, **article number**, and **your name** when sending your response via e-mail or fax.
- **Check** the metadata sheet to make sure that the header information, especially author names and the corresponding affiliations are correctly shown.
- **Check** the questions that may have arisen during copy editing and insert your answers/ corrections.
- **Check** that the text is complete and that all figures, tables and their legends are included. Also check the accuracy of special characters, equations, and electronic supplementary material if applicable. If necessary refer to the *Edited manuscript*.
- The publication of inaccurate data such as dosages and units can have serious consequences. Please take particular care that all such details are correct.
- Please **do not** make changes that involve only matters of style. We have generally introduced forms that follow the journal's style.
  Substantial changes in content, e.g., new results, corrected values, title and authorship are not allowed without the approval of the responsible editor. In such a case, please contact the Editorial Office and return his/her consent together with the proof.
- If we do not receive your corrections **within 48 hours**, we will send you a reminder.
- Your article will be published **Online First** approximately one week after receipt of your corrected proofs. This is the **official first publication** citable with the DOI. **Further changes are, therefore, not possible.**
- The **printed version** will follow in a forthcoming issue.

Please note

After online publication, subscribers (personal/institutional) to this journal will have access to the complete article via the DOI using the URL: http://dx.doi.org/[DOI].
If you would like to know when your article has been published online, take advantage of our free alert service. For registration and further information go to: http://www.link.springer.com.

Due to the electronic nature of the procedure, the manuscript and the original figures will only be returned to you on special request. When you return your corrections, please inform us if you would like to have these documents returned.

Metadata of the article that will be visualized in OnlineFirst

ArticleTitle	GANimation: One-Shot Anatomically Consistent Facial Animation	
Article Sub-Title		
Article CopyRight	Springer Science+Business Media, LLC, part of Springer Nature (This will be the copyright line in the final PDF)	
Journal Name	International Journal of Computer Vision	

Corresponding Author	Family Name	**Pumarola**
	Particle	
	Given Name	**Albert**
	Suffix	
	Division	
	Organization	Institut de Robòtica i Informàtica Industrial, CSIC-UPC
	Address	08028, Barcelona, Spain
	Phone	
	Fax	
	Email	apumarola@iri.upc.edu
	URL	
	ORCID	
Author	Family Name	**Agudo**
	Particle	
	Given Name	**Antonio**
	Suffix	
	Division	
	Organization	Institut de Robòtica i Informàtica Industrial, CSIC-UPC
	Address	08028, Barcelona, Spain
	Phone	
	Fax	
	Email	aagudo@iri.upc.edu
	URL	
	ORCID	
Author	Family Name	**Martinez**
	Particle	
	Given Name	**Aleix M.**
	Suffix	
	Division	
	Organization	The Ohio State University
	Address	Columbus, OH, 43210, USA
	Phone	
	Fax	
	Email	martinez.158@osu.edu
	URL	

	ORCID	
Author	Family Name	**Sanfeliu**
	Particle	
	Given Name	**Alberto**
	Suffix	
	Division	
	Organization	Institut de Robòtica i Informàtica Industrial, CSIC-UPC
	Address	08028, Barcelona, Spain
	Phone	
	Fax	
	Email	sanfeliu@iri.upc.edu
	URL	
	ORCID	
Author	Family Name	**Moreno-Noguer**
	Particle	
	Given Name	**Francesc**
	Suffix	
	Division	
	Organization	Institut de Robòtica i Informàtica Industrial, CSIC-UPC
	Address	08028, Barcelona, Spain
	Phone	
	Fax	
	Email	fmoreno@iri.upc.edu
	URL	
	ORCID	

	Received	31 January 2019
Schedule	Revised	
	Accepted	25 July 2019

Abstract	Recent advances in generative adversarial networks (GANs) have shown impressive results for the task of facial expression synthesis. The most successful architecture is StarGAN (Choi et al. in CVPR, 2018), that conditions GANs' generation process with images of a specific domain, namely a set of images of people sharing the same expression. While effective, this approach can only generate a discrete number of expressions, determined by the content and granularity of the dataset. To address this limitation, in this paper, we introduce a novel GAN conditioning scheme based on action units (AU) annotations, which describes in a continuous manifold the anatomical facial movements defining a human expression. Our approach allows controlling the magnitude of activation of each AU and combining several of them. Additionally, we propose a weakly supervised strategy to train the model, that only requires images annotated with their activated AUs, and exploit a novel self-learned attention mechanism that makes our network robust to changing backgrounds, lighting conditions and occlusions. Extensive evaluation shows that our approach goes beyond competing conditional generators both in the capability to synthesize a much wider range of expressions ruled by anatomically feasible muscle movements, as in the capacity of dealing with images in the wild. The code of this work is publicly available at https://github.com/albertpumarola/GANimation.

Keywords (separated by '-')	GAN - Face animation - Action-unit condition
Footnote Information	Communicated by M. Hebert.

International Journal of Computer Vision
https://doi.org/10.1007/s11263-019-01210-3

GANimation: One-Shot Anatomically Consistent Facial Animation

Albert Pumarola[1] · Antonio Agudo[1] · Aleix M. Martinez[2] · Alberto Sanfeliu[1] · Francesc Moreno-Noguer[1]

Received: 31 January 2019 / Accepted: 25 July 2019

Abstract
Recent advances in generative adversarial networks (GANs) have shown impressive results for the task of facial expression synthesis. The most successful architecture is StarGAN (Choi et al. in CVPR, 2018), that conditions GANs' generation process with images of a specific domain, namely a set of images of people sharing the same expression. While effective, this approach can only generate a discrete number of expressions, determined by the content and granularity of the dataset. To address this limitation, in this paper, we introduce a novel GAN conditioning scheme based on action units (AU) annotations, which describes in a continuous manifold the anatomical facial movements defining a human expression. Our approach allows controlling the magnitude of activation of each AU and combining several of them. Additionally, we propose a weakly supervised strategy to train the model, that only requires images annotated with their activated AUs, and exploit a novel self-learned attention mechanism that makes our network robust to changing backgrounds, lighting conditions and occlusions. Extensive evaluation shows that our approach goes beyond competing conditional generators both in the capability to synthesize a much wider range of expressions ruled by anatomically feasible muscle movements, as in the capacity of dealing with images in the wild. The code of this work is publicly available at https://github.com/albertpumarola/GANimation.

Keywords GAN · Face animation · Action-unit condition

1 Introduction

Being able to automatically and smoothly change the facial expression from a single image would open the door to many new exciting applications in different areas, including the movie industry, photography technologies, fashion and e-commerce business, to name but a few. As generative and adversarial networks have become more prevalent, this task has experienced significant advances, with architectures such as StarGAN (Choi et al. 2018), which is able not only to synthesize novel expressions, but also to change other attributes of the face, such as age, hair color or gender. Despite its generality, StarGAN can only change a particular aspect of a face among a discrete number of attributes defined by the annotation granularity of the dataset. For instance, for the facial expression synthesis task, Choi et al. (2018) is trained on the RaFD (Langner et al. 2010) dataset which has only 8 binary labels for facial expressions, namely sad, neutral, angry, contemptuous, disgusted, surprised, fearful and happy, respectively. The generation possibilities of Choi et al. (2018) are, in this case, limited by these eight expression categories.

Facial expressions, however, are the result of the combined and coordinated action of facial muscles that cannot be categorized in a discrete and low number of classes. Ekman and Friesen (1978) developed the Facial Action Coding System (FACS) for describing facial expressions in terms of the so-called action units (AUs), which are anatomically related to the contractions of specific facial muscles. Although the number of action units is relatively small (30 AUs were found to be anatomically related to the contraction of specific facial muscles), more than 7000 different AU combinations have been observed (Scherer 1982). For example, the facial

Communicated by M. Hebert.

✉ Albert Pumarola
apumarola@iri.upc.edu

Antonio Agudo
aagudo@iri.upc.edu

Aleix M. Martinez
martinez.158@osu.edu

Alberto Sanfeliu
sanfeliu@iri.upc.edu

Francesc Moreno-Noguer
fmoreno@iri.upc.edu

1 Institut de Robòtica i Informàtica Industrial, CSIC-UPC, 08028 Barcelona, Spain

2 The Ohio State University, Columbus, OH 43210, USA

expression for *fear* is generally produced with the following activation state: Inner Brow Raiser (AU1), Outer Brow Raiser (AU2), Brow Lowerer (AU4), Upper Lid Raiser (AU5), Lid Tightener (AU7), Lip Stretcher (AU20) and Jaw Drop (AU26) (Du et al. 2014). Depending on the magnitude of each AU, the expression will transmit the emotion of fear to a greater or lesser extent.

In this paper we aim at building a model for synthetic facial animation with the level of expressiveness of FACS, and being able to generate anatomically-aware expressions in a continuous domain, without the need to pre-compute the position of facial landmarks in the input images (Zafeiriou et al. 2017). For this purpose we leverage on the recent EmotioNet dataset (Benitez-Quiroz et al. 2016), which consists of one million images (we use 200,000 of them) of facial expressions of emotion in the wild annotated with discrete AUs' activation.[1] We build a GAN architecture which, instead of being conditioned with images of a specific domain as in Choi et al. (2018), it is conditioned on a one-dimensional vector indicating the presence/absence and the magnitude of each action unit. We train this architecture in an weakly supervised manner that only requires images with their activated AUs. To circumvent the need for pairs of training images of the same person under different expressions, we split the problem in two main stages. First, we consider an AU-conditioned bidirectional adversarial architecture which, given a single training photo, initially renders a new image under the desired expression. This synthesized image is then rendered-back to the original expression, hence being directly comparable to the input image. We incorporate very recent losses to enforce the photo-realism of the generated image. Additionally, our system also goes beyond state of the art in that it can handle images under changing backgrounds and illumination conditions. We achieve this by means of a self-learned attention layer that focuses the action of the network only in those regions of the image that are relevant to convey the novel expression.

As a result, we build an anatomically coherent facial expression synthesis method, able to render images in a continuous domain, and which can handle images in the wild with complex backgrounds and illumination conditions. As we will show in the results section, it compares favorably to other conditioned-GANs schemes, both in terms of the visual quality of the results, and the possibilities of generation. Figure 1 shows some example of the results we obtain, in which given one input image, we gradually change the magnitude of activation of the AUs used to produce a smile.

This paper is an extended of Pumarola et al. (2018) including a more exhaustive experimental evaluation and ablation studies. We have particularly analyzed the role of the attention mechanism we propose, which is a key ingredient of our architecture, and brings robustness to several artifacts. In this paper, we show that besides yielding robustness to cluttered backgrounds it is also effective to handle partial occlusions of the face. Finally, we also provide a user study to assess the quality of the generated results.

2 Related Work

2.1 Generative adversarial networks

GANs are a powerful class of generative models based on game theory. A typical GAN optimization scheme consists in simultaneously training a generator network to produce realistic fake samples and a discriminator network trained to distinguish between real and fake data. This idea is embedded by the so-called *adversarial loss*. Recent works (Arjovsky et al. 2017; Gulrajani et al. 2017) have shown improved stability relying on the continuous Earth Mover Distance metric (EMD), which we shall use in this paper to train our model. GANs have been shown to produce very realistic images with a high level of detail and have been successfully used for image translation (Isola et al. 2017; Kim et al. 2017; Zhu et al. 2017a), face generation (Karras et al. 2018; Radford et al. 2016), super-resolution imaging (Ledig et al. 2017; Wang et al. 2015), indoor scene modeling (Karras et al. 2018; Wang and Gupta 2016) and human pose editing (Pumarola et al. 2018).

2.2 Conditional GANs

An active area of research consists in designing GAN models that incorporate conditions and constraints into the generation process. Prior studies have explored combining several conditions, such as textual descriptions (Reed et al. 2016; Zhang et al. 2017; Zhu et al. 2017b) and class information (Mirza and Osindero 2014; Odena et al. 2017). Particularly interesting for this work are those methods exploring image-based conditioning as in image super-resolution (Ledig et al. 2017), future frame prediction (Mathieu et al. 2016), image in-painting (Pathak et al. 2016), image-to-image translation (Isola et al. 2017) and multi-target domain transfer (Choi et al. 2018).

2.3 Unpaired image-to-image translation

Similar to our framework, several works have tackled the problem of using unpaired training data. First attempts (Liu et al. 2017) relied on Markov random field priors for Bayesian based generation models, using images from the marginal distributions in individual domains. Others explored enhancing GANS with Variational Auto-Encoder

[1] The dataset was re-annotated with Baltrušaitis et al. (2015) to obtain continuous activation annotations.

Fig. 1 Facial animation from a single image. We propose GANimation, an anatomically coherent approach that is not constrained to a discrete number of expressions and can animate the face in a given image and render novel expressions in a continuum. In these examples, we are given solely the left-most input image $\mathbf{I}_{\mathbf{y}_r}$ (highlighted by a blue square), and the parameter α shown on the top denotes the degree of activation of the target action units involved in a smiling-like expression. Additionally, our system can handle images with complex illumination and non-human skin textures, such as the example in the bottom row

strategies (Kingma and Welling 2014; Liu et al. 2017). Later, several works (Li and Wand 2016; Pathak et al. 2016) have exploited the idea of driving the system to produce mappings transforming the style without altering the original input image content. Our approach is more related to those works exploiting cycle consistency to preserve key attributes between the input and the mapped image, such as CycleGAN (Zhu et al. 2017a), DiscoGAN (Kim et al. 2017) and StarGAN (Choi et al. 2018).

2.4 Face image manipulation

Face generation and editing is a well-studied topic in computer vision and generative models. Most works have tackled the task of attribute editing (Larsen et al. 2016; Perarnau et al. 2016; Shen and Liu 2017) trying to modify attribute categories such as adding glasses, changing color hair, gender swapping and aging. The works that are most related to ours are those synthesizing facial expressions. Early approaches addressed the problem using mass-and-spring models to physically approximate skin and muscle movement (Fischler and Elschlager 1973). The problem with this approach is that it is difficult to generate natural looking facial expressions as there are many subtle skin movements that are difficult to render with simple spring models. More related to current deep learning based methods, Susskind et al. (2008) leveraged on a deep belief network to generate new facial expression given a personal identity and the desired facial action unit. The results, however, revealed a lack of realism.

Other strategies constrain the set of possible expressions to those that can be generated using low-rank 3D Morphable Models (3DMMs). Early approaches along this line (Blanz and Vetter 1999; Yu et al. 2012) generated novel expressions by adjusting the initially estimated 3DMM parameters of a registered face. While simple, this approach produced strong image artifacts and could not convey shading and illumination effects. More recently, Nagano et al. (2018) and Thies et al. (2016) achieved impressive facial reenactment of a monocular video sequence but still required explicit estimation of the 3D geometry of the face, which can be very challenging for images and videos in the wild. Kim et al. (2018) improved the photo-realism by extending these geometry based on methods to pure data-driven deep learning techniques.

Journal: 11263 MS: 1210 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2019/8/10 Pages: 16 Layout: Large

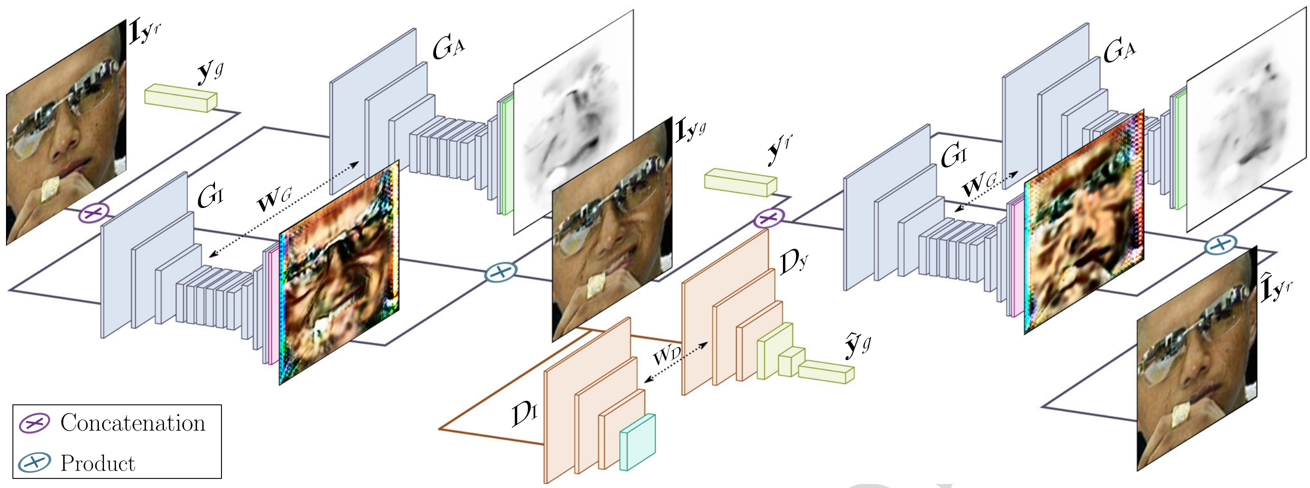

Fig. 2 Overview of our approach to generate photo-realistic conditioned images. The architecture of GANimation consists of three main blocks: a generator G to regress attention and color masks; a critic D_{I} to evaluate the quality of the generated image and its photo-realism; and finally, an expression estimator D_{y} to penalize differences between the desired conditioning expression $\mathbf{y}_g$ and its fulfillment $\hat{\mathbf{y}}_g$. It is worth noting that our scheme does not require supervision, i.e. neither pairs of images of the same person under different expressions, nor the target image $\mathbf{I}_{\mathbf{y}_g}$ are assumed to be known

Suwajanakorn et al. (2017) designed a deep network to synthesize novel and high quality expressions. Synthetic results were almost non-distinguishable from real videos, although the system was only suitable for a single actor for which the system had been trained. Subsequent approaches have focused on more generic alternatives, such as face editing for eyes inpainting (Dolhansky and Canton Ferrer 2018) and speech face animation (Zhou et al. 2019; Song et al. 2018; Vougioukas et al. 2018). These works, however, require large amounts of audio–visual clips (recall that GANimation is trained using sparse images).

Probably the works most closely related to ours are Choi et al. (2018), Li et al. (2016) and Odena et al. (2017), in the sense that they train highly complex convolutional networks able to handle images in the wild. These approaches, however, have been conditioned on discrete emotion categories (e.g. happy, neutral, and sad). Instead, our model resumes the idea of modeling skin and muscles, but we integrate it in modern deep learning machinery. More specifically, we learn a GAN model conditioned on a continuous embedding of muscle movements, allowing to generate a large range of anatomically feasible face expressions as well as smooth facial movement transitions in video sequences.

3 Problem Formulation

Let us define an input RGB image as $\mathbf{I}_{\mathbf{y}_r} \in \mathbb{R}^{H \times W \times 3}$, which represents the cropped face of a subject under an arbitrary expression. Every gesture expression is encoded by means of a set of N action units $\mathbf{y}_r = (y_1, \ldots, y_N)^\top$, where each y_n denotes a normalized value between 0 and 1 to module the magnitude of the nth action unit. This type of continuous representation is a key ingredient of our design, as a natural interpolation can be done between different expressions, allowing to render a wide range of realistic and smoothly changing facial expressions.

Our aim is to learn a mapping $\mathcal{M}$ to translate $\mathbf{I}_{\mathbf{y}_r}$ into an output image $\mathbf{I}_{\mathbf{y}_g}$ conditioned on an action-unit target $\mathbf{y}_g$, i.e. we seek to estimate the mapping $\mathcal{M} : (\mathbf{I}_{\mathbf{y}_r}, \mathbf{y}_g) \rightarrow \mathbf{I}_{\mathbf{y}_g}$. To this end, we propose to train $\mathcal{M}$ in a weakly supervised manner, using M training triplets $\{\mathbf{I}^m_{\mathbf{y}_r}, \mathbf{y}^m_r, \mathbf{y}^m_g\}^M_{m=1}$, where the target vectors $\mathbf{y}^m_g$ are randomly generated. Importantly, we neither require pairs of images of the same subject under different expressions, nor the expected target image $\mathbf{I}_{\mathbf{y}_g}$ to be known.

4 Our Approach

This section describes our novel approach to generate photo-realistic conditioned images, which, as shown in Fig. 2, consists of two main modules. On the one hand, a generator $G(\mathbf{I}_{\mathbf{y}_r}|\mathbf{y}_g)$ is trained to realistically transform the facial expression in image $\mathbf{I}_{\mathbf{y}_r}$ to the desired $\mathbf{y}_g$. Note that G is applied twice, first to map the input image $\mathbf{I}_{\mathbf{y}_r} \rightarrow \mathbf{I}_{\mathbf{y}_g}$, and then to render it back $\mathbf{I}_{\mathbf{y}_g} \rightarrow \hat{\mathbf{I}}_{\mathbf{y}_r}$. On the other hand, we use a WGAN-GP (Gulrajani et al. 2017) based critic $D_{\text{I}}(\mathbf{I}_{\mathbf{y}_g})$ to evaluate the quality of the generated image and an expression estimator $D_{\text{y}}(\mathbf{I}_{\mathbf{y}_g})$ to penalize differences between the desired and generated expression. We next describe in detail each one of these blocks.

Journal: 11263 MS: 1210 ☐TYPESET ☐DISK ☐LE ☐CP Disp.:2019/8/10 Pages: 16 Layout: Large

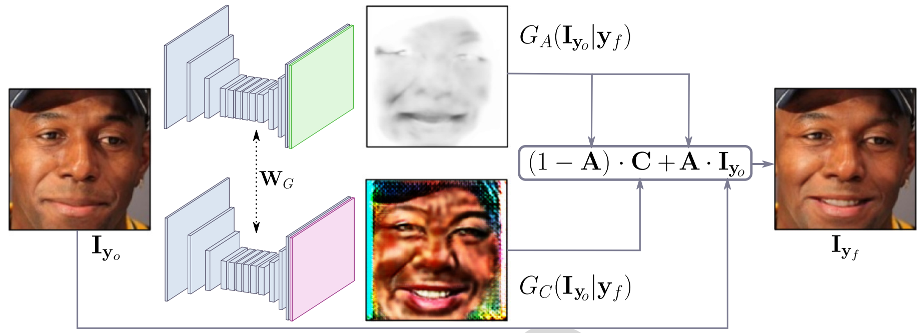

Fig. 3 Attention-based generator. Given an input image and the target expression, the generator regresses an attention mask $\mathbf{A}$ and an RGB color transformation $\mathbf{C}$ over the entire image. The attention mask defines a per pixel intensity specifying to which extend each pixel of the original image will contribute in the final rendered image

4.1 Network Architecture

4.1.1 Generator

Let G be the generator block. Since it will be applied bidirectionally (i.e. to map input image to desired expression and vice-versa) in the following discussion we use subscripts o and f to indicate *origin* and *final*.

Given the image $\mathbf{I}_{\mathbf{y}_o} \in \mathbb{R}^{H\times W\times 3}$ and the N-vector $\mathbf{y}_f$ encoding the desired expression, we form the input of the generator as a concatenation $(\mathbf{I}_{\mathbf{y}_o}, \mathbf{y}_o) \in \mathbb{R}^{H\times W\times (N+3)}$, where $\mathbf{y}_o$ has been represented as N arrays of size $H \times W$.

One key ingredient of our system is to make G focus only on those regions of the image that are responsible of synthesizing the novel expression and keep the rest elements of the image such as hair, glasses, hats or jewellery untouched. For this purpose, we have embedded an attention mechanism into the generator. Concretely, instead of regressing a full image, our generator outputs two masks, a color mask $\mathbf{C}$ and an attention mask $\mathbf{A}$. The final image can be obtained as:

$$\mathbf{I}_{\mathbf{y}_f} = (1-\mathbf{A})\cdot\mathbf{C} + \mathbf{A}\cdot\mathbf{I}_{\mathbf{y}_o}, \tag{1}$$

where $\mathbf{A} = G_A(\mathbf{I}_{\mathbf{y}_o}|\mathbf{y}_f) \in [0,1]^{H\times W}$ and $\mathbf{C} = G_C(\mathbf{I}_{\mathbf{y}_o}|\mathbf{y}_f) \in \mathbb{R}^{H\times W\times 3}$. The mask $\mathbf{A}$ indicates to which extend each pixel of $\mathbf{C}$ contributes to the output image $\mathbf{I}_{\mathbf{y}_f}$. In this way, the generator does not need to render static elements, and can focus exclusively on the pixels defining the facial movements, leading to sharper and more realistic synthetic images. This process is depicted in Fig. 3.

4.1.2 Conditional critic

This block is a network trained to evaluate the generated images in terms of their photo-realism. The structure of $D_\mathrm{I}(\mathbf{I})$ resembles that of the PatchGan (Isola et al. 2017) network mapping from the input image $\mathbf{I}$ to a matrix $\mathbf{Y}_\mathrm{I} \in \mathbb{R}^{H/2^6\times W/2^6}$, where $\mathbf{Y}_\mathrm{I}[i,j]$ is used as a partial function to compute the EMD between the distributions of real image patches and the overlapping patch ij of the generated image.

4.1.3 Expression estimator

Given an image $\mathbf{I}$ of a face, $D_\mathrm{y}(\mathbf{I})$ is an expression regression network responsible for estimating the AUs' activation $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_N)^\top$ in the image. Similar to the conditional critic, its structure resembles that of PatchGan. To reduce the number of parameters of the model, $D_\mathrm{y}(\mathbf{I})$ is implemented on top of the conditional critic as an auxiliary head sharing the weights of the first five layers.

4.2 Learning the Model

The parameters of the generator, conditional critic and expression estimator are simultaneously estimated. For this purpose we define a loss function made of four terms, namely an *image adversarial loss* (Arjovsky et al. 2017) with the modification proposed by Gulrajani et al. (2017) that pushes the distribution of the generated images to the distribution of the training images; the *attention loss* to drive the attention masks to be smooth and prevent them from saturating; the *conditional expression loss* that conditions the expression of the generated images to be similar to the desired one; and the *identity loss* that favors to preserve the person texture identity. In the following we describe these losses.

4.2.1 Image adversarial loss

In order to learn the parameters of the generator G, we use the modification of the standard GAN algorithm (Goodfellow et al. 2014) proposed by WGAN-GP (Gulrajani et al. 2017). Specifically, the original GAN formulation is based on the Jensen–Shannon (JS) divergence loss function and aims to maximize the probability of correctly classifying real and rendered images while the generator tries to foul the discriminator. This loss is potentially not continuous with respect to the generator's parameters and can locally saturate leading to vanishing gradients in the discriminator. This is addressed in WGAN (Arjovsky et al. 2017) by replacing JS with the continuous EMD. To maintain a Lipschitz constraint, WGAN-GP (Gulrajani et al. 2017) proposes to add a

Journal: 11263 MS: 1210 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2019/8/10 Pages: 16 Layout: Large

gradient penalty for the critic network computed as the norm of the gradients with respect to the critic input.

Formally, let $\mathbf{I}_{\mathbf{y}_o}$ be the input image with the initial condition $\mathbf{y}_o$, $\mathbf{y}_f$ the desired final condition, $\mathbb{P}_o$ the data distribution of the input image, and $\mathbb{P}_{\widetilde{I}}$ the random interpolation distribution. Then, the *critic loss* we use is:

$$\begin{aligned}\mathcal{L}_\mathrm{I}(G, D_\mathrm{I}, \mathbf{I}_{\mathbf{y}_o}, \mathbf{y}_f) = &-\mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}[D_\mathrm{I}(G(\mathbf{I}_{\mathbf{y}_o}|\mathbf{y}_f))] \\ &+ \mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}\left[D_\mathrm{I}(\mathbf{I}_{\mathbf{y}_o})\right] \\ &- \lambda_{\mathrm{gp}}\mathbb{E}_{\widetilde{I}\sim\mathbb{P}_{\widetilde{I}}}\left[(\|\nabla_{\widetilde{I}}D_\mathrm{I}(\widetilde{I})\|_2 - 1)^2\right],\end{aligned} \tag{2}$$

where λ_{gp} is a penalty coefficient.

4.2.2 Attention loss

When training the model we do not have ground-truth annotation for the attention masks $\mathbf{A}$. Similarly as for the color masks $\mathbf{C}$, they are learned from the resulting gradients of the critic module and the rest of the losses. However, the attention masks can easily saturate to 1 which makes that $\mathbf{I}_{\mathbf{y}_o} = G(\mathbf{I}_{\mathbf{y}_o}|\mathbf{y}_f)$, that is, the generator has no effect. To prevent this situation, we regularize the mask with a weight penalty. Additionally, to enforce a smooth spatial color transformation when combining the regions of the input image and those of the color transformation $\mathbf{C}$, we perform a *Total Variation Regularization* over $\mathbf{A}$. The attention loss can therefore be defined as:

$$\mathcal{L}_\mathbf{A}(G, \mathbf{I}_{\mathbf{y}_o}, \mathbf{y}_f) = \lambda_{\mathrm{TV}}\mathcal{L}_{\mathrm{TV}}(\mathbf{A}) + \mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}\left[\|\mathbf{A}\|\right], \tag{3}$$

where $\mathbf{A} = G_A(\mathbf{I}_{\mathbf{y}_o}|\mathbf{y}_f)$ and $\mathbf{A}_{i,j}$ represents the $[i, j]$ entry of $\mathbf{A}$. λ_{TV} is a penalty coefficient for the mask smoothing, being the corresponding loss defined as:

$$\mathcal{L}_{\mathrm{TV}}(\mathbf{A}) = \sum_{i,j}^{H,W}\left[(\mathbf{A}_{i+1,j} - \mathbf{A}_{i,j})^2 + (\mathbf{A}_{i,j+1} - \mathbf{A}_{i,j})^2\right].$$

4.2.3 Conditional expression loss

While reducing the *image adversarial loss*, the generator must also reduce the error produced by the AUs' regression head on top of D. In this way, G not only learns to render realistic samples but also learns to satisfy the target facial expression encoded by $\mathbf{y}_f$. This loss is defined with two components: an AUs regression loss with fake images used to optimize G, and an AUs regression loss of real images used to learn the regression head on top of D. This loss is computed as:

$$\begin{aligned}&\mathcal{L}_\mathrm{y}(G, D_\mathrm{y}, \mathbf{I}_{\mathbf{y}_o}, \mathbf{y}_o, \mathbf{y}_f) \\ &\quad= \mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}\left[\|D_\mathrm{y}(\mathbf{I}_{\mathbf{y}_o}) - \mathbf{y}_o\|_2^2\right] \\ &\quad+ \mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}\left[\|D_\mathrm{y}(G(\mathbf{I}_{\mathbf{y}_o}|\mathbf{y}_f))] - \mathbf{y}_f\|_2^2\right].\end{aligned} \tag{4}$$

4.2.4 Identity loss

With the previously defined losses the generator is enforced to generate photo-realistic face transformations. However, without ground-truth supervision, there is no constraint to guarantee that the face in both the input and output images correspond to the same person. Using a *cycle consistency loss* (Zhu et al. 2017a) we force the generator to maintain the identity of each individual by penalizing the difference between the original image $\mathbf{I}_{\mathbf{y}_o}$ and its reconstruction. The *identity loss* $\mathcal{L}_{\mathrm{idt}}(G, \mathbf{I}_{\mathbf{y}_o}, \mathbf{y}_o, \mathbf{y}_f)$ is defined as:

$$\mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}\left[\|G(G(\mathbf{I}_{\mathbf{y}_o}|\mathbf{y}_f)|\mathbf{y}_o) - \mathbf{I}_{\mathbf{y}_o}\|_1\right]. \tag{5}$$

To produce realistic images it is critical for the generator to model both low and high frequencies. Our *PatchGan* based critic D_I already enforces high-frequency correctness by restricting our attention to the structure in local image patches. To also capture low-frequencies it is sufficient to use l_1-norm. In preliminary experiments, we also tried replacing l_1-norm with a more sophisticated *Perceptual* loss (Johnson et al. 2016), although we did not observe improved performance.

4.2.5 Full loss

To generate the target image $\mathbf{I}_{\mathbf{y}_g}$, we build a loss function $\mathcal{L}$ by linearly combining all previous partial losses:

$$\begin{aligned}\mathcal{L} = &\mathcal{L}_\mathrm{I}(G, D_\mathrm{I}, \mathbf{I}_{\mathbf{y}_r}, \mathbf{y}_g) + \lambda_\mathrm{y}\mathcal{L}_\mathrm{y}(G, D_\mathrm{y}, \mathbf{I}_{\mathbf{y}_r}, \mathbf{y}_r, \mathbf{y}_g) \\ &+ \lambda_\mathrm{A}\left(\mathcal{L}_\mathrm{A}(G, \mathbf{I}_{\mathbf{y}_g}, \mathbf{y}_r) + \mathcal{L}_\mathrm{A}(G, \mathbf{I}_{\mathbf{y}_r}, \mathbf{y}_g)\right) \\ &+ \lambda_{\mathrm{idt}}\mathcal{L}_{\mathrm{idt}}(G, \mathbf{I}_{\mathbf{y}_r}, \mathbf{y}_r, \mathbf{y}_g),\end{aligned} \tag{6}$$

where λ_A, λ_y and λ_{idt} are the hyper-parameters that control the relative importance of every loss term. Finally, we can define the following minimax problem:

$$G^\star = \arg\min_{G}\max_{D\in\mathcal{D}}\mathcal{L}, \tag{7}$$

where $G^\star$ draws samples from the data distribution. Additionally, we constrain our discriminator D to lie on $\mathcal{D}$, that represents the set of 1-Lipschitz functions.

Journal: 11263 MS: 1210 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2019/8/10 Pages: 16 Layout: Large

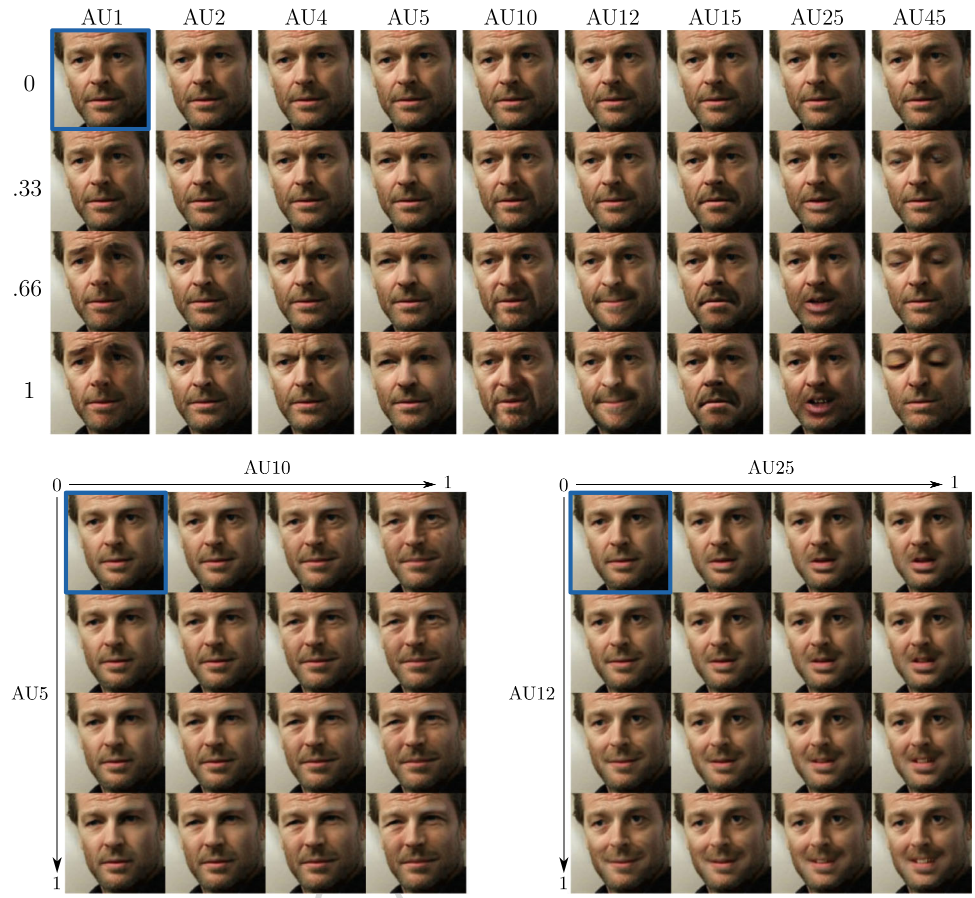

Fig. 4 Single and dual-AU edition. Top: Single AUs are activated at increasing levels of intensity (from 0.33 to 1). The first row corresponds to a zero intensity application of the AU which correctly produces the original image in all cases. Bottom: For every grid two specific AUs are activated at increasing levels of intensity (from 0 to 1). Left: Case in which the activation areas of the AUs (#10 and #5) do not overlap. Right: Both AUs activate overlapping areas of the face

5 Implementation Details

Our generator builds upon the variation of the network from Johnson et al. (2016) proposed by Zhu et al. (2017a) as it proved to achieve impressive results for image-to-image mapping. We have slightly modified it by substituting the last convolutional layer with two parallel convolutional layers, one to regress the color mask **C** and the other to define the attention mask **A**. We also observed that changing batch normalization in the generator by instance normalization improved training stability. For the critic we have adopted the *PatchGan* architecture of Isola et al. (2017), but removing feature normalization. Otherwise, when computing the gradient penalty, the norm of the critic's gradient would be computed with respect to the entire batch and not with respect to each input independently as is required by WGAN-GP.

The model is trained on the EmotioNet dataset (Benitez-Quiroz et al. 2016). We use a subset of 200,000 samples (over 1 million) to reduce training time. We use ADAM (Kingma and Ba 2015) with learning rate of 0.0001, beta1 0.5, beta2 0.999 and batch size 25. We train for 30 epochs and linearly decay the rate to zero over the last 10 epochs. Every five optimization steps of the critic network we perform a single optimization step of the generator. The weight coefficients for the loss terms in Eq. (6) are set to $\lambda_{gp} = 10$, $\lambda_A = 0.1$, $\lambda_{TV} = 0.0001$, $\lambda_y = 4000$, $\lambda_{idt} = 10$. To improve stability we tried updating the critic using a buffer with generated images in different updates of the generator, as proposed in Shrivastava et al. (2017), but we did not observe performance improvement.

Several design choices (e.g. sharing part of the weights between the conditional critic and the expression estimator) were done in order to fit the model into a single Nvidia® GTX 1080 Ti GPU with 11 GB of memory. The model is trained in 2 days on the 200,000 EmotioNet dataset samples. During testing only the regressors are necessary, and hence the size of the model is reduced to 813 MB. Inference can be done at 66 fps with an Nvidia ® GTX 1080 Ti GPU.

6 Experimental Evaluation

In this section we provide a thorough evaluation of the proposed architecture. Concretely, we evaluate GANimation's ability for single and multiple AUs editing, for discrete and continuous emotion editing, and compare it with existing techniques. We also provide a detailed analysis of the attention mechanism. Finally, we discuss the model's ability to deal with occlusions and its limitations and failure cases.

Journal: **11263** MS: **1210** ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:**2019/8/10** Pages: **16** Layout: **Large**

Fig. 5 Attention model. Details of the intermediate color mask **C** (first row) and the attention mask **A** (second row). The images in the bottom row are the synthesized expressions. Darker regions of the attention mask **A** show those areas of the image more relevant for each specific AU. Brighter areas are retained from the original image

It is worth pointing out that in some of the experiments the input faces are not cropped. In these cases we first use an off-the-shelf detector[2] to localize and crop the face, apply the expression transformation to that area with Eq. (1), and place the generated face back into its original image position. The attention mechanism is very helpful to process relatively high resolution images and a render smooth transitions between the morphed cropped faces and the original image.

6.1 Single Action Units Edition

We first evaluate our model's ability to activate AUs at different intensities while preserving the person's identity. Figure 4-top shows a subset of 9 AUs individually transformed with four levels of intensity (0, 0.33, 0.66, 1). For the case of 0 intensity it is desired not to change the corresponding AU. The model properly handles this situation and generates an identical copy of the input image for every case. The ability to apply an identity transformation is essential to ensure that non-desired facial movement is not be introduced.

For the cases with non-zero AU intensity, it can be observed how each AU is progressively accentuated. Note the difference between generated images at intensity 0 and 1. The model convincingly renders complex facial movements which in most cases are difficult to distinguish from real images. It is also worth mentioning that the independence of facial muscle clusters is properly learned by the generator. For instance, AUs relative to the eyes and the upper-half part of the face (AUs 1, 2, 4, 5, 45) do not affect the muscles of the mouth. Equivalently, mouth related transformations (AUs 10, 12, 15, 25) do not affect eyes nor eyebrow muscles.

Figure 5 shows, for the same experiment, the attention **A** and color **C** masks that produced the final result $\mathbf{I}_{\mathbf{y}_g}$. Note how the model has learned to focus its attention (darker area) onto the corresponding AU in a weakly supervised manner. In this way, it relieves the color mask from having to accurately regress each pixel value. Only the pixels relevant to convey the expression change are carefully estimated, the rest are just set to noise. For example, the attention is clearly obviating background pixels allowing to directly copy them from the original image. This is paramount to later being able to handle images in the wild with complex backgrounds (see Sect. 6.9).

6.2 Two Action Units Edition

In this subsection we evaluate the ability of our model to simultaneously activate two actions units. The model must not only be able to activate the desired AUs but also combine them in a realistic manner. The results of this experiment are shown in Fig. 4-Bottom. The left grid of the figure shows the case when the two AUs activate different areas of the face (AUs 5 is related to the chicks and 10 to the eyelids). In this case, since the muscles related to each AU are different, their effects are independent from one another. A more difficult case occurs when both AUs share facial muscles, see Fig. 4-Bottom-Right. In this specific case, when only activating $AU12$ (left column) the model draws a smile, but when we also activate $AU25$, in charge of controlling the distance between lips, the model produces a smile with the mouth open. Note that the generator hallucinates the teeth that would be visible when smiling with the lips apart.

6.3 Simultaneous Edition of Multiple AUs

We next push the limits of the GANimation model and evaluate it in the task of editing multiple AUs. Additionally,

[2] We use the face detector from https://github.com/ageitgey/face_recognition.

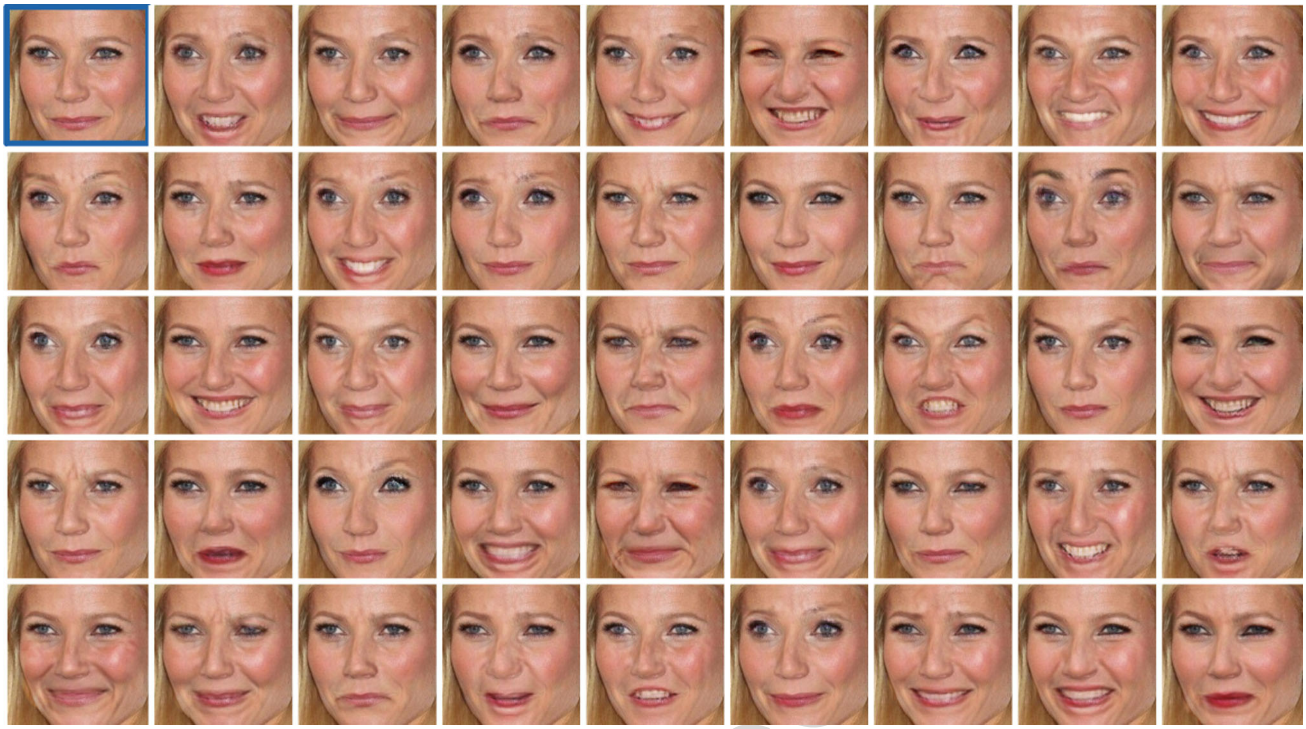

Fig. 6 Sampling the face expression distribution space. As a result of applying the AU-parametrization through the vector $\mathbf{y}_g$, we can synthesize, from the same source image $\mathbf{I}_{\mathbf{y}_r}$, a large variety of photo-realistic images

we also assess its ability to interpolate between two expressions. The results of this experiment are shown in Fig. 1 of Sect. 1. The first column is the original image with expression $\mathbf{y}_r$, and the right-most column is a synthetically generated image conditioned on a target expression $\mathbf{y}_g$. The rest of columns result from evaluating the generator conditioned with a linear interpolation of the original and target expressions: $\alpha\mathbf{y}_g+(1-\alpha)\mathbf{y}_r$. The outcomes show a very remarkable smooth and a consistent transformation across frames. We have intentionally selected challenging samples to show the robustness to complex lighting conditions and even, as in the case of the avatar, to non-real data distributions which were not previously seen by the model. These results are encouraging to further extend the model to video generation (Zhou et al. 2019; Nam et al. 2019; Song et al. 2018; Vondrick et al. 2016; Vougioukas et al. 2018) in future works.

6.4 High Expressions Variability

Given a single image, we next use GANimation to produce a wide range of anatomically feasible face expressions while conserving the person's identity. In Fig. 6 all faces are the result of conditioning the input image in the top-left corner with a desired face configuration defined by only 14 AUs. Note the large variability of anatomically feasible expressions that can be synthesized with only 14 AUs. Specially remarkable are some of the results in which parts of the face are not visible in the input image (e.g. teeth) need to be hallucinated.

6.5 Comparison with the State-of-the-Art

We next compare our approach against several baselines, namely DIAT (Li et al. 2016), CycleGAN (Radford et al. 2016), IcGAN (Perarnau et al. 2016) and StarGAN (Choi et al. 2018). For a fair comparison, we consider the results of these methods trained by the most recent work, StarGAN (Choi et al. 2018), on the task of rendering discrete emotions categories (e.g. happy, sad and fearful) trained and tested in the RaFD dataset (Langner et al. 2010). Face images in this dataset are properly cropped and aligned. Since DIAT (Li et al. 2016) and CycleGAN (Radford et al. 2016) do not allow conditioning, they were independently trained for every possible pair of source/target emotions. GANImation was also fine-tuned with the RaFD dataset. We next briefly discuss the main aspects of each approach:

DIAT (Li et al. 2016) Given an input image $x \in X$ and a reference image $y \in Y$, DIAT learns a GAN model to render the attributes of domain Y in the image x while conserving the person's identity. It is trained with the classic *adversarial loss* and a *cycle loss* $\|x - G_{Y\to X}(G_{X\to Y}(x))\|_1$ to preserve the person's identity.

CycleGAN (Radford et al. 2016) Similar to DIAT (Li et al. 2016), CycleGAN also learns the mapping between two

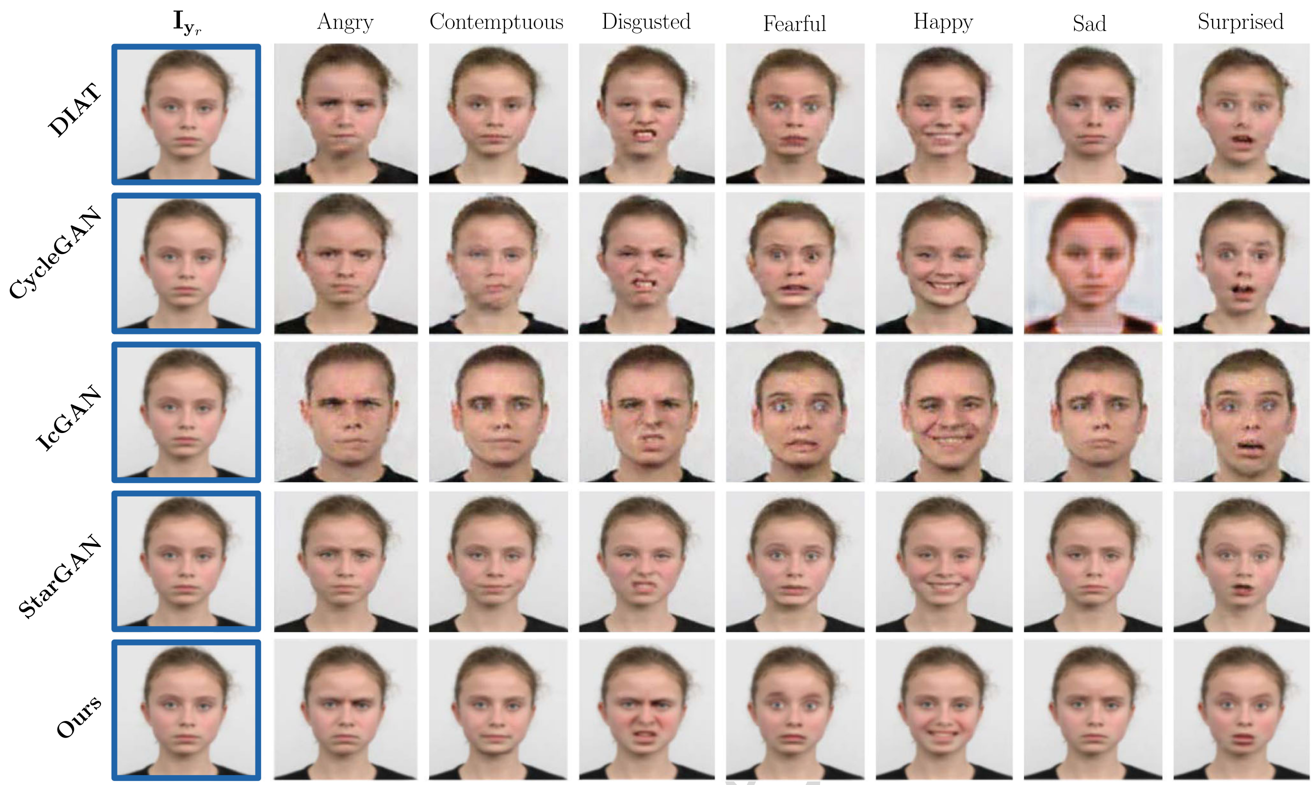

Fig. 7 Qualitative comparison with state-of-the-art. Facial expression synthesis results for: DIAT (Li et al. 2016), CycleGAN (Radford et al. 2016), IcGAN (Perarnau et al. 2016) and StarGAN (Choi et al. 2018); and our GANimation. In all cases, we represent the input image and seven different facial expressions. As it can be seen, our solution produces the best trade-off between visual accuracy and spatial resolution. Some of the results of StarGAN (Choi et al. 2018), the best current approach, show certain level of blur. Images of previous models were taken from Choi et al. (2018)

domains $X \rightarrow Y$ and $Y \rightarrow X$. To train the domain transfer, it uses a regularization term denoted *cycle consistency loss* that combines two cycles: $\|x - G_{Y\rightarrow X}(G_{X\rightarrow Y}(x))\|_1$ and $\|y - G_{X\rightarrow Y}(G_{Y\rightarrow X}(y))\|_1$.

IcGAN (Perarnau et al. 2016) Given an input image, IcGAN uses a pre-trained encoder–decoder to encode the image into a latent representation in concatenation with an expression vector $\mathbf{y}$ to then reconstruct the original image. It can modify the expression by replacing $\mathbf{y}$ with the desired expression before passing it through the decoder.

StarGAN (Choi et al. 2018) This approach is an extension of the *cycle loss* for simultaneously training between multiple datasets with different data domains. It uses a mask vector to ignore unspecified labels and to optimize only on known ground-truth labels. It yields more realistic results when training simultaneously with multiple datasets.

GANimation differs from these approaches in two main aspects. First, we do not condition the model on discrete emotions categories, but we learn a basis of anatomically feasible warps that allows generating a continuum of expressions. Secondly, the use of the attention mask allows applying the transformation only on the cropped face, and put it back onto the original image without producing transition artifacts. As shown in Fig. 7, besides estimating more visually compelling images than other approaches, this results on images with higher spatial resolution.

Table 1 Quantitative comparison with StarGAN (Choi et al. 2018)

Method	ACD ↓	IS ↑	User preference ↑ (%)
GANimation	0.31	**1.48**	**56**
StarGaN	**0.29**	1.41	44

The table reports the results of three metrics (described in the text): *Face Distance* (ACD Tulyakov et al. 2018, the lower the better), *Inception Score* (IS Salimans et al. 2016, the higher the better) and user preference (the higher the better)

Table 1 presents a quantitative analysis (that includes a user study) comparing GANimation and StarGAN (Choi et al. 2018), as a representative of current the state-of-the-art. We considered three metrics: the *Average Content Distance* (ACD) (Tulyakov et al. 2018), the *Inception Score* (IS) (Salimans et al. 2016) and the *User Preference*. ACD is the L_2-distance between feature vectors of the input and generated images extracted by a face classifier[2] (the lower the

Journal: **11263** MS: **1210** ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:**2019/8/10** Pages: **16** Layout: **Large**

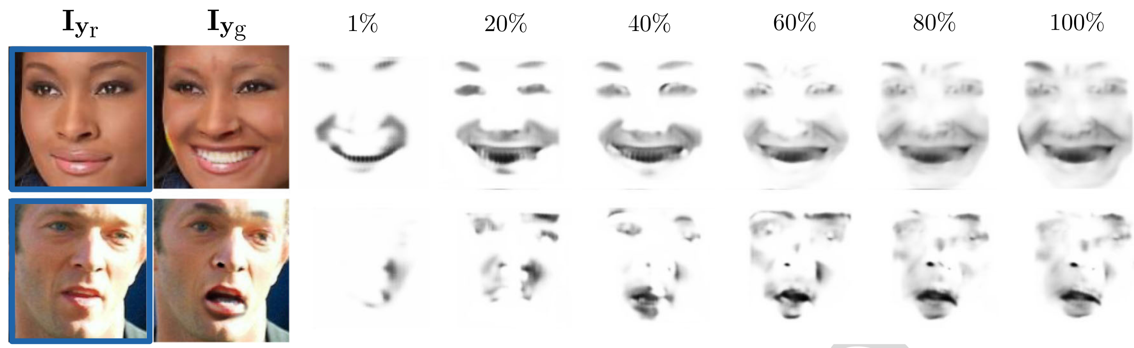

Fig. 8 Attention mask convergence (qualitative assessment). Evolution of the attention mask during training. Left to Right: Source $\mathbf{I}_{\mathbf{y}_r}$ and generated $\mathbf{I}_{\mathbf{y}_g}$ images, respectively; and the corresponding attention mask evolution (from 1 to 100%) of the total training epochs

better). IS is the metric used in previous approaches, that is higher for images with a large semantic content (the higher the better). For the study, we evaluated 100 randomly picked images from the RaFD dataset test set, each transformed to 5 randomly selected expressions. To compute the *User Preference* score we asked 20 human subjects to pick the most photo-realistic generated image among 20 randomly shuffled image pairs, one generated by each method. As shown in Table 1 both methods have a very similar performance in terms of the quality of the generated images. ACD is slightly favorable to StarGAN and GANimation is better in IS and User Preference. But recall that GANimation allows generating expressions in a continuum, while StarGAN is only able to render expressions from set of 8 emotion categories. We can conclude that GANimation retains/slightly improves the quality of StarGAN, while offering a much wider range of animation possibilities.

6.6 Attention Convergence

The most critical part when training GANimation is to ensure the correct convergence of the attention mask. The fact that we are not using ground-truth supervision can easily lead to the saturation of this mask, i.e., $\mathbf{A}_{H\times W} = (\mathbf{1})_{H\times W}$, meaning that $\mathbf{I}_{\mathbf{y}_o} - G(\mathbf{I}_{\mathbf{y}_o}|\mathbf{y}_f) = 0$, an hence the generator simply performs the identity mapping. Indeed, most terms in the loss function [see Eq. (6)] favor this situation, i.e., if the input image is not changed (identity generator) the photo-realism, the identity preservation and the smoothness of the attention mask are maximized. To avoid this from happening, we introduced the loss term $\mathcal{L}_A$ that explicitly enforces regularization over the attention mask and prevents it from saturating.

Figure 8 shows the convergence of the attention mask during training. We noted that in the first epochs the generator basically copies most parts of the original image (areas in white) and only introduces the basic lines that convey the new expression. After a while, the attention mask converges to a face segmentation mask that allows editing the fine details of the face such as color and shadows while leaving the original

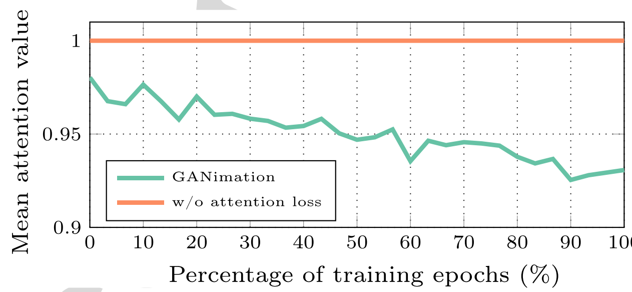

Fig. 9 Attention mask convergence (quantitative assessment). Mean value of the attention mask over training time

background unchanged. Figure 9 shows how the amount of newly created pixels (size of the darker regions in the attention mask) increases over the training time.

6.7 Ablation Study

To further analyze the GANimation's architecture and loss components we conducted an ablation study. Performing such ablation study, however, is not trivial, as most of the model elements are crucial for convergence. D_I and $\mathcal{L}_\mathrm{I}$ constrain the system to generate realistic images; D_y and $\mathcal{L}_\mathrm{y}$ ensure the proper expression conditioning when generating a new sample; and $\mathcal{L}_\mathrm{idt}$ enforces the model to preserve the person's identity. Removing any of these elements prevent the model from converging.

The only module that can be realistically ablated without catastrophically harming the network's performance is the attention mechanism $\mathbf{A} = G_A(\mathbf{I}_{\mathbf{y}_o}|\mathbf{y}_f)$ and its corresponding attention loss $\mathcal{L}_\mathrm{A}$. Figure 10 and Table 2 present a qualitative and quantitative ablation study of these two elements. For the quantitative results we considered three metrics: the *Average Content Distance* (ACD Tulyakov et al. 2018), the *Expression Distance* (ED) and the *User Preference*. ACD is the same metric as in Sect. 6.5 and ED is the l_1-distance between the generated and desired expressions (the lower the better). For the *User Preference* we have asked 20 human subjects to pick the most photo-realistic generated image (the

Fig. 10 Qualitative ablation study. Impact of the attention mechanism and the attention loss in the generated images. First row: Reference expressions. Second row: Results using the full GANimation pipeline. Third row: GANimation without the attention mechanism. Last row: GANimation without the attention loss

Table 2 Quantitative ablation study

Method	ACD ↓	ED ↓	User preference ↑ (%)
GANimation	0.4	**0.4**	**87**
w/o attention	0.4	**0.4**	13
w/o attention loss	**0.0**	4.8	0

Impact of the attention mechanism and the attention loss on the face generation results. Three metrics are considered (described in the text): *Face Distance* (ACD Tulyakov et al. 2018, the lower the better), *Expression Distance* (ED, the lower the better) and user preference (the higher the better)

higher the better). For the study, we evaluated 5000 randomly picked images from the CelebA (Liu et al. 2015) dataset, each transformed to 8 randomly selected expressions of the RaFD dataset. The model was not fine-tuned on CelebA. For the user study 20 randomly shuffled images were scored based on their photo-realism.

The quantitative results show that although we do not observe any gain on the face classification features nor on the estimated expressions, the proposed generation mechanism produces more photo-realistic images—better blended with the original background and better adjusted to the scene illumination. This is clearly reflected by the user study. When no attention is used, the cropped face bounding boxes are visible in the generated image and the illumination is not consistent (see Fig. 10-*w/o attention*). By contrasts, when using the proposed generator the background is perfectly blended and the illumination of the background and the generated image are consistent (see Fig. 10-*GANimation*).

The ablation study also demonstrates the necessity of introducing the proposed attention loss $\mathcal{L}_A$ for the proper convergence of the model (see Table 2). When removing it the obtained ACD metric is 0.0, meaning $\mathbf{I}_{\mathbf{y}_r} = \mathbf{I}_{\mathbf{y}_g}$, that is, the output image is identical to the input image.

6.8 Dealing with Occlusions

We next explicitly evaluate the robustness of the proposed approach to partial occlusions of the input face. The results are shown in Fig. 11. Interestingly, the attention mask tags the occluded pixels in white, meaning that these pixels will not be changed by the generator when creating the new expression. This is another interesting property of the attention mechanism, which besides learning a smooth foreground–background blending function, it also learns to ignore the

Journal: **11263** MS: **1210** TYPESET DISK LE CP Disp.:**2019/8/10** Pages: **16** Layout: **Large**

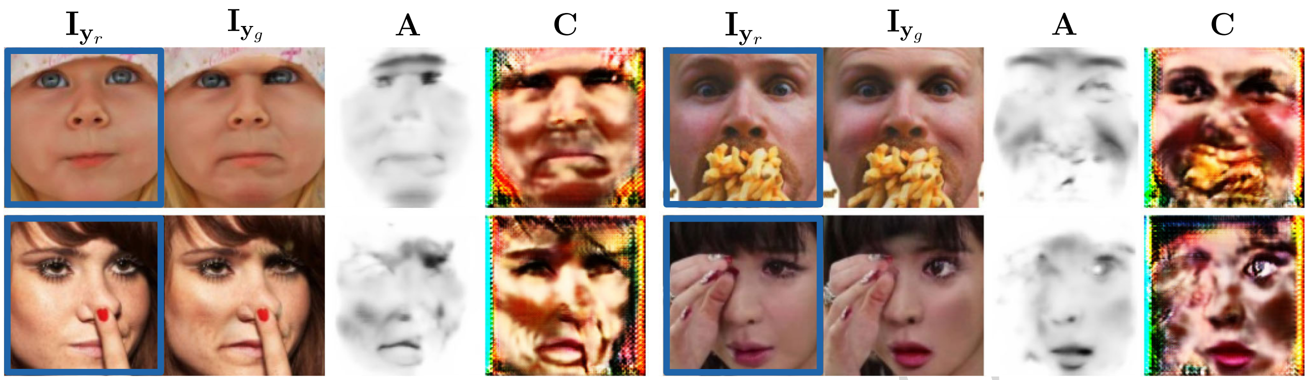

Fig. 11 Dealing with occlusions. Facial editing when dealing with input images containing occlusions. In all cases, we represent (from left to right) the source image $\mathbf{I}_{\mathbf{y}_r}$; the target image $\mathbf{I}_{\mathbf{y}_g}$; the attention mask $\mathbf{A}$; and the color mask $\mathbf{C}$. Top: Occlusions created by external interfering objects (had and french-fries). Bottom: Self-occlusions created by other parts of the body (hands and hair)

Fig. 12 Qualitative evaluation on images in the wild. Top: We represent an image (left) from the film "*Pirates of the Caribbean*" and an its generated image obtained by our approach (right). Bottom: In a similar manner, we use an image frame (left) from the series "*Game of Thrones*" to synthesize five new images with different expressions

static element of the image that do not participate in the generation of the facial expression, like hats, glasses, hands or interfering objects. Recall that this is learned in a weakly supervised manner.

6.9 Images in the Wild

As previously seen, the attention mechanism not only learns to focus on specific areas of the face but also allows smoothly merging the original and the generated image background. This allows our approach to be easily applied to images in the wild while still maintaining the resolution of the original images. For these images we follow the detection and cropping scheme we described before. Figure 12 shows two examples on these challenging images: the first example illustrates our model's performance on a multiple-face editing task with complex illumination; the second example deals with a non-human-like facial skin texture distribution, which is obviously not observed at training time. Note how the attention mask allows for a smooth and unnoticeable merging between the entire frame and the generated faces.

6.10 Pushing the Limits of the Model

We next push the limits of our network and discuss the model limitations when dealing with extreme situations such as stone like skin, drawings and face sketch abstractions. We have split success cases into six categories which we summarize in Fig. 13-top. The first two examples (top-row) correspond to human-like sculptures and non-realistic drawings. In both cases, the generator is able to maintain the artistic effects of the original image. Also, note how the atten-

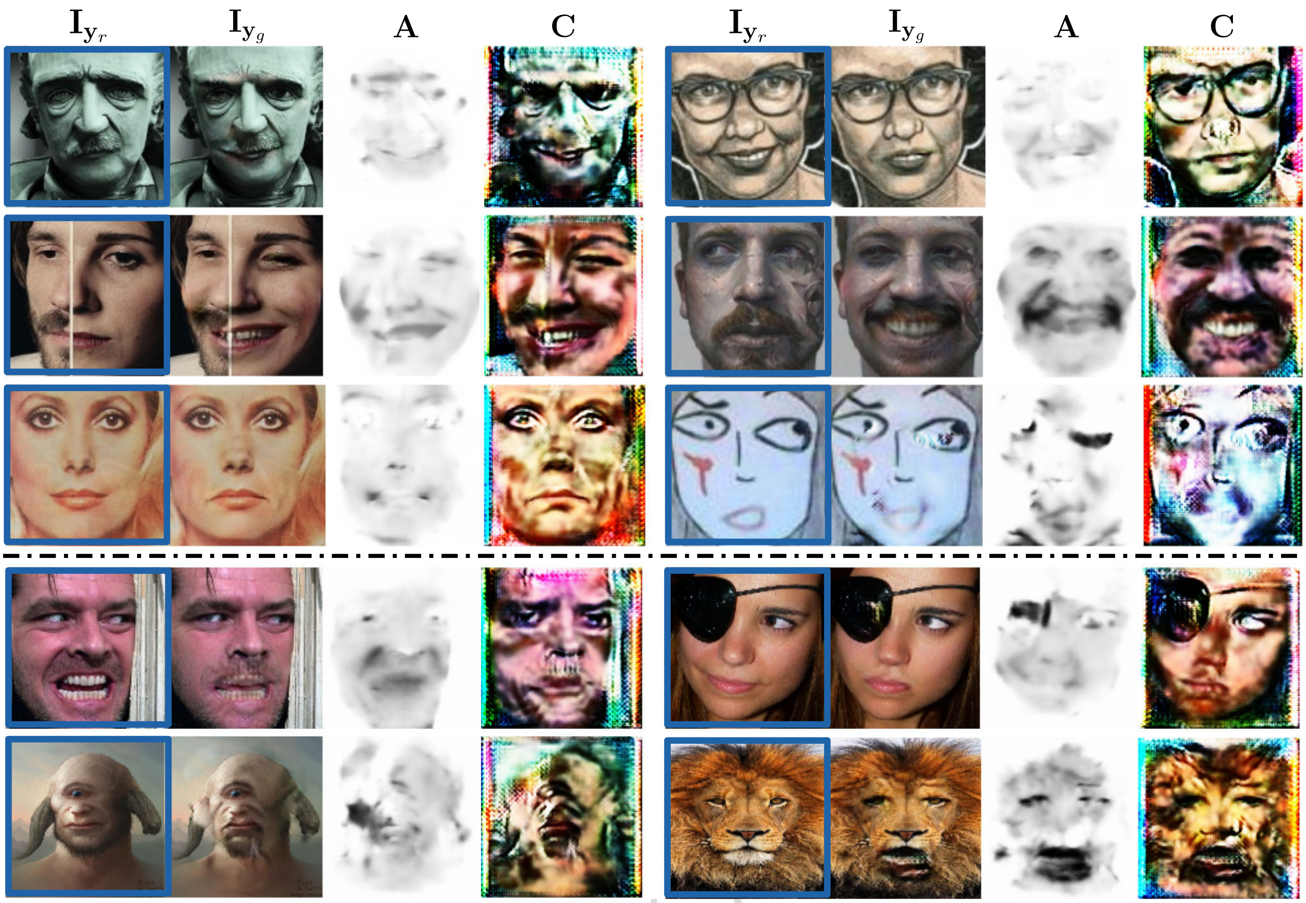

Fig. 13 Success and failure cases. In all cases, we represent the source image $\mathbf{I}_{\mathbf{y}_r}$, the target image $\mathbf{I}_{\mathbf{y}_g}$, the attention mask $\mathbf{A}$ and the color mask $\mathbf{C}$. Top: Some success cases in extreme situations. Bottom: Several failure cases

tion mask ignores artifacts such as the pixels occluded by the glasses. The third example (second-row, left) shows robustness to non-homogeneous textures over the face. Observe that the model is not trying to homogenize the texture by adding/removing the beard's hair. The second-row, right example, corresponds to an anthropomorphic face with non-real texture. As for the Avatar image, the network is able to warp the face without affecting its texture. The next category (third-row, left) is related to non-standard illuminations/colors for which the model has already been shown robust in Fig. 1. The last and most surprising category is face-sketches (third-row, right). Although the generated face suffers from some artifacts, it is impressive how GANimation is still capable of finding sufficient features on the face to transform its expression from worried to excited.

The fourth and fifth rows of Fig. 13 show a number of failure cases. The first case is related to errors in the attention mechanism when given extreme input expressions. The attention does not weight sufficiently the color transformation causing transparencies. The second case (fifth row, right) shows failures with non-previously seen occlusions such as an eye patch causing artifacts in the missing face attributes. The model also fails when dealing with non-human anthropomorphic distributions as in the case of cyclopes. Also, in this case, the face detection failed to detect the Cyclopes face forcing the generator to directly modify the original image without previously cropping the face. Lastly, we tested the model behavior when dealing with animals and observed artifacts like human face features.

7 Conclusions

We have presented GANimation, a novel GAN model for face animation in the wild that can be trained in a weakly supervised manner. It advances current works which, so far, had only addressed the problem for discrete emotions category editing and portrait images. Our model encodes anatomically consistent face deformations parameterized by means of Action Unit (AUs). Conditioning the GAN model on these AUs allows the generator to render a wide range of expressions by simple interpolation. Additionally, we embed an attention model within the network which allows focusing only on those regions of the image relevant for every specific expression. By doing this, we can easily process images in the wild, with distracting backgrounds, illumination artifacts and occlusions. We have thoroughly evaluated the model capabilities and limits in the EmotioNet (Benitez-Quiroz et al. 2016) and RaFD (Langner et al. 2010) datasets; conducted a quantitative and qualitative ablation study; studied the self-learned attention behaviour and its convergence; and finally demonstrated our model ability to deal with occlusions and images in the wild. The results are very promising, and show smooth transitions between different expressions. This opens the possibility of applying our approach to video sequences, which we plan to do in the future.

Acknowledgements This work is partially supported by an Amazon Research Award, by the Spanish Ministry of Economy and Competitiveness under Projects HuMoUR TIN2017-90086-R, ColRobTransp DPI2016-78957 and María de Maeztu Seal of Excellence MDM-2016-0656; by the EU Project AEROARMS ICT-2014-1-644271; and by the Grant R01-DC- 014498 of the National Institute of Health. We also thank Nvidia for hardware donation under the GPU Grant Program.

References

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. arXiv preprint arXiv:1701.07875

Baltrušaitis, T., Mahmoud, M., & Robinson, P. (2015). Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *FG*.

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 *CVPR*.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on computer graphics and interactive techniques* (pp. 187–194). New York: ACM Press.

Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*.

Dolhansky, B., & Canton Ferrer, C. (2018). Eye in-painting with exemplar generative adversarial networks. In *CVPR*.

Du, S., Tao, Y., & Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, *111*(15), E1454–E1462.

Ekman, P., & Friesen, W. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Altom: Consulting Psychologists Press.

Fischler, M. A., & Elschlager, R. A. (1973). The representation and matching of pictorial structures. *IEEE Transactions on Computers*, *22*(1), 67–92.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *NIPS*.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein GANs. In *NIPS*.

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A.: Image-to-image translation with conditional adversarial networks. In *CVPR* (2017)

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *ECCV*.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*.

Kim, T., Cha, M., Kim, H., Lee, J., & Kim, J. (2017). Learning to discover cross-domain relations with generative adversarial networks. In *ICML*.

Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, N., et al. (2018). Deep video portraits. *ACM Transactions on Graphics*, *37*, 163.

Kingma, D., & Ba, J. (2015). ADAM: A method for stochastic optimization. In *ICLR*.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In *ICLR*.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & Van Knippenberg, A. (2010). Presentation and validation of the radboud faces database. *Cognition and Emotion*, *24*(8), 1377–1388.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In *ICML*.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*.

Li, C., & Wand, M. (2016). Precomputed real-time texture synthesis with Markovian generative adversarial networks. In *ECCV*.

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

Liu, M. Y., Breuel, T., & Kautz, J. (2017). Unsupervised image-to-image translation networks. In *NIPS*.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *ICCV*.

Mathieu, M., Couprie, C., & LeCun, Y. (2016). Deep multi-scale video prediction beyond mean square error. In *ICLR*.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784

Nagano, K., Seo, J., Xing, J., Wei, L., Li, Z., Saito, S., et al. (2018). paGAN: Real-time avatars using dynamic textures. *ACM Transactions on Graphics*, *37*(6), 258:1–258:12.

Nam, S., Ma, C., Chai, M., Brendel, W., Xu, N., & Joo Kim, S. (2019). End-to-end time-lapse video synthesis from a single outdoor image. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1409–1418).

Odena, A., Olah, C., & Shlens, J. (2017). Conditional image synthesis with auxiliary classifier GANs. In *ICML*.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *CVPR*.

Perarnau, G., van de Weijer, J., Raducanu, B., & Álvarez, J. M. (2016). Invertible conditional GANs for image editing. arXiv preprint arXiv:1611.06355

Pumarola, A., Agudo, A., Martinez, A. A., Sanfeliu, A., & Moreno-Noguer, F. (2018). Ganimation: Anatomically-aware facial animation from a single image. In *ECCV*.

Pumarola, A., Agudo, A., Sanfeliu, A., & Moreno-Noguer, F. (2018) Unsupervised person image synthesis in arbitrary poses. In *CVPR*.

Radford, A., Metz, L., & Chintala, S. (2016). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICLR*.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee., H. (2016). Generative adversarial text to image synthesis. In *ICML*.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. In *Advances in neural information processing systems* (pp. 2234–2242). 4

Scherer, K. R. (1982). Emotion as a process: Function, origin and regulation. *Social Science Information*, *21*, 555–570.

Shen, W., & Liu, R. (2017). Learning residual images for face attribute manipulation. In *CVPR*.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., & Webb, R. (2017). Learning from simulated and unsupervised images through adversarial training. In *CVPR*.

Song, Y., Zhu, J., Wang, X., & Qi, H. (2018). Talking face generation by conditional recurrent adversarial network. arXiv preprint arXiv:1804.04786

Susskind, J. M., Hinton, G. E., Movellan, J. R., & Anderson, A. K. (2008). Generating facial expressions with deep belief nets. In *Affective computing*. Rijeka: InTech.

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

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2face: Real-time face capture and reenactment of RGB videos. In *CVPR*.

Tulyakov, S., Liu, M. Y., Yang, X., & Kautz, J. (2018). Mocogan: Decomposing motion and content for video generation. In *CVPR*.

Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Generating videos with scene dynamics. In *Advances in neural information processing systems* (pp. 613–621).

Vougioukas, K., Petridis, S., & Pantic, M. (2018). End-to-end speech-driven facial animation with temporal gans. In *BMVC*.

Wang, X., & Gupta, A. (2016). Generative image modeling using style and structure adversarial networks. In *ECCV*.

Wang, Z., Liu, D., Yang, J., Han, W., & Huang, T. (2015). Deep networks for image super-resolution with sparse prior. In *ICCV*.

Yu, H., Garrod, O. G., & Schyns, P. G. (2012). Perception-driven facial expression synthesis. *Computers and Graphics*, *36*(3), 152–162.

Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., & Shen, J. (2017). The menpo facial landmark localisation challenge: A step towards the solution. In *CVPRW*.

Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., & Metaxas, D. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*.

Zhou, H., Liu, Y., Liu, Z., Luo, P., & Wang, X. (2019). Talking face generation by adversarially disentangled audio–visual representation. In *AAAI*.

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017a). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.

Zhu, S., Fidler, S., Urtasun, R., Lin, D., & Loy, C. C. (2017b). Be your own prada: Fashion synthesis with structural coherence. In *ICCV*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Journal: 11263
Article: 1210

Author Query Form

Please ensure you fill out your response to the queries raised below and return this form along with your corrections

Dear Author

During the process of typesetting your article, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query	Details required	Author's response
1.	Kindly check and confirm the article title is correctly identified.	
2.	As per the information provided by the publisher, Fig. 1 will be black and white in print; hence, please confirm whether we can add "colour figure online" to the caption.	
3.	Please specify the significance of the bold reflected inside Tables 1 and 2 by providing a description in the form of a table footnote. Otherwise, kindly amend if deemed necessary.	
4.	Please update the complete details in the Refs. Salimans et al. (2016), Susskind et al. (2008) and Vondrick et al. (2016).	