Self-supervised Deformation Modeling for Facial Expression Editing

ShahRukh Athar, Zhixin Shu, Dimitris Samaras
1 Stony Brook University
2 Adobe Research
Email: {sathar, zhshu, samaras}@cs.stonybrook.edu

Abstract—Deep generative models have recently demonstrated impressive results in photo-realistic facial image synthesis and editing. Existing neural network-based approaches usually only rely on texture generation to edit expressions and largely neglect the motion information. However, facial expressions are inherently the result of muscle movement. In this work, we propose a novel end-to-end network that disentangles the task of facial editing into two steps: a “motion-editing” step and a “texture-editing” step. In the “motion-editing” step, we explicitly model facial movement through an image deformation, warping the image into the desired expression. In the “texture-editing” step, we generate the necessary textures, such as teeth and shading effects, for a photo-realistic result. Our physically-based task-disentanglement system design allows each step to learn a focused task, and thus need not generate texture to hallucinate motion. Our system is trained in a self-supervised manner, requiring no ground truth deformation annotation. Using Action Units [8] as the representation for facial expression, our method improves the state-of-the-art facial expression editing performance in both qualitative and quantitative evaluations.\footnote{The code can be found at https://github.com/cvlab-stonybrook/DefGAN}.

I. INTRODUCTION

Editing facial expressions of faces found ‘in-the-wild’ is a problem of great interest in Computer Vision and Computer Graphics, with a wide variety of applications in industries ranging from cinema to photography to e-commerce. An ideal facial expression editing system would allow the user to seamlessly change the expression of an input face without affecting invariant attributes such as facial identity, age, etc. Advances in adversarial learning [11] have recently enabled such expression editing in many cases [23], [6].

It is well known that facial expressions result from complex, constrained movements of facial muscles in three dimensions. Thus accurate expression modeling and encoding in 2D pixel space is challenging. In 1978, Eckman and Friesen [8] developed the Facial Action Coding System (FACS) that ‘can be used to describe any facial movement (observed in photographs, motion picture film, or videotape) in terms of anatomically based action units’. Each Action Unit (AU) corresponds to a change in some specific region of the face and any anatomically possible expression can be represented as a vector of AU detections or intensities. For example, a smile (possibly corresponding to happiness) can be encoded by the intensity of AU6 and AU12 and a sad expression can be encoded by the intensity of AU1, AU4, and AU15. Due to their interpretability and universality AUs are an ideal representation for editing expressions. Ground truth AU annotation of images is generally done by trained experts and is rather time-consuming. Over the years, however, there have been a number of learning-based methods that predict the AU activations of a face in any given image with low error rates, thus making the AU annotation of large scale datasets feasible. Recent work, such as GANimation [23] relies on AUs as a weak supervision signal to learn a facial expression model that allows one to seamlessly transition between expressions by interpolating in the AU space. Despite the AU supervision signal, GANimation [23] does not explicitly model facial movement and instead uses texture synthesis to mimic facial movement effects. A downside of this are possible artifacts

Figure 1: Expression Editing using DefGAN. The image on the top shows an input face image being edited to the target expression ‘disgust’, where $\alpha$ controls the activation of the target AUs [8]. DefGAN edits the image in two phases. First, in the ‘Motion editing’ phase, we deform the input image to conform to the target expression as can be seen in the top row of Fig. 1a. Next, in the ‘Texture editing’ phase we hallucinate the necessary textures to give us the final image (bottom row Fig. 1a). As can be seen, the deformation models facial movements and performs most of the editing work. In contrast, as shown in Fig. 1b, GANimation [23] edits the image entirely using texture hallucinations leading to artifacts in the final image (bottom right image of Fig. 1b)
In the editing results such as the disappearing eyebrows and beard changes seen in Fig. 1.

In this work, we propose DefGAN, a method that separates the task of editing expressions into two sequential phases – a ‘Motion Editing’ phase which models facial movements as an image deformation followed by a ‘Texture Editing’ phase to hallucinate final details that arise from the appearance or disappearance of texture (such as teeth). In the ‘Motion Editing’ phase, a Convolutional Neural Network (CNN) models facial movement explicitly by predicting a deformation field that deforms the input face to the target expression (for example, curving the region around the mouth into a crescent to generate a smile). Based on recent work [24], we model facial movements with an offset based deformation field. In the ‘Texture Editing’ phase, we hallucinate the necessary textures (such as teeth, shadows) using another CNN (on top of the deformed image) to complete the editing process and obtain the final edited image.

We weakly supervise our networks using easily obtainable AU annotations on images from the EmotionNet dataset [9] and use an adversarial learning framework similar to [23], [6] for training. The generator is tasked with editing a given input image towards a desired target expression while ensuring the edited image can be mapped back to the original image using a cyclic transformation [23], [6], [31]. Meanwhile, the discriminator ensures that the edited image looks real and also conforms to the target expression.

To summarize, our method learns anatomically consistent, expression-conditioned deformation fields (without ground-truth deformation annotations) through the explicit disentanglement of the editing process into ‘Motion editing’ and ‘Texture editing’ phases. The user studies we have conducted show that facial expression editing methods such as GANimation [23] and ours produce realistic expressions (with an average plausibility score of 3.5 out of maximum 4.0). The scores have a bimodal distribution. Encouraged by these results, we carried out a user study to directly compare the quality of DefGAN’s editing results with GANimation [23] and found that, on average, users prefer the editing results of DefGAN over GANimation [23]. In addition to the user study, we also carry extensive expression edits on a large variety of faces in-the-wild and show that editing expressions by explicit disentanglement of facial movement and texture synthesis lead to more realistic results while better preserving expression-invariant features of the face.

II. RELATED WORK

Over the past few years, there has been a significant amount of work on editing expressions and more broadly in transferring images from one domain to another.

**Face Manipulation.** The earliest face expression models relied on mass-and-skin models for facial movement [10]; such models, however, could not model finer skin movements often involved in facial expressions. Another line of research used registered 3D face scans to create linear low-dimensional face embeddings [5] by explicitly taking into account variations due to expressions. Though such models often produce realistic expression edits, they require expensive 3D scans of the same person with different expressions and cannot easily scale to learn from larger datasets. Other work using detailed 3D Scans of the human face to edit expressions includes [29], [28], [27].

Recent developments in generative adversarial networks have allowed the training of Convolutional Neural Networks not only to be used for generating very photorealistic face images [18] but also used for to the development of a number of unsupervised and weakly supervised facial manipulation methods such as [21], [25]. More specifically, in the case of expression editing, adversarial learning has used landmarks [26], discrete expression labels [6] and continuous Action Units [23] as weak supervision. In this work, we choose Action Units [8] as a weak supervision source due their interpretability (each AU corresponds to a change in some region of the face) and wide applicability (AUs can encode any anatomically possible facial expression [8]).

**Generative Adversarial Networks.** Generative Adversarial Networks (GANs) [11] are a powerful class of generative models that have essentially become the standard for unconditional image generation. GANs work by pitting neural networks against each other; the generator network is tasked with producing samples that are indistinguishable from the real data distribution while the discriminator is tasked with distinguishing between the samples generated by the generator and the real data. Follow up work [3], [12], has significantly improved GAN training stability.

**Conditional GANs.** Conditional GANs are a subset of GANs that are used to sample from conditional distributions instead of unconditional distributions. Conditional GANs have been very successful in a variety of computer vision tasks such as image in-painting [14], super-resolution [22], and domain transfer [31], [6], [19].

**Deformation modelling.** There is growing interest in learning deformations from large scale datasets. It is now possible to model face images as deformations of texture templates in a completely unsupervised manner [24]. Early work on using deformations to edit expressions includes Expression flow [30] which edits expressions by warping the input image using a 2D flow field, estimated using another image of the same person with the desired expression. We also build on prior work on incorporating deformations within convolutional networks such as [16], [7].

**Unpaired Image-to-Image Translation.** We cast the problem of editing expressions in the unpaired image to image translation framework, where the target image is the image of the input person having the target expression. The advent of GANs [11] has made it possible to produce very photorealistic results when translating from one domain to
another. For example, pix2pix [15] generates high quality images from mere sketches or segmentation maps as inputs. More relevant to our work are methods [31], [6], [19], [23] that transfer various facial attributes, including expressions.

III. METHOD

Consider an input image \( I_x \) with some expression \( x \). We aim to change the expression of the person in \( I_x \) to some target expression \( y \) to get the image \( I_y \). Here, expressions are encoded by AU [8] intensities, \( x = (x_1, \ldots, x_n) \), where each \( x_i \) is the intensity of the \( i \)th AU scaled between 0 and 1. We edit this expression in two stages. First, in the motion editing phase, we deform the input image to conform to the target expression. We use a deformation generator \( G_{\text{Def}} \) to produce the deformed image \( I_y^* \). Although this image could have been appropriately deformed to achieve the target expression, it might still lack the necessary texture modifications to look realistic. For example, if we had to edit a face with a neutral expression to a face with a grin, the best a deformed image could give would be a wide (and possibly unrealistic) smile, and we would still need to hallucinate the texture of the teeth to get the correct target expression. We hallucinate the necessary texture, in the texture editing phase, using another convolutional network:

\[
I_y = G_{\text{Texture}}(I_y^*, y)
\]  

where, \( G_{\text{Texture}} \) is the texture hallucination network. In the interest of brevity, we denote the entire transformation from input image to the final image \( I_x \to I_y \), as follows:

\[
I_y = G_{\text{Comp}}(I_x, x, y)
\]  

where, \( G_{\text{Comp}} := G_{\text{Texture}} \circ G_{\text{Def}} \) denotes the composition of (1) and (2).

A. Architecture

DefGAN consists of three convolutional networks. The deformation generator \( G_{\text{Def}}^W \), the texture hallucination network \( G_{\text{Texture}} \), and a discriminator \( D \) with a critic output, \( D_c \), and an AU regression output, \( D_{\text{exp}} \).

1) Motion Editing Phase: The first stage of DefGAN’s expression editing involves deforming the input image, \( I_x \) to \( I_y^* \) such that \( I_y^* \) closely approximates the target expression. We use a deformation generator \( G_{\text{Def}}^W \) that first predicts a deformation grid and then warps the input image using it. More specifically, the transformation can be written as:

\[
\mathcal{G} = G_{\text{Def}}(I_x, x, y)
\]

\[
I_y^* = \text{Warp}(\mathcal{G}, I_x)
\]  

with \( \mathcal{G} \) the predicted deformation grid, \( G_{\text{Def}}^W \), the deformation generator, is the composition of the above two operations

\[
G_{\text{Def}}^W(I_x, x, y) := \text{Warp}(G_{\text{Def}}(I_x, x, y), I_x)
\]  

\[
I_y^* = G_{\text{Def}}^W(I_x, x, y)
\]  

We use an offset based deformation grid as proposed in [24] with a maximum offset of 5 pixels.

2) Texture Editing Phase: The second stage of DefGAN’s expression edit, the texture editing phase, hallucinates the necessary features that cannot be modelled by a deformation to improve final image realism and fidelity to the target expression. As in [23], we found that a masking mechanism helps the texture generator improve the edit quality

\[
T, M = G_{\text{Texture}}(I_y^*, y)
\]

\[
I_y = M \odot T + (1 - M) \odot I_y^*
\]  

with \( T \) the hallucinated texture map and \( M \) the attention mask. \( G_{\text{Texture}} \), the texture network, is the composition of the above two operations.

B. Training

Similar to prior work [23], [6] we rely on a GAN based framework [11] to train the deformation network, \( G_{\text{Def}}^W \), and the texture network, \( G_{\text{Texture}} \). In addition to an adversarial loss, we train DefGAN to also minimize an expression loss, a cycle consistency loss, a facial identity loss and a regularization loss on the deformation grid.

Adversarial Losses: In order to ensure DefGAN’s image edits look natural we train the generators to minimize an adversarial loss [11]. We use the WGAN-GP loss [12] which minimizes the Earth Mover Distance between the generated and the real distribution. Specifically, let \( I_x \) be the input image with expression \( x \), let \( y \) be the target expression and let \( P_r \) be the real distribution of images. The critic loss for the discriminator, \( D \) is:

\[
L_{\text{critic}}^D = \mathbb{E}_{I_x \sim P_r} \left[ D_c(G_{\text{Comp}}(I_x, x, y)) \right] - \mathbb{E}_{I_x \sim P_r} \left[ D_c(I_x) \right] + \lambda_{\text{gp}} \mathbb{E}_{\hat{x} \sim P} \left[ \left\| \nabla_\hat{x} D_c(\hat{x}) \right\|_2 - 1 \right]^2
\]  

where \( D_c \) the critic output of the discriminator \( D \), \( \lambda_{\text{gp}} \) the gradient penalty coefficient and \( P \) the interpolated distribution. The generators, \( G_{\text{Def}}^W \) and \( G_{\text{Texture}} \) are trained to ‘please’ the critic by maximizing the critic score. This loss is:

\[
L_{\text{critic}}^G = - \mathbb{E}_{I_x \sim P_r} \left[ D_c(G_{\text{Comp}}(I_x, x, y)) \right]
\]  

Expression Losses: To ensure that the generators, while producing a realistic image, are also generating the correct target expression we add a loss that penalizes deviations from the target expression. This loss is defined using the AU output of the discriminator, \( D_{\text{exp}} \), that is trained to predict the AU intensities for any given input image \( I_x \):

\[
L_{\text{exp}}^{G_{\text{Def}}} = \lambda_{\text{exp}}^{G_{\text{Def}}} \mathbb{E}_{I_x \sim P_r} \left[ \|D_{\text{exp}}(G_{\text{Def}}(I_x, x, y)) - y\|_2 \right]
\]

\[
L_{\text{exp}}^{G_{\text{Comp}}} = \lambda_{\text{exp}}^{G_{\text{Comp}}} \mathbb{E}_{I_x \sim P_r} \left[ \|D_{\text{exp}}(G_{\text{Comp}}(I_x, x, y)) - y\|_2 \right]
\]

\[
L_{\text{exp}}^{G} = L_{\text{exp}}^{G_{\text{Def}}} + L_{\text{exp}}^{G_{\text{Comp}}}
\]  

\[
G_{\text{Def}}^W(I_x, x, y) := \text{Warp}(G_{\text{Def}}(I_x, x, y), I_x)
\]  

\[
I_y^* = G_{\text{Def}}^W(I_x, x, y)
\]
Here, $\lambda_{G_{\text{comp}}}$ and $\lambda_{G_{\text{def}}}$ are the coefficients of each term. We apply the expression loss both on the final image output $I_{y} = G_{\text{Comp}}(I_{x}, x, y)$ and the intermediate image output $I_{x} = G_{\text{Def}}^{W}(I_{x}, x, y)$. The AU output, $D_{\text{exp}}$, is trained to minimize the AU prediction error on real images

$$L_{\text{exp}}^{D} = \mathbb{E}_{I_{x} \sim P_{r}} \left[ \left\| D_{\text{exp}}(I_{x}) - x \right\|_{2}^{2} \right]$$

(10)

During training we set $\lambda_{G_{\text{exp}}} = 1000.0$, and $\lambda_{G_{\text{comp}}} = 4000$.

**Cycle Loss.** In order to preserve subject identity as DefGAN edits the image, we enforce a cycle consistency loss on the generator networks as follows:

$$L_{\text{cycle}}^{G} = \lambda_{\text{cycle}} \mathbb{E}_{I_{x} \sim P_{r}} \left[ \left\| G_{\text{Comp}}(G_{\text{Comp}}(I_{x}, x, y), y, x) - I_{x} \right\|_{1} \right]$$

(11)

In the absence of ground truth images for each person with different annotated expressions, we found that this cycle loss ensures the identity of the person does not change as the expression changes. During training we set $\lambda_{\text{cycle}} = 100.0$

**Face Identity Loss.** We regularize the deformation generator, $G_{\text{Def}}^{W}$, by ensuring that it preserves the identity of the person as it deforms the input image. More specifically, we maximize the cosine similarity between the OpenFace [2] embedding of the deformed input image, $T_{x}$, and the input image $I_{x}$. This loss can be expressed as

$$L_{\text{faceID}}^{G_{\text{Def}}} = \mathbb{E}_{I_{x} \sim P_{r}} \left[ 1 - \cos(\text{OpenFace}(T_{x}), \text{OpenFace}(I_{x})) \right]$$

(12)

**Composition Loss.** Imposing a composition loss on the generated deformation grid $G$ was useful in producing realistic expression edits. The grid composition loss is:

$$L_{\text{comp}}^{G_{\text{Def}}} = \lambda_{\text{comp}} \mathbb{E}_{I_{x} \sim P_{r}} \left[ \left\| \text{Warp}(G_{\text{cycle}}, G) - I_{D_{\text{def}}} \right\|_{2}^{2} \right]$$

(13)

where, $G_{\text{cycle}}$ is the deformation grid produced during the cycle transformation and $I_{D_{\text{def}}}$ is the identity deformation grid. We use $\lambda_{\text{comp}} = 10.0$ during training.

**Regularization.** To ensure smoothness of the generated deformation fields we add a TV-regularization term on the deformation grid, $G$ as defined in (4), and also penalize the difference between $G$ and the identity deformation:

$$L_{reg}^{G_{\text{Def}}} = \lambda_{T_{V}}^{G_{\text{Def}}} \mathbb{E}_{I_{x} \sim P_{r}} \left[ \sum_{i,j} (M_{i+1,j} - M_{i+1,j})^2 + (M_{i,j+1} - M_{i,j})^2 \right]$$

$$+ \lambda_{T_{V}}^{G_{\text{def}}} \mathbb{E}_{I_{x} \sim P_{r}} \left[ \left\| M \right\|_{2} \right]$$

(14)

where, $I_{D_{\text{def}}}$ is the identity deformation grid. We also add a similar regularization to the mask, $M$ that is generated during the texture editing phase:

$$L_{reg}^{M} = \lambda_{T_{V}}^{M} \mathbb{E}_{I_{x} \sim P_{r}} \left[ \sum_{i,j} (M_{i+1,j} - M_{i,j})^2 + (M_{i,j+1} - M_{i,j})^2 \right]$$

$$+ \lambda_{T_{V}}^{M} \mathbb{E}_{I_{x} \sim P_{r}} \left[ \left\| M \right\|_{2} \right]$$

(15)

We use $\lambda_{T_{V}}^{G_{\text{Def}}} = 1e-5, \lambda_{T_{V}}^{G_{\text{def}}} = 0.1$ and $\lambda_{T_{V}}^{M} = 1e-5$ during training.

**Final Loss.** The total loss on the generators is

$$L_{\text{Total}}^{G} = L_{\text{exp}}^{G} + L_{\text{cycle}}^{G} + L_{\text{faceID}}^{G_{\text{Def}}} + L_{\text{comp}}^{G_{\text{Def}}} + L_{\text{reg}}^{G_{\text{Def}}} + L_{\text{reg}}^{M}$$

(16)

We minimize this loss over the parameters of $G_{\text{Def}}$ and $G_{\text{Texture}}$ to convergence:

$$G_{\text{Def}}, G_{\text{Texture}}^{*} = \arg\min_{G_{\text{Def}}, G_{\text{Texture}}} L_{\text{Total}}^{G}$$

(17)

The total loss on the discriminator is:

$$L_{\text{Total}}^{D} = L_{\text{exp}}^{D} + L_{\text{cycle}}^{D}$$

(18)

The discriminator is trained to minimize this loss:

$$D^{*} = \arg\min_{D} L_{\text{Total}}^{D}$$

(19)
IV. EXPERIMENTS AND RESULTS

We evaluated our model on a variety of facial identities and on a range of expression editing tasks to test its quality and robustness. We used 168 different facial identities from the CelebA-HQ dataset [17] for evaluation. In addition to CelebA-HQ [17] we scraped 40 more images from the internet with variations in pose, illumination and facial attributes to test the robustness of our model on more challenging input images. First, we show the results of expression editing on a number of in-the-wild faces and measure the change in facial identity after the expression edit by comparing the OpenFace [2] embeddings of the input image and the edited image. Next, we show the results of manipulating single Action Units [8] using our model and finally, we discuss the results of a user study conducted to determine which model among DefGAN and GANimation [23] produced better expression edits on in-the-wild images as judged by humans.

A. Training Details

DefGAN. We trained DefGAN on a subset of the EmotionNet dataset [9] containing 190k images. We used the Adam optimizer [20] with an initial learning rate of $1e^{-4}$, $\beta_1 = 0.5$, $\beta_2 = 0.999$ and a batch size of 25. The model was trained for 40 epochs with the learning rate decaying to 0 over the last 20 epochs. The deformation generator and the texture generator were optimized jointly. The critic was trained for 10 steps for every step of the generators.

GANimation. GANimation [23] was trained on the same subset of the EmotionNet dataset [9] containing 190k images. We used the hyperparameters used by the authors in [23] with the only difference that we train for 40 epochs.

B. Expression Synthesis on Faces in-the-Wild

We first tested our method by performing expression edits on in-the-wild images from CelebA-HQ dataset [17] and on 40 more images scraped from the internet. The AU representation for each expression was computed by running the OpenFace AU detector [4] on all the peak expression images of a randomly selected person from the MUG dataset [1], which consists of seven labeled expressions (anger, disgust, fear, neutral, happy, sad and surprise) for 84 persons. Fig. 4 shows the results of expression edits performed on a few images from the internet and from CelebA-HQ [17]. As can be seen, our model consistently performs better edits across all expressions and faces. In particular, we noticed that GANimation [23] tends to distort the face by either producing artifacts (for example, the result of ‘Disgust’ in row 3 of Fig. 4) or by ‘over-editing’ (for example, the results of ‘Happy’ in row 1, 2 and 4 of Fig. 4) which we posit is due to its complete reliance on the hallucination mechanism. In contrast, since DefGAN uses a deformation to warp the face to conform to the target expression, it only hallucinates the necessary details and does not produce such artifacts. Table I shows the Fréchet Inception Distance (FID) [13] score of the images edited by DefGAN and GANimation. The FID score [13] is a standard evaluation measure for the realism of generative models. The lower the FID score of a model, the more realistic are its images. Results in Table I further suggest that the editing results of DefGAN are more realistic.

Fig. 5 shows the absolute pixel-wise difference between the input image and the edited image across a few target expressions. Ideally, we would only see differences in regions that correspond to the expression change. As can be seen in Fig. 5, the edits made by DefGAN are significantly more concentrated to the regions relevant to the final expression than the edits made by GANimation which tend to be more spread out. Fig. 3 shows the distance between the edited image and the input image in the OpenFace [2] embedding space on fifty different randomly chosen representations of each expression. DefGAN retains the input facial identity much better than GANimation across all the six target expressions. The retention of facial identity is also seen visually in Fig. 4, where attributes such as the beard and eyebrows of the edited images (results of ‘Disgust’ in all rows, the results of ‘Happy’ in row 2) have greater fidelity to the input with DefGAN’s edits than GANimation’s edits which tend to either erase or thicken them. Fig. 1 shows how deformations can be especially helpful in certain expression edits, such as going from a close to neutral expression to a ‘disgust’ expression. In this expression transformation, we see that the deformation in DefGAN does most of the work of converting the input face to the ‘disgust’ expression while the hallucination only adds minor details to the final image.

<table>
<thead>
<tr>
<th>Method Name</th>
<th>FID Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>GANimation [21]</td>
<td>4.75</td>
</tr>
<tr>
<td>DefGAN</td>
<td>3.82</td>
</tr>
</tbody>
</table>

Table I: FID Scores of GANimation [23] and DefGAN.
Figure 4: Editing Facial Expressions. Here we show images edited by DefGAN and GANimation [23] to various target expressions. GANimation [23] tends to produce artifacts (results of ‘Anger’ and ‘Disgust’ in row 3) or ends up hallucinating inaccurate textures (results of ‘Happy’ in row 2 and results of ‘Anger’ in row 4). In contrast, the editing results of DefGAN are more consistent with fewer artifacts and more accurate textures.

C. Learnt facial movements conditioned on Action Units

In this section, we analyze the effect of changing individual AUs (AU4, AU5, AU14, and AU26) on an input face. We show their effect on the deformed image. As can be seen from Fig. 6, DefGAN successfully learns to faithfully model facial movement through its deformation mechanism. For example, when increasing the intensity of AU5 (Upper Lid Raiser) we clearly see the eyebrows raising up while other regions of the face remain unchanged. Coherent facial movements resulting from deformations can also be seen as we change the intensities of AU4 (Brow Lowerer), AU14 (Dimpler) and AU26 (Jaw Drop).
Figure 5: Difference Images. This figure shows the pixel-wise absolute difference between images edited by DefGAN or GANimation and the input image. DefGAN only changes the parts of the input necessary to attain the target expression while GANimation [23] changes larger portions of the input image regardless of target expression.

Figure 6: Learnt Action Unit conditioned facial movements. **Left:** Here we show the effects of single AU activations on the deformed image $I^*$. **Right:** Here we show the regions of the face affected by the change in intensity of the corresponding AU. As can be seen, changing the intensity of any particular AU causes smooth changes in the corresponding facial regions akin to the results of true facial muscle movement as encoded by that AU.

**D. User Study**

We conducted an anonymous user study to evaluate the quality of expression edits made by DefGAN and GANimation [23]. We performed the user study in two stages: In the first stage where 55 users participated, we evaluate how realistic are the editing results of each method, without directly comparing them. We randomly sample 10 images edited by each method and show them to the users, asking them to rate the plausibility of the image from 1 (Definitely implausible) to 4 (Definitely plausible). Each image was assigned a random target expression from the following expressions: happy, disgust, sad, fear, angry, and surprise. The results of our method and GANimation [23] were placed side-by-side and users were asked to judge the quality of each edited image with respect to its fidelity to the facial identity of the input image, the closeness to the target expression and the overall plausibility of the image. The options are presented to the users in random orders to alleviate bias. The results of the user study are shown in Fig. 7. Users mostly preferred the results of DefGAN over GANimation [23] on the whole and across most expressions. The results were close when the target expression was ‘angry’ or ‘happy’ but were overwhelmingly in favor of DefGAN for all other expressions. The results of the user study provide further evidence that expression editing results of DefGAN are not only more realistic but also preserve facial identity and attain the target expression better than GANimation does.

**V. Conclusion**

To summarize, we presented a novel method for facial expression editing that can produce high-quality expression edits on in-the-wild images. Based on the most recent advances in deformation modeling [24] and expression editing [23], we created a method that is able to learn facial movements as deformations without using ground-truth deformation annotations. The explicit use of a deformation in the “motion-editing” phase allows DefGAN to perform targeted edits on the input face which, extensive evaluations show, not only produces very high quality edited images but also better retains expression invariant facial attributes. In future work, we hope to improve expression modeling by also taking into account the temporal nature of expressions and possibly extending expression editing to in-the-wild images were chosen for this stage. These images were selected to have a close to neutral input expression and with non-extreme poses. Each image was assigned a random target expression from the following expressions: happy, disgust, sad, fear, angry, and surprise. The results of our method and GANimation [23] were placed side-by-side and users were asked to judge the quality of each edited image with respect to its fidelity to the facial identity of the input image, the closeness to the target expression and the overall plausibility of the image. The options are presented to the users in random orders to alleviate bias. The results of the user study are shown in Fig. 7. Users mostly preferred the results of DefGAN over GANimation [23] on the whole and across most expressions. The results were close when the target expression was ‘angry’ or ‘happy’ but were overwhelmingly in favor of DefGAN for all other expressions. The results of the user study provide further evidence that expression editing results of DefGAN are not only more realistic but also preserve facial identity and attain the target expression better than GANimation does.
video sequences.

VI. ACKNOWLEDGMENTS

We would like to thank Francesc Moreno-Noguer for his valuable comments. This work was supported by a gift from Adobe, NSF grants CNS-1718014 and IIS-1763981, NIH grant NICHD 1R21 HD93912-01A1, the Partner University Fund, and the SUNY2020 Infrastructure Transportation Security Center.

REFERENCES