

Learning How to Smile: Expression Video Generation with Conditional Adversarial Recurrent Nets

Wei Wang, Xavier Alameda-Pineda, *Senior Member IEEE*, Dan Xu, Elisa Ricci,
and Nicu Sebe, *Senior Member IEEE*

Abstract—While several research studies have focused on analyzing human behavior and, in particular, emotional signals from visual data, the problem of synthesizing face video sequences with specific attributes (*e.g.* age, facial expressions) received much less attention. This paper proposes a novel deep generative model able to produce face videos from a given image of a neutral face and a label indicating a specific facial expression, *e.g.* spontaneous smile. Our framework consists of two main building blocks: an image generator and a frame sequence generator. The image generator is implemented as a deep neural model which combines generative adversarial networks and variational auto-encoders, while the sequence generator is a label-conditioned recurrent neural network. In the proposed framework, given as input a neutral face and a label, the sequence generator outputs a set of hidden representations with smooth transitions corresponding to video frames. Then, the image generator is used to decode the hidden representations into the actual face images. To impose that the net generates videos consistent with the given label, a novel identity adversarial loss is proposed. Our experimental results demonstrate the effectiveness of the framework and the advantage of introducing an adversarial component into recurrent models for face video generation.

Index Terms—video generation, gated recurrent unit, smile.

I. INTRODUCTION

THE automatic recognition of facial expressions [1]–[4], as one of the most prominent human social signals [5], has been widely studied in computer vision. In particular, the problem of analyzing facial expressions in videos is typically addressed with discriminative approaches, which aim to learn classification models able to distinguish between sequences belonging to different categories. Often, these approaches are specifically designed to analyze the visual facial dynamics of different expressions. Although the performance of such methods is impressive, they do not possess the ability to

Conditioning label (**posed**, creepy, nervous, spontaneous, over-reacting)

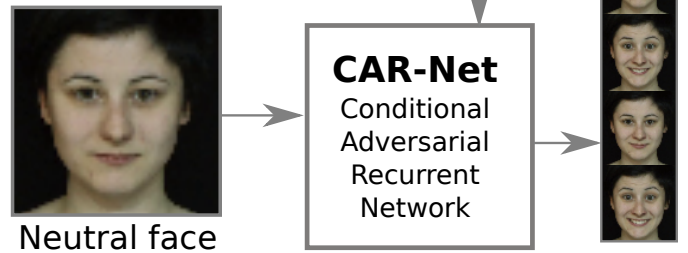


Figure 1: The proposed conditional adversarial recurrent networks (CAR-Nets) generate a sequence of a person smiling from a neutral face while preserving the identity and conditioned to a given label (*e.g.* posed).

reproduce the visual dynamics associated to, for instance, a smile. Bridging this gap, *e.g.* learning how to generate smiling face videos, is the motivation of the present study.

Image generation has become popular in the past few years, mostly thanks to the proliferation of deep neural architectures, and in particular of generative adversarial networks (GAN) [6], [7] and variational auto-encoders (VAE) [8]. The literature on video generation is less populated. Few studies focus on video generation [8], [9]. Even GAN and VAE have exhibited promising results in many image and a few video generation tasks, it is still unclear how to address the generation of videos containing human dynamics, such as facial expressions. Intuitively, this is mainly due to the fact that the generated frames need to be realistic not only *individually*, but also *collectively*, meaning that the transitions between adjacent frames need to be smooth.

Generating facial expression videos would have a positive impact in different fields. For instance, face verification systems would have much more data to be trained on, thus increasing their robustness to noise and outliers and reducing their failure chances. Likewise for facial expression recognition systems, an expression generation framework would reduce not only the time to collect data, but also to annotate it, which is highly desirable in the ongoing deep learning era. Importantly, systems implementing artificial and virtual agents impersonated by an avatar would also clearly benefit from an expression generating framework that produces realistic outputs, both steady images and video sequences. For instance, one may want to be able to generate spontaneous smiles, over-

Wei Wang is with CVLab, EPFL, BC 307, Station 14, CH-1015 Lausanne.
Xavier Alameda-Pineda is with INRIA, Perception Team, Grenoble, 38334, FR.

Dan Xu is with VGG group, University of Oxford. Parks Road, Oxford, OX1 3PJ, UK.

Elisa Ricci is with University of Trento & Fondazione Bruno Kessler, Trento, 38123, IT.

Nicu Sebe is with University of Trento, Italy, 38123 IT.
e-mail: wei.wang@epfl.ch, danxu@robots.ox.ac.uk, niculae.sebe@unitn.it, xavier.alameda-pineda@inria.fr, eliricci@fbk.eu.

Xavier Alameda-Pineda acknowledges the support from the IDEX of Université Grenoble Alpes and from the Agence Nationale de la Recherche on the project ML3RI.

Manuscript received December 27, 2018; revised July 14, 2019 and September 16, 2019; accepted December 21, 2019.

reacting smiles, nervous smiles, posed smiles, creepy smiles, etc. Generally speaking, the developed methodology must be able to condition the output to be a facial expression that follows certain subtle requirements.

In this paper, we propose a novel approach for generating videos of smiling people. Specifically, we introduce a framework which generates a sequence of images (i) both individually and collectively realistic, (ii) allowing a smooth transition from the neutral original image into a smiling image, (iii) satisfying certain subtle conditioning characteristics (e.g. posed expression), and (iv) preserving the face identity. To that aim, we first learn a face manifold representation exploiting recent advances on VAEs and GANs [10]. We then learn how to generate posed and spontaneous smiles by means of a new architecture named *Conditional Recurrent Adversarial Network* (CAR-Net, see Figure 1). The recurrent network learns the statistical behavior of a smiling sequence when represented in the face manifold, and thus learns how to generate a sequence of smiling embeddings starting from a neutral (original) face image. By temporally averaging the generated face embeddings and comparing them to the averaged sequence of real face embeddings, we are able to discriminate between real and fake smiling videos while preserving the identity. Our approach is evaluated on three publicly available datasets, namely the UvA-NEMO Smile [11], the DISFA [12] and DISFA+ [13] datasets. The reported experiments demonstrate that the proposed method is effective at generating videos of smiling expressions and that the generated sequences can be successfully exploited for improving the performance of discriminative models detecting spontaneous and posed smiles.

II. RELATED WORK

In the last few years, deep generative models such as generative adversarial networks [6] and variational auto-encoders [14] exhibited spectacular results in image generation. A combination of VAEs and GANs has been also proposed [10]. Both GANs and VAEs have been applied to the problem of face synthesis. For instance, in [14] Hou *et al.* proposed a variation of traditional VAEs which considers a perceptual loss and demonstrated that this model is effective for capturing the information that encodes facial expressions. Yan *et al.* [15] addressed the task of generating images from visual attributes and demonstrated that their approach can be used to synthesize faces considering attributes like age, gender or expressions. Similarly, in [16] a two-stage deep generative model is proposed for transferring specific attributes to faces. Conditional GAN has also been proposed to generate faces with arbitrary poses [17] and expression [18]. In these works, the conditioning labels are denoted by one hot vectors with only one item being 1 which represents the label while the rest of the items are 0s. Actually, the conditioning vector does not have to be one hot vector, it could also be represented by continuous values such as action unit intensity score corresponds to the expressions [19] or the embeddings of conditioning poses from arbitrary person [20]. Another category of methods for face synthesis is based on morphable models using 3d model [21]–[23]. The 3d model can disentangle the identity

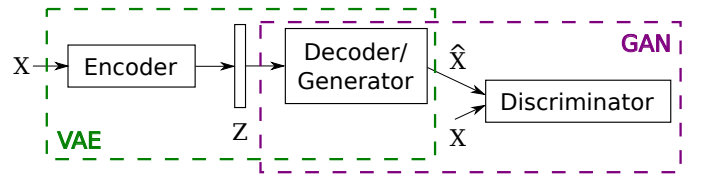


Figure 2: VAE-GAN structure. Image X is *encoded* to a hidden variable Z and then *decoded* back to a real image. The decoding process corresponds to the *generator* of the adversarial structure, competing with the *discriminator* to make the generated image \hat{X} more realistic.

and expression of the person. In such way, the expressions could be transferred from the source image to the target image. However, these methods are usually time consuming as they need to use the 3D model to render the image and the creation of the 3D model is very time consuming.

Recently, deep networks have been also explored for video generation [24]–[27]. In order to model temporal dependencies, previous works either employ a spatio-temporal network which synthesizes all the frames simultaneously, or they take advantage of recurrent neural networks (RNNs) to generate images sequentially [28], [29]. A 3D spatio-temporal DCGAN [30] is employed to generate videos from scratch in [24]. Similarly, in [25] a deep network is proposed to generate all the video frames at the same time. Other works have considered long-short term memory (LSTM) networks [8] and convolutional LSTMs [26] to predict a set of future frames. However, these methods cannot generate videos conditioned on specific labels and none of these methods impose that the generated videos should be realistic, *i.e.* there is no adversarial component in their pipelines, and most of these approaches [24] do not consider conditioning category labels but require different models for different categories.

III. GENERATING SMILES WITH CAR-NETS

Our framework first learns a compact representation of the face manifold and the corresponding embedding from the image space, to later on learn the statistical dynamics of smiling sequences in this compact representation space. This strategy has one prominent advantage: reducing the computational complexity to learn the dynamics.

Since utilizing a good compact representation is crucial to our success, we inspire from the powerful VAE-GAN framework, recently introduced in [10] and sketched in Figure 2. Let us assume the existence of a set of training video sequences $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_n, \dots, \mathbf{X}_N\}$, where $\mathbf{X}_n = [\mathbf{X}_{n0}, \dots, \mathbf{X}_{nT_n}]$ denote the frames of the n -th sequence and T_n is the number of frames. The VAE-GAN is trained with all the frames of the training set, and produces a set of compact representations, denoted by $\mathbf{Z} = \{\mathbf{Z}_1, \dots, \mathbf{Z}_n, \dots, \mathbf{Z}_N\}$, where each element of \mathbf{Z} is a sequence of embeddings: $\mathbf{Z}_n = [\mathbf{Z}_{n0}, \dots, \mathbf{Z}_{nT_n}]$ corresponding to the original sequence \mathbf{X}_n . The details of the structure of the adopted VAE-GAN and on the associated training procedure are provided in the experimental section.

The proposed conditional adversarial recurrent network (CAR-Net) is specifically designed to address the challenging task of learning how to generate a sequence of T embeddings

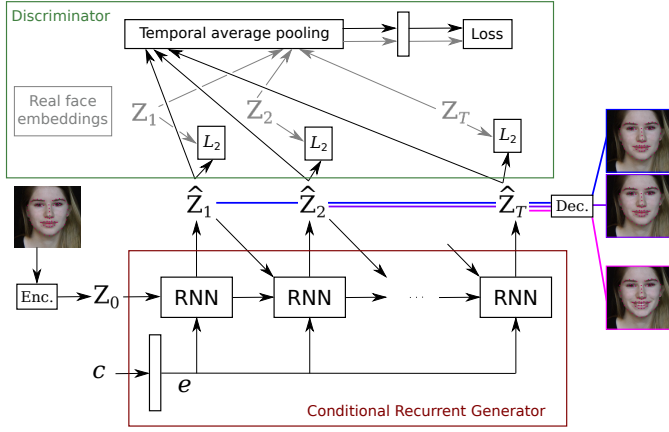


Figure 3: CAR-Net flow diagram. The input image is embedded into the face manifold (\mathbf{Z}_0). The *conditional recurrent generator* produces a sequence of face embeddings $\hat{\mathbf{Z}}_1, \dots, \hat{\mathbf{Z}}_T$. The identity adversarial loss combines (i) a pair-wise L_2 norm ensuring that the smiling dynamics are respected with (ii) a discriminative loss on the sequence’s temporal average that guarantees the sequence preserves the identity.

$\hat{\mathbf{Z}}^{I,c} = [\hat{\mathbf{Z}}_1^{I,c}, \dots, \hat{\mathbf{Z}}_T^{I,c}]$ from an input neutral face image \mathbf{I} and a conditioning label (i.e. posed/spontaneous) c , that correspond to the person in \mathbf{I} smiling and that are individually as well as collectively realistic. Once this sequence of embeddings is satisfactorily generated, the decoder of the VAE-GAN structure is used to generate the sequence of face images: $\hat{\mathbf{X}}^{I,c} = [\hat{\mathbf{X}}_1^{I,c}, \dots, \hat{\mathbf{X}}_T^{I,c}]$.

CAR-Net (see Figure 3) is constructed with a recurrent generator conditioned to a label c and a video discriminator. On the one hand the discriminator aims to make the difference between generated and real sequences, with special emphasis on preserving the identity and on the transitions’ smoothness. On the other hand, the generator tries to fool the discriminator from the input image and conditioning label. Remarkably, the generator possesses two contributions.

First, the generator produces a *sequence* from a still image. Second, the generator is *conditioned* on a user-chosen input label c , selecting between different types of smile (e.g. posed, spontaneous, etc.). More formally, given the embedding of the initial face image \mathbf{Z}_0^I and the conditioning label c , the generator produces a sequence of embeddings $\hat{\mathbf{Z}}^{I,c} = [\hat{\mathbf{Z}}_1^{I,c}, \dots, \hat{\mathbf{Z}}_T^{I,c}] = \mathbf{G}(\mathbf{Z}_0^I, c)$. In the following we first describe the proposed identity adversarial loss, to later on explain how it is co-articulated with the conditional recurrent generator.

A. Identity Adversarial Loss

As outlined before, the discriminator needs to penalize those sequences that (i) either do not follow the dynamics of a smiling sequence or (ii) the face identity changes along the sequence. We remark that, while the dynamics of the sequence change over time, the identity is the same, and therefore the final loss must be a combination of two losses. Our intuition is that, a smiling sequence of one person will draw a certain path on the face manifold *around* the identity of person. We argue that the temporal average of an identity-preserving sequence of face embeddings follows a different pattern that

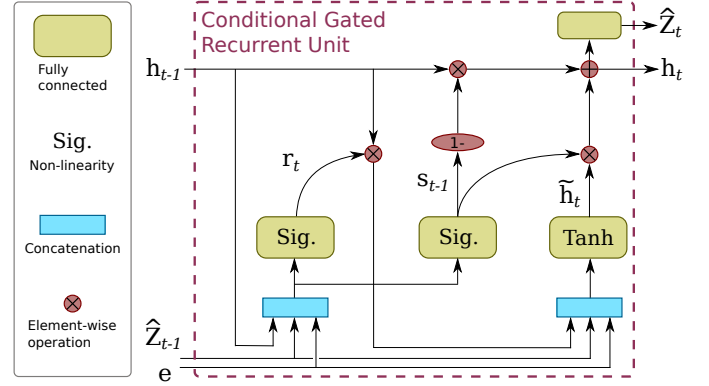


Figure 4: Structure of the conditional gated recurrent unit (Con-GRU). The hidden state \mathbf{h}_t is conditioned by not only by \mathbf{h}_{t-1} and $\hat{\mathbf{Z}}_{t-1}$, but *also* by \mathbf{e} , the representation of c . Red blocks indicate element-wise operations, yellow blocks indicate fully connected layers with non-linearities and blue blocks indicate vector concatenation.

the temporal average of a sequence of face embeddings that does not preserve the identity. It is therefore natural to try to discriminate between these two patterns. However, this is insufficient to force the sequence generator to learn how to encompass the dynamics of smiling sequences. In order to address this issue, we further push the discriminator to penalize those sequences that do not follow the required dynamics with a pair-wise loss between the frames of the generated and real embedding sequences.

In practice, the identity adversarial loss (IAL) is a combination of a cross-entropy loss discriminating between generated and real sequences and a pair-wise loss enforcing the natural dynamics. In order to train the network with this loss and different values of the conditioning c , we require a set of training sequences per each different value of c . For ease of notation, we will assume that all training sequences are available for all values of c , although this is not a requirement of the method, and write \mathbf{Z}^c to denote the training embeddings for the label c . Analogous notation holds for the n -th training sequence \mathbf{Z}_n^c and its frames \mathbf{Z}_{nt}^c . Formally, the identity adversarial loss has the following expression:

$$J_{\text{VAL}} = \sum_{c,n=1}^{C,N} \log \mathbf{D}(\bar{\mathbf{Z}}_n^c) + \log(1 - \mathbf{D}(\bar{\mathbf{Z}}_n^c)) + \sum_{t=1}^{T_n} \|\mathbf{Z}_{tn}^c - \hat{\mathbf{Z}}_{tn}^c\|_2, \quad (1)$$

where $\bar{\mathbf{Z}}$ denotes the temporal averaging of the sequence \mathbf{Z} and \mathbf{D} denotes the discriminator. Our discriminator consists on two fully connected layers (with tanh and σ non-linearities respectively).

$$\mathbf{D}(\bar{\mathbf{Z}}) = \sigma(U_2 \tanh(U_1 \bar{\mathbf{Z}})) \quad (2)$$

where U_1 and U_2 are the weights of the two layers, and the bias terms are absorbed in U_1 and U_2 .

Remarkably, the first two terms of (1) correspond to the identity-preserving cross-entropy loss and the third terms forces the conditional recurrent generator, whose structure is discussed in the next section, to learn the smile dynamics.

B. Conditional Recurrent Generator

The generator is designed to recurrently produce a sequence of outputs that are conditioned on the input label c . We construct the conditional recurrent generator from a building block named conditional gated recurrent units (Con-GRU).

A functional diagram of the Con-GRU, as used in this study, is shown in Figure 4, where we can see that the entire recurrent process is conditioned on \mathbf{e} , which is the internal learned representation of the label c . Indeed, \mathbf{e} is the output after two fully connected layers:

$$\mathbf{e} = \tanh(V_2 \tanh(V_1 c)), \quad (3)$$

where V_1 and V_2 denote the weights of the fully connected layers. The bias terms are absorbed in V_1 and V_2 . In addition to \mathbf{e} , the generated face embedding at time t , $\hat{\mathbf{Z}}_t$, is conditioned, not only to the hidden recurrent state, \mathbf{h}_{t-1} as in standard recurrent networks, but also to the previously generated face embedding $\hat{\mathbf{Z}}_{t-1}$. Generally speaking we write:

$$\mathbf{h}_t = f(\mathbf{h}_{t-1}, \hat{\mathbf{Z}}_{t-1}, \mathbf{e}), \quad (4)$$

where f denotes the non-linearity associated to the Con-GRU that will be detailed in the following.

Gated recurrent units in general (and Con-GRUs in particular) rely on an intermediate representation of \mathbf{h}_t , denoted by $\tilde{\mathbf{h}}_t$ and a selecting variable \mathbf{s}_t . Each component of \mathbf{h}_t is copied from \mathbf{h}_{t-1} if the corresponding entry in \mathbf{s}_t is set to 0, or copied from $\tilde{\mathbf{h}}_t$ if the corresponding entry in \mathbf{s}_t is set to 1. In plain words, when the entries of \mathbf{s}_t are set to 1, the network forgets the knowledge at previous time steps, and focuses on the newly generated representation $\tilde{\mathbf{h}}_t$. Otherwise, it keeps the previous representation \mathbf{h}_{t-1} . Formally:

$$\mathbf{h}_t = (1 - \mathbf{s}_t) \otimes \mathbf{h}_{t-1} + \mathbf{s}_t \otimes \tilde{\mathbf{h}}_t, \quad (5)$$

where \otimes indicates the element-wise product,

$$\tilde{\mathbf{h}}_t = \tanh(W_{hr}\mathbf{r}_t \otimes \mathbf{h}_{t-1} + W_{hz}\hat{\mathbf{Z}}_{t-1} + W_{he}\mathbf{e}), \quad (6)$$

$$\mathbf{r}_t = \sigma(W_{rh}\mathbf{h}_{t-1} + W_{rz}\hat{\mathbf{Z}}_{t-1} + W_{re}\mathbf{e}), \quad (7)$$

and σ indicates the sigmoid function. We highlight the explicit dependency on the internal representation of the conditioning label \mathbf{e} .

The selecting variable explicitly depends on the conditioning label as well:

$$\mathbf{s}_t = \sigma(W_{sh}\mathbf{h}_{t-1} + W_{sz}\hat{\mathbf{Z}}_{t-1} + W_{se}\mathbf{e}), \quad (8)$$

which concludes the definition of f . Once \mathbf{h}_t is computed, the face embedding is generated through a linear layer:

$$\hat{\mathbf{Z}}_t = W_{zh}\mathbf{h}_t. \quad (9)$$

Summarizing, the parameters of the generator are the weights of the Con-GRU, plus the two fully connected layers transforming c into \mathbf{e} . During training, we need to optimize both the parameters of the generator and the discriminator.

C. CAR-Net training

As for any adversarial network, training CAR-Nets requires the estimation of the parameters of both the generator θ_G and the discriminator θ_D . The three terms of the identity adversarial loss in 1 depend on θ_G and θ_D in a different manner. Indeed, the operation \mathbf{D} depends on θ_D and the generated sequence $\hat{\mathbf{Z}}_n^c = \mathbf{G}(\mathbf{Z}_{n0}, c)$ depends on θ_G . In addition, since the generator and the discriminator are competing, the first tries to minimize the loss, while the second tries to maximize it. We write the optimization of θ_G as:

$$\min_{\theta_G} \sum_{c,n=1}^{C,N} \log(1 - \mathbf{D}(\overline{\mathbf{G}(\mathbf{Z}_{n0}, c; \theta_G)}; \theta_D)) + \ell(\mathbf{Z}_n^c, \mathbf{G}(\mathbf{Z}_{n0}, c; \theta_G)), \quad (10)$$

where ℓ is the sum of L_2 distances between generated and real sequences. The optimization problem of the discriminator writes:

$$\max_{\theta_D} \sum_{c,n=1}^{C,N} \log \mathbf{D}(\overline{\mathbf{Z}}_n^c; \theta_D) + \log(1 - \mathbf{D}(\overline{\mathbf{G}(\mathbf{Z}_{n0}, c; \theta_G)}; \theta_D)). \quad (11)$$

The discriminator and generator compete with each other and, in order to balance this competition during training, their respective parameters θ_G and θ_D are alternatively updated. Intuitively, if one of them is too strong, we will further iterate the update of the other, until meeting a certain stopping criterion. In order to formalize this, we inspire from [7] and define two internal performance-monitoring measures:

$$\rho_R = - \sum_{c,n=1}^{C,N} \log \mathbf{D}(\overline{\mathbf{Z}}_n^c), \quad \rho_G = - \sum_{c,n=1}^{C,N} \log(1 - \mathbf{D}(\overline{\mathbf{Z}}_n^c)). \quad (12)$$

Intuitively, when the error over the real or generated sequences is too low, i.e. the discriminator is doing very well, we skip training the discriminator. Analogously, when the error is too big we skip training the generator. We just need to be careful to avoid the situation in which neither of them is trained. Therefore, the strategy is the following. If τ_D and τ_G are two boolean variables indicating whether or not the discriminator and the generator respectively need to be updated, we initialize $\tau_D = \tau_G = 1$ and at every iteration recompute them as:

$$\begin{cases} \tau_D = 0, & \text{if } \rho_R < \gamma, \rho_G < \gamma \\ \tau_G = 0, & \text{if } \rho_R > 1 - \gamma, \rho_G > 1 - \gamma \\ \tau_D = \tau_G = 1 & \text{if } \tau_D = \tau_G = 0, \end{cases} \quad (13)$$

where γ being a hysteresis parameter, set to $\gamma = 0.3$ as in [7].

IV. EXPERIMENTAL VALIDATION

A. Experimental setup

In this section we first discuss the smile video datasets, then the pre-processing procedure to extract smiling videos. We also briefly discuss the still image datasets used, together with the video datasets, to train the VAE-GAN, and finally detail the train-test splitting protocol.

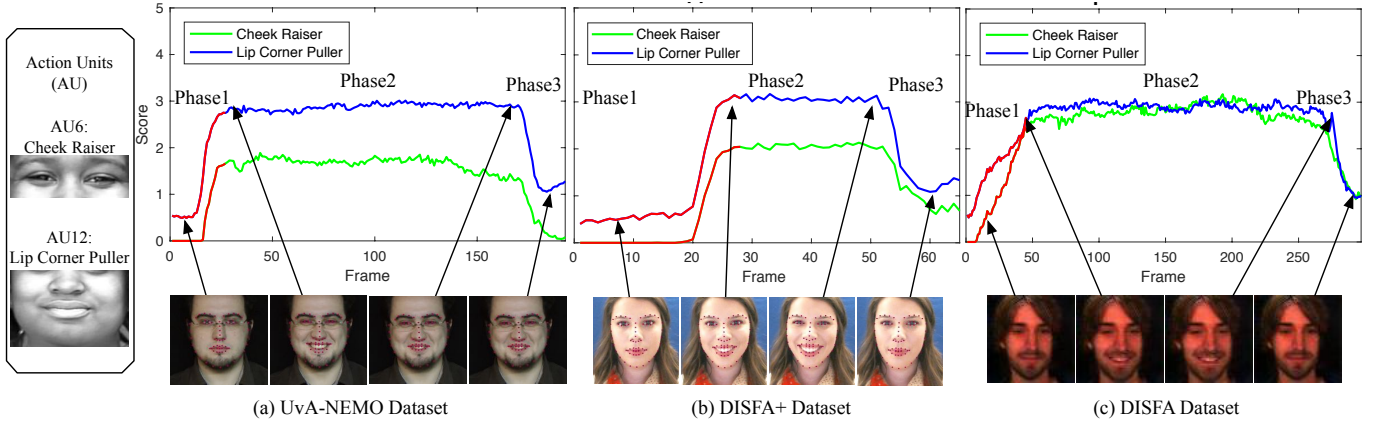


Figure 5: Action unit score *w.r.t.* the facial dynamic. The ‘cheek raiser’ and ‘lip corner puller’ units reflect the three main smiling phases. The action unit images are extracted from <https://www.cs.cmu.edu/~face/facs.htm>. The video frames from the UvA-NEMO and DISFA+ datasets follow the three phase strictly. However, each video in DISFA dataset contains multiple expressions. Therefore, a preliminary segmentation is required to identify segments corresponding to smiles.

1) *Smile Video Datasets*: The UvA-NEMO dataset [11], the DISFA dataset [12] and the DISFA+ dataset [13] are used to validate the proposed CAR-Net for smile video generation.

The UvA-NEMO dataset contains 1240 smile videos (643 posed videos and 597 spontaneous videos). There are 400 subjects in the videos (215 male and 185 female) with different ages which range from 8 to 76. Among all people, 50 wear glasses. The videos have a resolution of 1920×1080 pixels sampled at 50 FPS. Each video starts and ends with a neutral expression. The average duration is 3.9 seconds.

The DISFA dataset is a spontaneous facial expression dataset. It contains videos of 27 subjects (12 females and 15 males) with different ethnicities. The videos are recorded with a resolution of 1024×768 at 20 FPS. Each subject has been recorded for 4 minutes and each video includes multiple expressions, such as happiness, disgust and surprise. As our work specifically focuses on smile expression, we manually segmented the videos to isolate the short sequences containing spontaneous smiles (see Subsection IV-A2). After segmentation, we obtain 17 videos.

The DISFA+ dataset is an extension of DISFA. It contains posed smile sequences for 9 individuals from the DISFA dataset. Different from DISFA, the DISFA+ dataset provides 20 video sequences associated to smiles, thus a procedure for filtering out data corresponding to other facial expressions is unnecessary. While our method is generic, we present experiments considering two types of smiles: posed and spontaneous. While the UvA-NEMO dataset includes sequences for both categories, for the DISFA and DISFA+ we build a small dataset pairing sequences of corresponding targets.

2) *Extracting smiling face videos*: The targeted application requires an automatic procedure to extract smiling sequences from face videos. Indeed, given a video we need to (i) segment the part in which the person is smiling and (ii) spatially align the sequences to learn the smile dynamics.

To overcome the first issue, we rely on action units [31]. Indeed, smiling mostly involves two action units: *cheek raiser* and *lip corner puller*. By employing [32], we are able to estimate the intensity of the two action units, that we plot in Fig. 5. We can clearly see that smiling is well correlated with these

two action units, specially with *lip corner puller*. We used this action unit to segment the subsequence corresponding to the moving from neutral to smile (first phase in Fig. 5).

Once the smiling videos are segmented, we need to align the faces. Many face alignment algorithms are available [33]–[36]. In this paper, we align the faces with respect to the center of the two eyes horizontally and to the vertical line passing through the center of the two eyes and the mouth using OpenFace [37].

After applying the smiling face video extraction procedure described in the previous section, we notice that in the UvA-NEMO dataset the average smile length is 32 frames, with tiny variations. In our experiments we used this value as the length of the generated sequences.

3) *Face Image Datasets*: In order to train a robust and generic VAE-GAN for face encoding and generation, we used five different still face image datasets, namely: CelebA, CornellKin, Family101, TSKinFace and UBKin, in addition to the previous three video datasets (NEMO-UvA, DISFA and DISFA+). CelebA consists of 202,599 images from 10,177 different identities, with large head pose and occlusion variability. CornellKin contains frontal face images of 150 pairs of public figures and celebrities along with their parents/children. Family101 contains 101 different families with distinct family names, including 206 nuclear families, 607 individuals, and 14,816 images in total. TSKinFace collects almost 1,000 family photos with different family grouping (Father-Mother-Daughter or Father-Mother-Son). Finally, UBKin contains images of 400 people with child-parent or parent-grand parent pairs. Overall, training with such variety of images improves the robustness of the proposed method to unseen images.

4) *Train and Test Setup*: The video dataset, UvA-NEMO, DISFA and DISFA+ are split into train and test sets. The train split is used both for the VAE-GAN training (together with other face image datasets) and for training the recurrent block of the proposed method. All images are resized to 64×64 for training the VAE-GAN. In case of the UvA-NEMO dataset we follow the splitting protocol of [11] and use 1 tenth for test and the rest for training. For the DISFA and DISFA+ sequences, we randomly select two thirds for training and the rest for test.

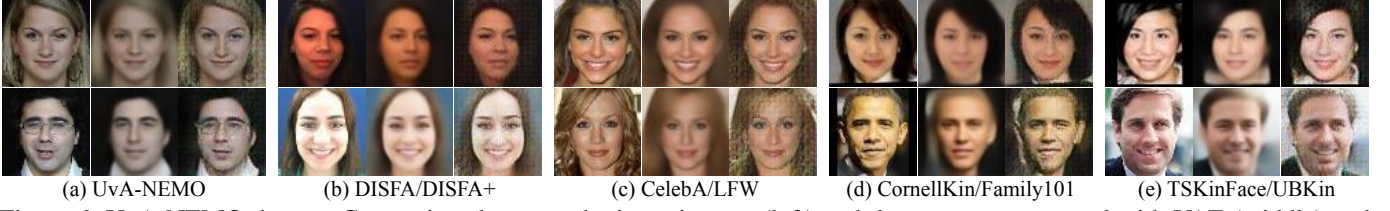


Figure 6: UvA-NEMO dataset. Comparison between the input images (left) and the ones reconstructed with VAE (middle) and VAE-GAN (right). VAE-GAN can reconstruct sharper images with detailed information compared with the VAE.

B. CAR-Net architecture and optimization

As stated above, while our method is generic, we present experiments considering two types of smiles: pose and spontaneous. These two conditioning classes are represented in two-dimensional binary vectors which are transmitted to the Con-GRUs through two fully connected layers of output dimension 512 and 1024 respectively. The internal state of the recurrent units is of dimension $\mathbf{h}_t \in \mathbb{R}^{1024}$, and we set \mathbf{h}_0 to zero. The face embedding representation is of dimension 512 (the precise VAE-GAN architecture is detailed below). Consequently, the parameters of the Con-GRU have dimensions as follows: $W_{hr}, W_{rh}, W_{sh} \in \mathbb{R}^{1024 \times 1024}$, $W_{hz}, W_{rz}, W_{sz}, W_{he}, W_{re}, W_{se}, W_{zh} \in \mathbb{R}^{512 \times 1024}$. Finally, the two fully connected layers have output dimension 256 and 1 (generated/real classification). We employ stochastic optimization with Adam [38], with a base learning rate of 0.001 and an exponential decay of 0.95 every 500 iterations.

1) *VAE-GAN architecture*: The VAE-GAN consists of 3 parts: the encoder, the decoder/generator and the discriminator. The *encoder* has 3 convolutional layers with kernel size of 5 and 64, 128 and 256 output channels respectively; plus a fully connected layer with 2048 output units. The encoder is a probability model which is forced to produce latent vectors that follows the Gaussian distribution. Given an input image, the encoder outputs μ and σ which represent the mean and variance of a Gaussian distribution. Then the hidden representation z can be computed by, $z = \mu + \exp(\sigma) * e$, where e is a reparameterization value which samples a latent vector z from the distribution. Next, the latent vector will be fed to the decoder to generate images. e is a randomly sampled variable which follows Gaussian distribution with 0 mean and unit standard deviation. The *decoder/generator* has one fully connected layer with $256 \times 8 \times 8$ output units, reshaped into a tensor with 256 channels. This layer is followed by 4 deconvolutional layers of kernel size 5 and 256, 128, 64, 3 output channels. The *discriminator* has a symmetric structure with respect to the generator (4 convolutional layers plus a fully connected layer), followed by the final real/generated classification layer. Optimization is done as in [10].

C. Qualitative Analysis

We first provide some qualitative results on the UvA-NEMO dataset and show the effect of the main components *i.e.* the VAE-GAN component, the adversarial component, and the conditional recurrent component. We also demonstrate that our framework is flexible and it is able to generate very realistic sequences by exploiting the data collected in the wild.

1) *VAE-GAN component*: Figure 6 shows the reconstructed images generated with VAE and VAE-GAN. As expected (and in agreement with the findings in [10]) the images obtained with VAE-GAN have higher quality and contain more clear details than the ones reconstructed using VAE. For instance, the wrinkles (Fig. 6, row 1, column 1), teeth (Fig. 6, row 2, column 2), and hair (Fig. 6, row 2, column 3) can be better reconstructed with VAE-GAN. Our results confirm the interest of adding a discriminator on top of the VAE, and of exploiting this powerful framework to construct an efficient face embedding for the targeted application.

2) *Adversarial component*: We compare three different version of CAR-Net, the standard, the one without the identity adversarial loss (CAR-Net no-ID) and a CAR-Net whose discriminator acts only in the last frame of the RNN instead of on the temporal average (CAR-Net ID-LR).

Figure 7 shows generated examples for the UvA-NEMO (top blocks) and the DISFA/DISFA+ (bottom blocks) datasets, for spontaneous (left) and posed (right) smiles. Each block shows a few frames along the sequence, and each row corresponds to one method. From Figure 7, we can observe that both *CAR-Net ID-LR* and *CAR-Net no-ID* can not well preserve the identity. In the following qualitative analysis section, we have implemented a user study to evaluate whether the identity can be well preserved for each method. Besides, we have also viewed the identity distance between the face in the first frame and other generated frames using FaceNet.

3) *Conditional recurrent component*: To our knowledge, there are no existing large-scale video generation models which can be conditioned on different labels. For instance, the recent VideoGAN approach in [24] can only generate videos given an input frame but does not employ conditioning labels. However, in order to perform a comparison with [24], in this work we train two video-GANs which correspond to two different smiling labels. In addition to VideoGAN we also compare the proposed CAR-Net with the Ground Truth and with the Interpolated Sequence. Note that the interpolated sequence requires the initial and final frames, and therefore has far more input information than CAR-Net and VideoGAN.

From Figure 7, we notice that VideoGAN generates sequences with visual artifacts, leading to very unnatural images. Interpolating the sequence may seem a good idea at first sight, but we rapidly understand that works well only in sequence in which the smile dynamics are linear, see Figure 7 (c) and (d). However, it fails to reproduce non-linear dynamics such as in Figure 7 (a) and (b). When a person smiles, his or her action unit score changes in a non-linear manner. As can be seen in Figure 6 (b), in the beginning, the score of AU12 (*i.e.*, Lip corner puller) increases slowly, and

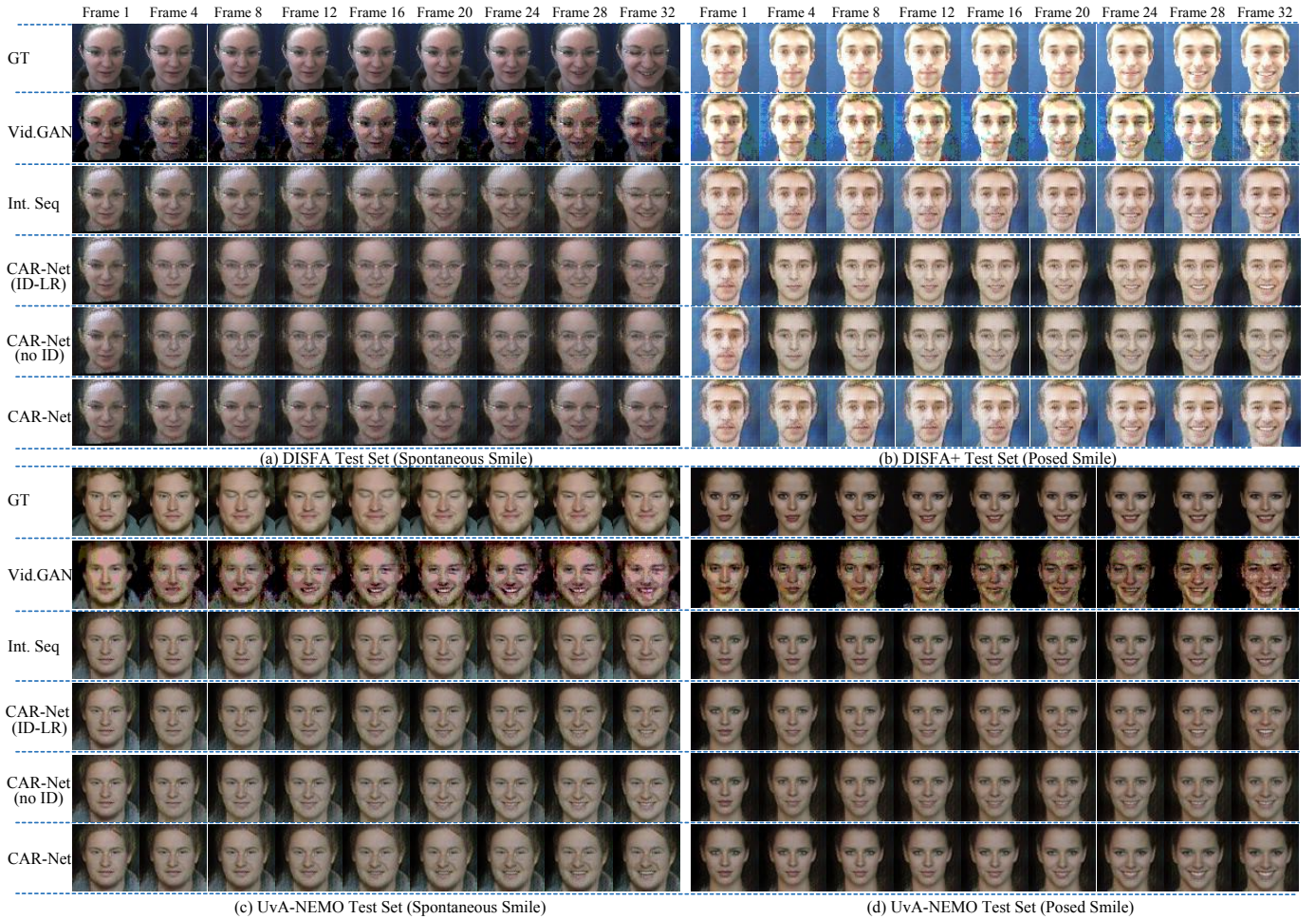


Figure 7: Comparison between the Ground Truth, VideoGAN, Interpolation and CAR-Nets on DISFA/DISFA+ (a) and (b) and UvA-NEMO (c) and (d), for both spontaneous (a) & (c) and posed (b) & (d) smiles.

then at a certain point, the score increases quickly to its peak value. Finally, the score becomes stable and changes slowly until it reaches its peak value. In other words, during the smile process, the action unit, such as the lip corner wont be raised linearly with a constant speed. It changes in a non-linear manner with accelerations. As previously discussed, compared to VideoGAN and CAR-Net, interpolating requires also to have the smile face image. Regarding the different versions of CAR-Net, we observe that the standard one exhibits less artifacts than the other two in all four sequences.

4) *Comparison with other baselines:* To provide further insights on the challenges of the task of automatic smile video generation, we also report additional qualitative results. In particular, we consider the Temporal Generative Adversarial Nets (TGAN) [25] and MoCoGAN [39]. Similarly to the VideoGAN experiments, as TGAN and MoCoGAN does not employ conditioning labels, we train two different networks corresponding to the two smiling labels. It is worth noting that, opposite to VideoGAN, TGAN and MoCoGAN cannot be directly compared with our approach as this model cannot be conditioned on an input image. In other words, TGAN and MoCoGAN can generate a smile sequence but cannot generate a video corresponding to a specific person.

Figure 8 shows the video sequences generated with

TGAN [25] and MoCoGAN [39]. It is easy to observe that all the generated video frames depict a transition from neutral facial expression to smile (spontaneous/posed). Furthermore, by comparing the generated spontaneous smile videos with the posed ones, it is evident that the faces corresponding to posed smiles usually have their mouths wide open, indicating a very exaggerated (and fake) behaviour compared to spontaneous smiles. However, while TGAN and MoCoGAN can learn correct smile patterns, Fig. 8 also shows that the sequences generated by TGAN and MoCoGAN have severe artifacts, most of which propagate to several neighboring frames. The underlying reason is that the discriminator in TGAN and MoCoGAN is a 3D-discriminator on top of the whole sequence. Therefore, even though the generated videos may look realistic from the perspective of the video discriminator, there is no guarantee that each individual frame looks good.

5) *Ability to generalize to realistic conditions:* One limitation of the UvA-NEMO, DISFA and DISFA+ datasets is that these datasets are collected under laboratory conditions. Even if the recurrent generator must be trained with smile videos, the VAE-GAN can be trained on any face image dataset. As previously discussed, we train the VAE-GAN with five face image datasets, in addition to the three video smile datasets. Figure 9 shows the generated spontaneous and posed smile face sequences for people in the LFW dataset. We can see

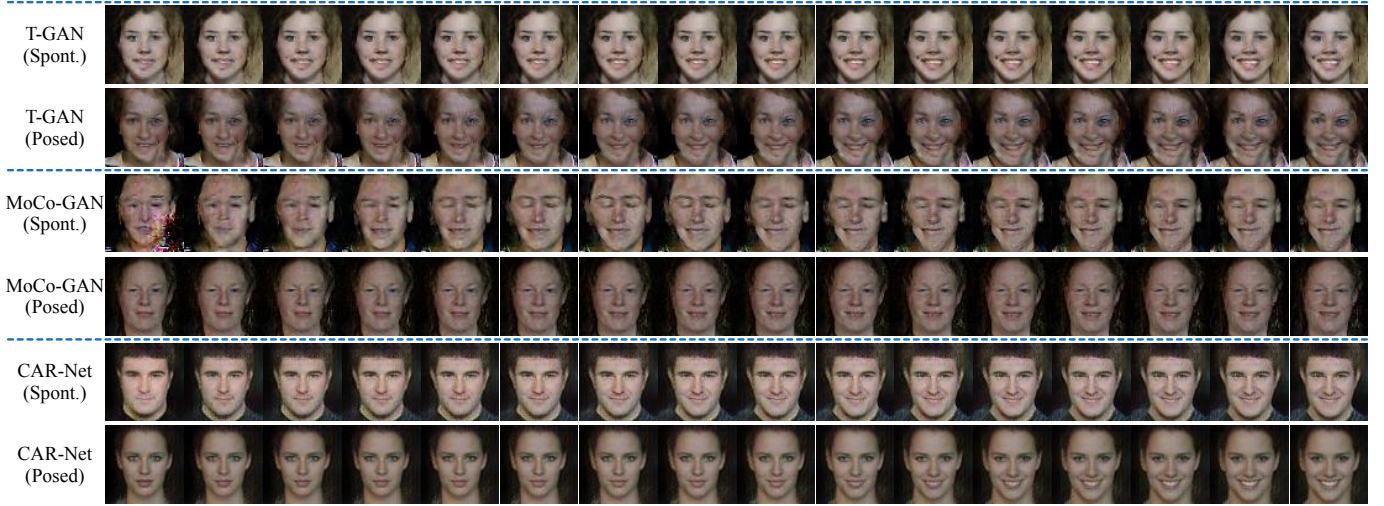


Figure 8: Generated video frames using TGAN [25] for spontaneous (a) and posed (b) smiles.

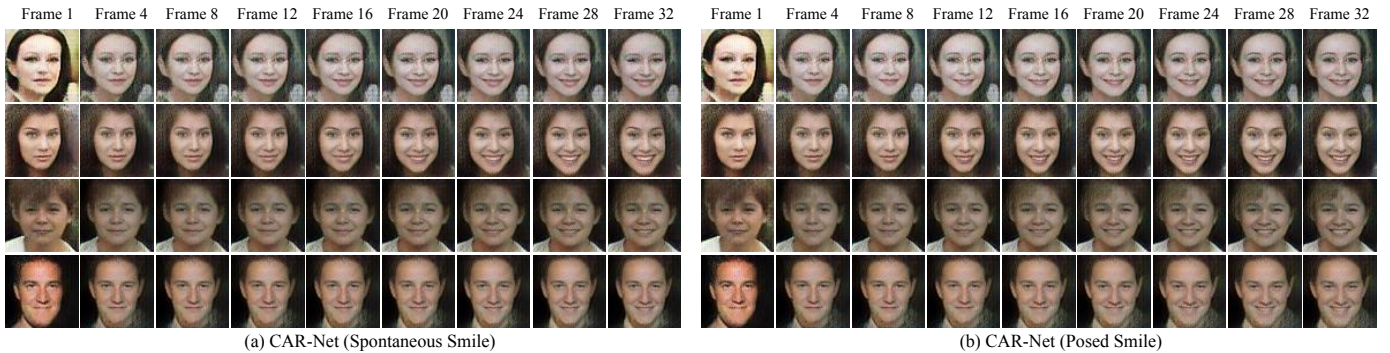


Figure 9: Sequences generated by CAR-Net on images from the LFW dataset (obtained in the wild).

Table I: Percentage (%) of generated video preference.

| Models Preference | Spontaneous Smile | Posed Smile |
|-------------------|-------------------|-------------|
| VideoGAN [24] | 27.72 | 21.62 |
| CAR-Net | 64.73 | 66.80 |
| ~ | 7.55 | 11.58 |
| CAR-Net (no ID) | 14.04 | 16.16 |
| CAR-Net | 53.78 | 48.43 |
| ~ | 32.18 | 35.43 |

that the images generated by CAR-Net are realistic, do not contain artifacts and represent spontaneous and posed smiles. This demonstrates that CAR-Nets successfully generalizes to images collected in the wild.

D. Quantitative Analysis

We conduct three types of quantitative analysis. First, we perform a user-study to compare the videos generated by CAR-Net with those obtained with VideoGAN [24]. Second, we perform an identity preservation test, using a pre-trained face recognition model. Third, we evaluate whether the data generated by CAR-Net could be used to improve the performance of a discriminative model for detecting automatically spontaneous vs posed expressions.

1) *User Study*: Following [24], we quantitatively evaluate and compare CAR-Net with VideoGAN by performing a user study. We prepared 37 video pairs and invited 40 subjects with different backgrounds to do the evaluation. Both our model and baseline model generate videos conditioning on the images

from the test set of UvA-NEMO, DISFA and DISFA+ datasets. Next, we randomly select 37 videos with 18 spontaneous smile videos and 19 posed smile videos generated by our model and the corresponding ones generated by our baseline models. To compare with VideoGAN, we show a subject a pair of videos (one generated by CAR-Net and the other by VideoGAN) and ask her *Which video looks more realistic?*. We collected 1480 ratings from all the subjects. To further demonstrate the effectiveness of the proposed framework, we also collect user ratings comparing video pairs generated by CAR-Net with and without the identity adversarial loss. Again, we show a subject a pair of videos (obtained with the CAR-Net with and without adversarial loss) and ask them *Which video better preserves the identity of the person?*.

Table I shows the preferences expressed by the annotators (%) both for spontaneous and posed smiles. The symbol ~ in Table I means that the two videos are rated as similar. When we compare the CAR-Net with the VideoGAN baseline, most annotators prefer the videos generated by our CAR-Net. This is not surprising: by visually inspecting the frames shown in Fig. 7 it is clear that several artifacts are present in the sequences generated with VideoGAN. Furthermore, when we evaluate the contribution of the identity adversarial loss, we observe that most annotators think that the CAR-Net with identity adversarial loss can better preserve the identity. This is in accordance with qualitative results shown in Fig. 7.

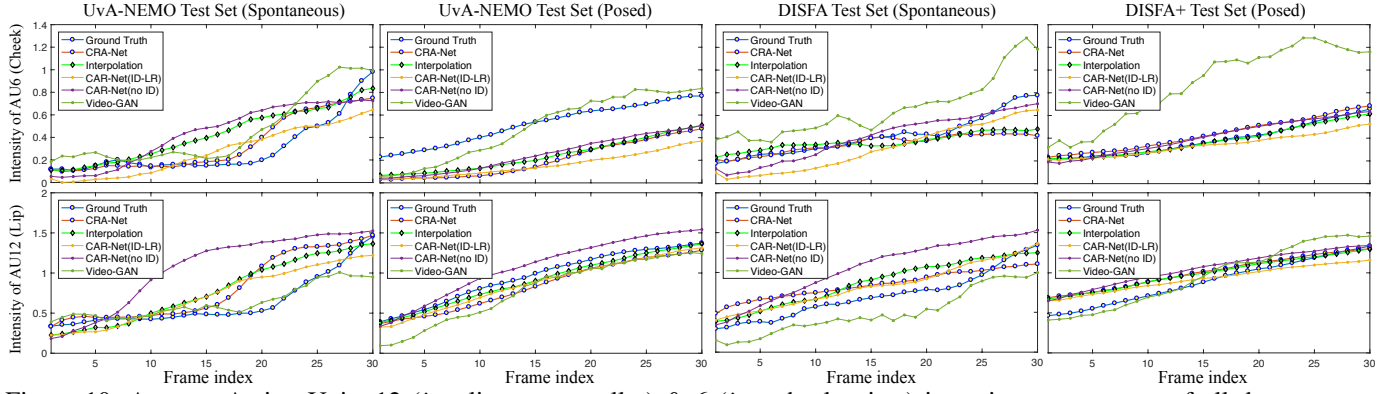


Figure 10: Average Action Units 12 (*i.e.*, lip corner puller) & 6 (*i.e.*, cheek raiser) intensity score curves of all datasets as a function of the video frame. The score is computed as the average of every i -th frame across all test videos.

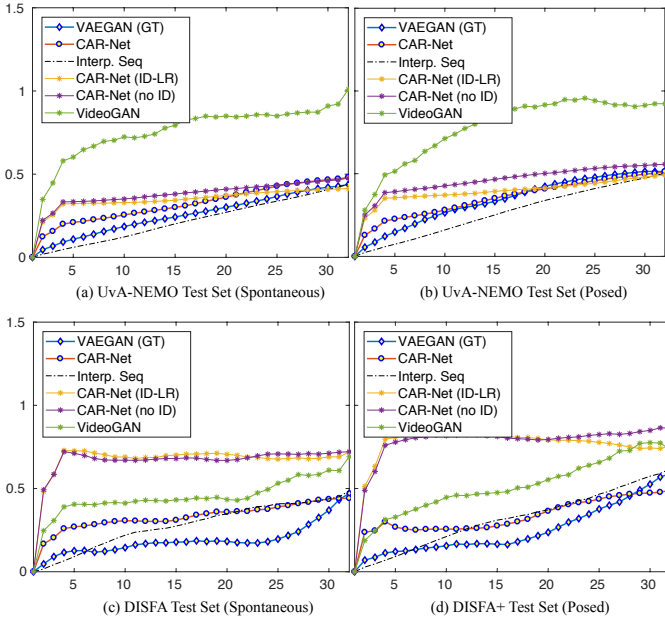


Figure 11: Average distance between the face in the first frame and other frames as seen by FaceNet [40].

2) *Identity preservation test*: We also performed additional experiments to quantitatively assess whether the face identity is preserved along a video sequence. We relied on a pre-trained face recognition model, *i.e.*, FaceNet [40]), which adopts a triplet loss. This loss encourages the face embeddings corresponding to the same identity to be similar, while it pushes the face embeddings associated to different people far away from each other. The Euclidean distance between face embeddings is employed for face recognition. In this work we used the distance between the first frame and all the other frames in the video to verify whether the face identity is preserved. Specifically, for each generated sequence, we first computed the distance between the embeddings of the first frame and the other frames. For each frame index t , we averaged the computed distance over all the sequences of the data set, thus obtaining how close to the original face is the t -th frame of a generated sequence on average. Figure 11 plots the average distances of each model, for both posed and spontaneous smiles. In detail, we benchmark VideoGAN with the two versions of CAR-Net: including and excluding the ID preservation loss (CAR-Net and CAR-Net-noID, respectively).

As shown in Fig. 11, the distance between first frame and the generated frames grows almost monotonically with respect to the frame index. This observation shows that the generated faces lose its identity gradually for all the methods. We can also observe that the VideoGAN loses the identity much faster compared with CAR-Net, and CAR-Net is the slowest one to lose the identity. We also notice that the benefit of using the ID preservation loss is more important for the spontaneous smiles. One possible explanation is that spontaneous smiles are inherently more variate and diverse than pose smiles, and therefore harder to learn. In this regard, CAR-Net seems to perform equivalently in spontaneous and posed smiles, thank to the ID preservation loss. Overall, this experiment demonstrates the importance of ID preservation loss.

3) *Action Unit Recognition*: In order to investigate if CAR-Nets are able to faithfully learn the smile dynamics, we performed an experiment in which we estimate the intensity of AU12 (lip corner puller) & AU6 (cheek raiser). To do that we employ the method in [37], and we average over the entire test set. Results are plot in Figure 10 for UvA-NEMO, DISFA and DISFA+ datasets, for all variants of CAR-Net, VideoGAN, interpolation and the Ground Truth. We first remark that all methods have a monotonically increasing average behavior, which is generally a good property. When comparing the two datasets, we clearly see that the variance on the DISFA/DISFA+ is larger than in UvA-NEMO. This could be an indication that learning smile dynamics on DISFA/DISFA+ is slightly more difficult than on the UvA-NEMO. Comparing the different methods in those curves is a little bit difficult, since most of the methods lie around the average ground truth curve in most of the four plots. Exceptions to these rule are VideoGAN and CAR-Net (no-ID) for UvA-NEMO and DISFA, and mostly every method for DISFA+.

4) *Posed vs Spontaneous classification*:

a) *Smile Classification*: We also conduct an experiment to see if the generated video sequences can be used in the context of posed vs spontaneous smile classification [11]. Similar to a very recent experimental protocol for face analysis [27], we augment the data with synthetic images and train a SVM smile classifier under the three following experimental conditions: using only real data, using only synthetic data and combining both real and synthetic data. In our case, synthetic data is generated with the trained CAR-Net, and we generate

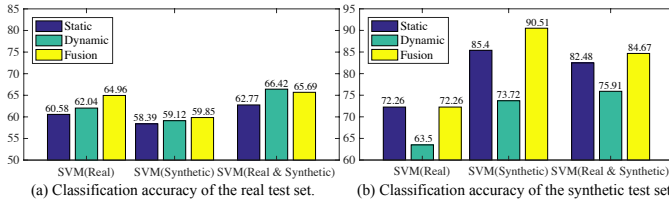


Figure 12: Spontaneous/posed classification accuracy (%) for (a) real joint test sets of UvA-NEMO, DISFA and DISFA+ and (b) synthetic joint test set. The SVM classifiers are trained with different training sets (real, synthetic and both) using different visual features (static, dynamic, fusion).

the exact same amount of synthetic data as we have real data.

In this experiment, we also want to understand how the classification performs when using static features, dynamic features, or fusing both features. To this aim we exploit a simple descriptor inspired by VLAD [41], and a linear SVM. In more detail, we divide each 32 VAE embedding sequence into four groups of 8 VAE embeddings. For each group, we consider the first embedding as the reference embedding, and the video static feature is obtained by concatenating the four group reference embeddings. In order to take into account temporal dynamics, we compute the average difference between embeddings and the reference. The four average differences (one per group) are concatenated to obtain the dynamic features. We then train separately with static and with dynamic features. Late fusion is employed to jointly exploit both static and dynamic descriptors.

Figure 12 (a) reports the classification accuracy under different training conditions. We evaluated the impact of using static, dynamic and fused features, as well as training with real, synthetic, or both data. We can observe that fusion are generally the best performing features (except when training with real and synthetic data), and exploiting dynamic features is systematically better than using static features. In terms of the training set, we see an (expected) decrease of performance when training only with synthetic data, but we also see a clear increase of performance when jointly training with real and synthetic data. This demonstrates one of the advantages of CAR-Nets: the ability of generate synthetic data to increase the performance of face expression classification tasks.

Figure 12 (b) reports the classification accuracy for the joint synthetic test set. Similar to Figure 12 (a), we can observe that fusion are the best performing features for the synthesized test set. This observations have reinforced the statement that CAR-Nets have the ability of generate synthetic data to increase the performance of face expression classification tasks. The augmented synthetic training data always brought performance gain for the synthetic test set. Besides, it is interesting to observe that sometimes if we train the classifier only using the synthetic training data, it can lead to better performance on the synthetic test set. The underlying explanation is that both the synthetic training and test set are generated using the same model (CAR-Nets). Therefore, they have the similar patterns and it is much easier for the classifiers to learn the shared similarities between them.

V. CONCLUSIONS

This paper introduces a novel CAR-Net to address the task of smile video generation from a neutral image and a conditioning label. We first learn the face embedding with the VAE-GAN. To make the VAE-GAN more robust to the images in the wild, we employ five extra face datasets. We then introduce conditional GRU as our basic building block and train them to generate face embedding sequences conditioned on a given label. To preserve the identity, an identity adversarial loss is proposed to confront a discriminator with the conditional recurrent generator. We validate the CAR-Net using the UvA-NEMO, DISFA and DISFA+ datasets, and show extensive qualitative and quantitative results. Further evaluation is conducted on face preservation, action unit score for smile dynamics and in posed/spontaneous smile classification, where we demonstrated the effectiveness of the CAR-Net in data augmentation.

REFERENCES

- [1] G. Zen, L. Porzi, E. Sangineto, E. Ricci, and N. Sebe, "Learning personalized models for facial expression analysis and gesture recognition," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 775–788, 2016.
- [2] S. Y. Park, S. H. Lee, and Y. M. Ro, "Subtle facial expression recognition using adaptive magnification of discriminative facial motion," in *ACM International Conference on Multimedia*, 2015.
- [3] B. Gong, Y. Wang, J. Liu, and X. Tang, "Automatic facial expression recognition on a single 3d face by exploring shape deformation," in *ACM International Conference on Multimedia*, 2009.
- [4] F. Zhang, Q. Mao, M. Dong, and Y. Zhan, "Multi-pose facial expression recognition using transformed dirichlet process," in *ACM International Conference on Multimedia*, 2016.
- [5] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.
- [6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Int. Conf. on Neural Information Processing Systems*, 2014.
- [7] E. L. Denton, S. Chintala, R. Fergus *et al.*, "Deep generative image models using a laplacian pyramid of adversarial networks," in *Int. Conf. on Neural Information Processing Systems*, 2015.
- [8] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *Int. Conf. on Machine Learning*, 2015.
- [9] W. Wang, X. Alameda-Pineda, D. Xu, P. Fua, E. Ricci, and N. Sebe, "Every smile is unique: Landmark-guided diverse smile generation," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2018.
- [10] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Int. Conf. on Machine Learning*, 2016.
- [11] H. Dibeklioğlu, A. Salah, and T. Gevers, "Are you really smiling at me? spontaneous versus posed enjoyment smiles," *Eur. Conf. on Computer Vision*, 2012.
- [12] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.
- [13] M. Mavadati, P. Sanger, and M. H. Mahoor, "Extended disfa dataset: Investigating posed and spontaneous facial expressions," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition Workshops*, 2016.
- [14] X. Hou, L. Shen, K. Sun, and G. Qiu, "Deep feature consistent variational autoencoder," in *IEEE Int. Conf. on Winter App. of Computer Vision*, 2017.
- [15] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2image: Conditional image generation from visual attributes," in *Eur. Conf. on Computer Vision*, 2016.
- [16] M. Li, W. Zuo, and D. Zhang, "Deep identity-aware transfer of facial attributes," 2016, arXiv preprint:1610.05586.
- [17] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2017.

- [18] H. Ding, K. Sricharan, and R. Chellappa, "Exprgan: Facial expression editing with controllable expression intensity," in *AAAI Conference on Artificial Intelligence*, 2018.
- [19] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *Eur. Conf. on Computer Vision*, 2018.
- [20] O. Wiles, A. Sophia Koepke, and A. Zisserman, "X2face: A network for controlling face generation using images, audio, and pose codes," in *Eur. Conf. on Computer Vision*, 2018.
- [21] V. Blanz, T. Vetter *et al.*, "A morphable model for the synthesis of 3d faces," in *Siggraph*, 1999.
- [22] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2016.
- [23] H. Kim, P. Carrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *ACM Transactions on Graphics (TOG)*, 2018.
- [24] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Int. Conf. on Neural Information Processing Systems*, 2016.
- [25] M. Saito, E. Matsumoto, and S. Saito, "Temporal generative adversarial nets with singular value clipping," in *IEEE Int. Conf. on Computer Vision*, 2017.
- [26] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh, "Action-Conditional Video Prediction using Deep Networks in Atari Games," in *Int. Conf. on Neural Information Processing Systems*, 2015.
- [27] I. Ö. Ertugrul and H. Dibeklioglu, "What will your future child look like? modeling and synthesis of hereditary patterns of facial dynamics," in *IEEE Int. Conf. on Face and Gesture Recognition*, 2017.
- [28] W. Wang, Z. Cui, Y. Yan, J. Feng, S. Yan, X. Shu, and N. Sebe, "Recurrent face aging," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2016.
- [29] W. Wang, Y. Yan, Z. Cui, J. Feng, S. Yan, and N. Sebe, "Recurrent face aging with hierarchical autoregressive memory," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 654–668, 2018.
- [30] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *Int. Conf. on Learning Representations*, 2016.
- [31] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, 2001.
- [32] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *IEEE Int. Conf. on Face and Gesture Recognition*, 2015.
- [33] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.
- [34] X. Liu, "Discriminative face alignment," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1941–1954, 2009.
- [35] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a practical face recognition system: Robust alignment and illumination by sparse representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 372–386, 2012.
- [36] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013.
- [37] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *IEEE Int. Conf. on Winter App. of Computer Vision*, 2016.
- [38] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, arXiv preprint:1412.6980.
- [39] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2018.
- [40] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2015.
- [41] R. Arandjelović and A. Zisserman, "All about VLAD," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2013.



Wei Wang received his master degree from Southern Denmark University, and PhD degree from Multimedia and Human Understanding Group (MHUG) at the Department of Electrical and Computer Engineering, University of Trento, Italy. He is currently a Postdoc in CVLab at EPFL, Switzerland. His research interests include machine learning and its application to computer vision, augmented reality and multimedia analysis.



Xavier Alameda-Pineda is a research scientist at Inria Grenoble Rhône-Alpes, with the Perception team. He received M.Sc. degrees in mathematics (2008), in telecommunications (2009) and in computer science (2010) and a Ph.D. in mathematics and computer science (2013) from Université Joseph Fourier. He served as Area Chair at ICCV'17, at ICIAP'19 and at ACM MM'19. He will be Program Chair at ACM MM'22. He is the recipient of several paper awards and of the ACM SIGMM Rising Star Award in 2018. He is the coordinator of the H2020 ICT project SPRING. His scientific interests lie in computer vision, machine learning and signal processing for robotics and multimodal social behavior analysis.



Dan Xu received his PhD degree from Multimedia and Human Understanding Group (MHUG) at the Department of Electrical and Computer Engineering, University of Trento, Italy. He was a research assistant in the Department of Electronic Engineering at The Chinese University of Hong Kong. He is currently a Postdoc in VGG group at the University of Oxford. His research focuses on computer vision, multimedia and machine learning.



Elisa Ricci is an associate professor with the Department of Information Engineering and Computer Science, University of Trento and a researcher at Fondazione Bruno Kessler. She received her PhD from the University of Perugia in 2008. She has since been a post-doctoral researcher at Idiap, Martigny and the Fondazione Bruno Kessler, Trento. She was also a visiting student at University of Bristol during her PhD. Her research interests are mainly in the areas of computer vision and machine learning.



Nicu Sebe is with the Department of Information Engineering and Computer Science, University of Trento, Italy, where he is leading the research in the areas of multimedia information retrieval and human behavior understanding. He was GC of FG 2008 and ACM Multimedia 2013, and PC of ACM Multimedia 2007 and 2011, ECCV 2016 and ICCV 2017. He is a Senior Member of ACM and IEEE and a fellow of IAPR.