Intelligent Systems with Applications 16 (2022) 200139

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Facial expression video generation based-on spatio-temporal convolutional GAN: FEV-GAN

Hamza Bouzid *, Lahoucine Ballihi

LRIT-CNRST URAC 29, Mohammed V University in Rabat, Faculty Of Sciences, Rabat, Morocco

ARTICLE INFO

Keywords:
Facial expression generation
Video generation
Deep learning
Generative adversarial networks
Spatio-temporal convolutional networks

MSC:
41A05
41A10
65D05
65D17

ABSTRACT

Facial expression generation has always been an intriguing task for scientists and researchers all over the globe. In this context, we present our novel approach for generating videos of the six basic facial expressions. Starting from a single neutral facial image and a label indicating the desired facial expression, we aim to synthesize a video of the given identity performing the specified facial expression. Our approach, referred to as FEV-GAN (Facial Expression Video GAN), is based on Spatio-temporal Convolutional GANs, that are known to model both content and motion in the same network. Previous methods based on such a network have shown a good ability to generate coherent videos with smooth temporal evolution. However, they still suffer from low image quality and low identity preservation capability. In this work, we address this problem by using a generator composed of two image encoders. The first one is pre-trained for facial identity feature extraction and the second for spatial feature extraction. We have qualitatively and quantitatively evaluated our model on two international facial expression benchmark databases: MUG and Oulu-CASIA NIR&VIS. The experimental results analysis demonstrates the effectiveness of our approach in generating videos of the six basic facial expressions while preserving the input identity. The analysis also proves that the use of both identity and spatial features enhances the decoder ability to better preserve the identity and generate high-quality videos. The code and the pre-trained model will soon be made publicly available.

1. Introduction

Facial expressions have always been considered one of the essential tools for human interaction. Integrating the ability to recognize and synthesize facial expressions to machines provides a natural and smooth interaction. Which opens the door to many exciting new applications in different fields, including the movie industry, e-commerce, and even in the medical field. Motivated by this, researches have studied facial expression recognition and have already reached a high level of precision, while facial expression generation has been more demanding and less studied in the state of the art. Recently, with the success of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) in data generation, in particular image generation, the task of generating facial expressions has seen tremendous progress.

However, the dynamic facial expressions synthesis is even less studied due to the difficulty of the tasks: **(1)** learning the dataset distribution (facial structure, background), **(2)** representing a natural and smooth evolution of facial expressions (Temporal representation), and **(3)** preserving the same input identity. To address the high complexity of these three tasks, most existing methods tend to treat facial expression generation as a two-steps process. One step for the low dimensional temporal generation (motion) and the other for the spatial generation (content). Such methods (Otberdout et al., 2019; Tulyakov et al., 2018; Wang et al., 2018a) are mostly based on **(1)** the generation of motion as codes in a latent space, thereafter **(2)** combining it with the input image embedding to generate frames individually, through the use of an Image-to-Image translation network. These methods are efficient in learning facial structure and identity preservation, but they are flawed when it comes to modeling spatio-temporal consistency and appearance retainment. This is caused by the independence between frames generation.

Motivated by the success of Deep spatio-temporal neural network models in recognition and prediction tasks (Al Chanti and Caplier, 2018; Ali et al., 2021, 2022; Tran et al., 2018), researchers have proposed a variety of one-step methods (Jang et al., 2018; Vondrick et al., 2016; Wang et al., 2020a, 2020b) that use fractionally strided 3D convolutions. Videos generated by these kinds of methods show more spatio-temporal consistency, however, lower video quality, more noise

* Corresponding author.
E-mail addresses: hamza-bouzid@um5r.ac.ma (H. Bouzid), lahoucine.ballihi@fsr.um5.ac.ma (L. Ballihi).

https://doi.org/10.1016/j.iswa.2022.200139
Received 20 July 2022; Received in revised form 21 September 2022; Accepted 6 October 2022
Available online 8 October 2022

and distortions, and more identity preservation issues compared to two-steps methods. We claim this is due to the high complexity of the three tasks combined in a single network (learning **1.** the spatial presentation, **2.** the temporal representation, **3.** identity preservation). This requires a large network with high potential complexity and a large amount of data, which significantly increases the difficulty of model optimization.

To solve the issues of the low quality, noise and identity preservation capability faced by one-step methods, We propose encoding the input image into two codes in the latent space, using two feature extractors (E_{Id} identity feature extractor, E_s spatial feature extractor). We also suggest exploiting the high performance of state-of-the-art facial recognition systems, by utilizing a pre-trained facial recognition feature extractor as our identity encoder E_{Id}. This grants identity related features that help with identity preservation. In addition, the use of a pre-trained feature extractor allows for applying the optimization process only on the other encoder E_S, that is used to extract other spatial features in order to maintain sufficiently good quality while reconstructing facial expression video.

In summary, our contributions include the following aspects:

1. We propose a conditional GAN, with a single generator and a single discriminator, that generates at each time step a dynamic facial expression video, corresponding to the desired class of expressions. The generated videos present a realistic appearance, and preserve the identity of the input image.
2. We investigate the influence of utilizing two encoders E_{Id} and E_S, where E_{Id} is a facial identity feature extractor and E_S is a spatial feature extractor.
3. We exploit the high potential of state-of-the-art facial recognition systems. We use a pre-trained face recognition model as our generator encoder E_{Id}, which will ensure strongly related identity features. This aims to facilitate the task of the decoder by providing meaningful and structured features.
4. We deeply evaluate our model, quantitatively and qualitatively, on two public facial expressions benchmarks: MUG facial expression database and Oulu-CASIA NIR&VIS facial expression database. We compare it with the recent state-of-the-art approaches: VGAN (Vondrick et al., 2016), MoCoGAN (Tulyakov et al., 2018), ImaGINator (Wang et al., 2020b) and (Otberdout et al., 2019).

2. Related work

Static facial expression generation - Facial expressions synthesis was initially achieved through the use of traditional methods, such as geometric interpolation (Pighin et al., 2006),Parameterization (Raouzaiou et al., 2002), Morphing (Beier and Neely, 1992), etc. These methods show success on avatars, but they fall short when dealing with real faces, and they are unable to generalize a flow of movement for all human faces due to the high complexity of natural human expressions and the variety of identity-specific characteristics. To face these limitations, neural networks based methods have been applied on facial expressions generation, including RBMs (Zeiler et al., 2011), DBNs (Sabzevari et al., 2010) and Variational Auto-Encoders (Kingma and Welling, 2014),etc. These methods learn acceptable data representations and better flow between different data distributions compared to prior methods, but they face problems such as the lack of precision in controlling facial expressions, low resolution, and blurry generated images.

With the appearance of GANs, multiple of its extensions have been dedicated to facial expressions generation. Makhzani et al. (2015) and Zhou and Shi (2017) exploit the concept of adversity with auto-encoders to present Adversarial Auto Encoders. Zhu et al. (2017) propose a conditional GAN, namely CycleGAN, that uses Cycle-Consistency Loss to preserve the key attributes (identity) of the data. Choi et al. (2018) address the inefficiency of creating a new generator for each type of transformation, by proposing an architecture that can handle different transformation between different datasets. Wang et al. (2018b) suggest exploiting the U-Net architecture as GAN generator, in order to increase the quality and the resolution of generated images. US-GAN (Akram and Khan, 2021) uses a skip connection, called the ultimate skip connection, that links the input image with the output of the model, which allows the model to focus on learning expression-related details only. the model outputs the addition of the input image and the generated expression details, improving therefore the identity preservation, but displaying artifacts in areas related to the expression (mouth, nose, eyes). The studies above established the task of generating classes of facial expressions (sad, happy, angry, etc.), but in reality, the intensity of the expression inhibits the understanding of the emotional state of the person. ExprGAN (Ding et al., 2017) used an expression controller module to control the expression intensity continuously from weak to strong. Methods like GANimation (Pumarola et al., 2018), EF-GAN (Wu et al., 2020) used Action Units (AUs) in order to learn conditioning the generation process which offers more diversity in the generated expressions. Other methods such as G2-GAN (Song et al., 2018) and GC-GAN (Qiao et al., 2018) exploited Facial Geometry as a condition to control the facial expression synthesis. The objective of the latter models is to take as input a facial image and facial Landmarks in form of binary images or landmarks coordinates, then learn to generate a realistic face image with the same identity and the target expression. Kollias et al. (2020) and Bai et al. (2022) utilize labels from the 2D Valence-Arousal Space, in which the valence is how positive or negative is an emotion and the arousal is the power of the emotion activation (Russell, 1980), to guide the process of facial expression generation, enhancing the variety and the control of the generated expressions. All these approaches and others have established the task of facial expression generation, but have not considered the dynamicity of these expressions.

Dynamic facial expression generation - Facial expressions are naturally dynamic actions that contain more information and details than a single pose, e.g. the speed of facial expression transformation, head movements when displaying the expression, etc. This information can be significant in understanding the emotional state of a person. To achieve this, methods like (Ha et al., 2020; Li et al., 2021; Tang et al., 2021; Tu et al., 2021; Vowels et al., 2021) focus on facial expression transfer, in which the facial expression is transferred from a driver to a target face, while aiming to preserve the target identity even in situations where the facial characteristics of the driver differs widely from those of the target. In other methods, the motion is generated separately as codes in the latent space, these codes are then fed to the generator in order to generate frames of the video individually. For example, MoCoGAN (Tulyakov et al., 2018) decomposes the video into content and motion information, where the video motion is learned by Gated RNN (GRU) and the video frames are generated sequentially by a GAN. RV-GAN (Gupta et al., 2022) uses a transpose (upsampling instead of downsampling) convolutional LSTMs as GAN generator to generate frames individually. However, the results of both models present content and motion artifacts, and they both could only be applied to seen-before identities, and a finite number of expressions. In (Fan et al., 2019), the principle of MoCoGAN is extended by adding an encoder that helps preserving the input identity, and a coefficient to control the degree of the expression continuously. The authors of Wang et al. (2018a) utilize a Multi-Mode Recurrent Landmark Generator to learn generating variant sequences of facial landmarks of the same category (e.g. different ways to smile), translated later to video frames. In Otberdout et al. (2019), the authors exploit a conditional version of manifold-valued Wasserstein GAN to model the facial landmarks motion as curves encoded as points on a hypersphere. The W-GAN learns the distribution of facial expression dynamics of different classes, from which new facial expression motions are synthesized and transformed to videos by TextureGAN. Other works have investigated guiding facial expression generation by speech audio data, such as Chen et al. (2020), Guo et al. (2021), Liang et al. (2022) and Wang et al. (2022), or by a combination of audio and

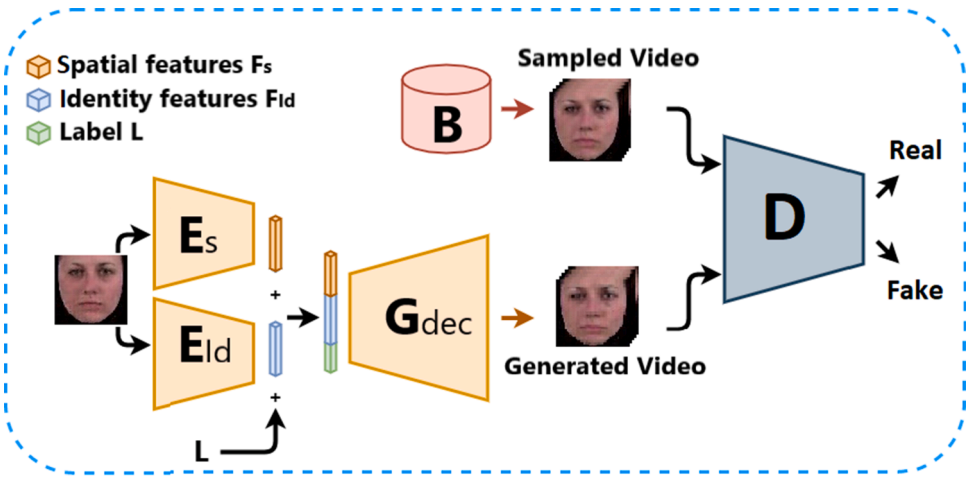

Fig. 1. Overview of the proposed model for generating facial expressions: **FEV-GAN**. The input image is mapped through two encoders, identity (E_{Id}) and spatial (E_S) encoders, into the latent codes F_{Id} and F_S. Both codes and the label L are fed to the decoder G_{dec} that transforms them to a video of the input identity performing the desired expression. While the discriminator D has the purpose of classifying realistic scenes from synthesized scenes.

facial landmark information, like Sinha et al. (2022), Wang et al. (2021) and Wu et al. (2021). All the methods mentioned before are methods that generate a single frame at a time-step, which lowers the dependency between the video frames causing the lack of spatio-temporal consistency. Contrasted to the methods mentioned before, methods like VGAN (Vondrick et al., 2016), G3AN (Wang et al., 2020a) and ImaGINator (Wang et al., 2020b) use a single step for the generation of the whole facial expression video, by employing fractionally strided spatio-temporal convolutions to simultaneously generateappearance, as well as motion. VGAN decomposes the generated videos into two parts, a static section (the background) and a dynamic section, which imposes the use of a generator composed of two streams, for generating the background and the foreground, that are combined in the output to generate the whole video. G3AN aims to model appearance and motion in disentangled manner. This is accomplished by decomposing appearance and motion in a three-stream Generator, where the main stream models spatio-temporal consistency, while the two auxiliary streams enhance the main stream with multi-scale appearance and motion features, respectively. Both VGAN and G3AN are unconditional models that start from a Gaussian noise input, causing the lack of identity preservation and of control over the generated expression. In order to avoid these problems, ImaGINator uses a blend of auto-encoder architecture and spatio-temporal fusion mechanism, where the low-level spatial features in the encoder are sent directly to the decoder (the same concept as U-Net (Ronneberger et al., 2015)). It also uses two discriminators, one processes the whole video and the other processes frame by frame. Videos generated by these kinds of methods show more spatio-temporal consistency but lower video quality, more noise and less identity preservation compared to two-steps methods.

Motivated by the discussed above, we present a novel one-step approach for facial expression generation based on fractionally strided spatio-temporal convolutions. The rest of the paper is organized as follows. In section 3 we introduce our new FEV-GAN model. Section 4 shows the experimental settings and the quantitative and qualitative analysis of the model. Section 5 concludes the paper and provides perspectives for future research.

3. Proposed approach

As stated in the introduction, our main aim is to establish a model that generates dynamic facial expression videos from appearance information and expression category. Thus, we formulate our goal as learning a function $G : \{I, L\} \longrightarrow \widehat{Y}$, where I is the input image, L is the label vector and $\widehat{Y}$ is is the generated video.

To achieve this objective, we propose a Framework consisting of the following components: a Generator network G built on an encoder-decoder architecture. The encoders E_{Id} and E_S take as input a single image I and extract identity features F_{Id} and spatial features F_S, respectively. The decoder G_{dec} utilizes the extracted features (F_{Id},F_S) and a label L to generate a realistic video $\widehat{Y}$. Finally a discriminator D assists the learning of the Generator for both appearance and expression category. The overview of our approach is shown in Fig. 1.

3.1. FEV-GAN model description

In the following, the architecture of our network is described, and details on the generator G and the discriminator D are provided.

Generator E_{Id}, E_S, G_{dec}: As shown in Fig. 1, our generator consists of three networks, a pre-trained image identity encoder E_{Id}, a randomly initialized encoder E_S and a video decoder G_{dec}. The encoder E_{Id} is a well known state-of-the-art facial recognition model, VGG-FACE (Parkhi et al., 2015) feature extractor. It takes a ($64 \times 64 \times 3$) RGB image I as input and transforms it into 1024 feature maps of (14×14), containing the facial identity features F_{Id}. The encoder E_S is initialized with a Gaussian noise. It extracts 512 feature maps of (14×14) features F_S that contain more spatial details beside the identity features. Then, the features F_{Id}, F_S and the label vector L are concatenated and utilized by the decoder G_{dec} to generate the new video. The decoder combines spatio-temporal convolutions and fractionally strided convolutions to transform the input tensor into the high dimensional generated video. Tri-dimensional convolution offers spatial and temporal invariance. Fractional convolution is an efficient over-sampling tool, allowing to transform the latent tensor into a ($32 \times 64 \times 64 \times 3$) video.

Discriminator D: The objectives of the discriminator is to learn the ability (i) to classify realistic scenes from synthesized scenes, (ii) to recognize realistic motion, and (iii) to detect the difference between the different classes of motion. To achieve these three tasks we use a five-layer spatio-temporal convolutional network for the purpose of learning both visual content and motion modeling. The network takes ($32 \times 64 \times 64 \times 3$) videos Y from the dataset or $\widehat{Y}$ generated by the generator G and the labels vectors as input. It then checks if the video is realistic and both the motion and the label are convenient. The discriminator architecture is designed to almost invert G_{dec} by replacing fractionally strided spatio-temporal convolutions with direct spatio-temporal convolutions (to sub-sample instead of over-sample), and adjust the latter layer to produce a binary real or fake classification.

Input: Ground-truth Video $Y = [Y_0, Y_1, ..., Y_n]$; Input Image $I = [I_0, I_1, ..., I_n]$; Target expression label $L = [L_0, L_1, ..., L_n]$;
Output: FEV-GAN model;
1 $E_{Id} \longleftarrow$ initialized with non-learnable pre-trained VGG-FACE parameters;
2 $E_s, G_{dec}, D \longleftarrow$ initialized with learnable parameters sampled from a Gaussian distribution;
3 **for** *the number of epochs* **do**
4 **for** *the number of iterations in an epoch* **do**
5 $F_{Id}(I_i) \longleftarrow E_{Id}(I_i)$;
6 $F_s(I_i) \longleftarrow E_s(I_i)$;
7 $\hat{Y}_i \longleftarrow G_{dec}(F_{Id}(I_i), F_s(I_i), L_i)$;
8 $E_{real} \longleftarrow D(Y_i, L_i)$;
9 $E_{recon} \longleftarrow D(\hat{Y}_i, L_i)$;
10 $\mathcal{L}_D \longleftarrow \log(E_{real}) + (1 - \log(1 - E_{recon}))$;
11 $D \longleftarrow D - \alpha(\partial \mathcal{L}_D / \partial D)$;
12 $\mathcal{L}_{rec} \longleftarrow \| Y_i - \hat{Y}_i \|_1$;
13 $\mathcal{L}_{id} \longleftarrow \| F_{vgg}(Y_i) - F_{vgg}(\hat{Y}_i) \|_1$;
14 $\mathcal{L}_G \longleftarrow \log(E_{real}) + \mathcal{L}_{rec} + \mathcal{L}_{id}$;
15 $G \longleftarrow G - \alpha(\partial \mathcal{L}_G / \partial G)$;
 end
end

Algorithm 1. Our Learning Algorithm for the proposed model FEV-GAN

3.2. Loss function

To train our model $G : \{I, L\} \longrightarrow \widehat{Y}$, both network D and G are optimized using our objective function Eq. 1.

$$\mathscr{L}_{total}(G, D) = \mathscr{L}_{adv}(G, D) + \lambda_1 \mathscr{L}_{rec}(G) + \lambda_2 \mathscr{L}_{id}(G), \tag{1}$$

which consists of three losses: Adversarial loss $\mathscr{L}_{adv}$ that helps the generator learn the database distribution. Reconstruction loss $\mathscr{L}_{rec}$ that captures the overall structure of the video and improves the quality. Identity loss $\mathscr{L}_{id}$ which ensures facial identity details preservation. Due to the large difference we found between the three losses values, λ_1 and λ_2 parameters are used to help stabilize the training and balance the optimization of our model.

We therefore aim to solve

$$G^* = arg \min_G \max_D \mathscr{L}_{total}. \tag{2}$$

We note that the parameters of E_{Id} are frozen in the training phase, as it is a pre-trained model that already extracts facial features.

Adversarial loss: Our conditional adversarial loss is a cross entropy loss, that is applied on both G and D, with the aim that $G(I, L)$ learns to generate videos $\widehat{Y}$ that look similar to real videos Y, and D learns to distinguish between real samples Y and generated samples $\widehat{Y}$. We train the model so as G aims to minimize the function while D aims to maximize it.

$$\begin{aligned} \min_G \max_D \mathscr{L}_{adv} = \min_G \max_D \; & \mathbb{E}_{Y \sim P_{data}(Y)} [logD(Y; L)] \\ & + \mathbb{E}_{z \sim P_z(z)} [1 - logD(G(I; L); L)]. \end{aligned} \tag{3}$$

Reconstruction loss: Our reconstruction loss at the video level is defined by

$$\mathscr{L}_{rec} = [\, ||Y - \widehat{Y}||_1 \,], \tag{4}$$

with the purpose of capturing the overall structure, video consistency and helping preserve the identity details. This loss is a L_1 norm loss between the generated videos $\widehat{Y}$ and the ground truth videos Y. By combining this loss with $\mathscr{L}_{adv}$, it helps the generator G create more realistic videos and reconstruct smooth expression motion.

Identity loss: The Identity Loss is used for identity preservation. It is a L_1 norm loss similar to the $\mathscr{L}_{rec}$, but while $\mathscr{L}_{rec}$ aims to minimize the $L1$ distance between pixel values of the generated and the ground truth videos, $\mathscr{L}_{id}$ aims to minimize the $L1$ distance between the identity features of the input image and the frames of generated video. We exploit the same architecture of our VGG-FACE encoder to extract the identity features from both input and output data. This loss is formalized as:

$$\mathscr{L}_{id} = \sum_{i=0}^{N} [\, ||F_{vgg}(I) - F^i_{vgg}(\widehat{Y})||_1 \,], \tag{5}$$

where $F_{vgg}(I)$ are the identity features of the input image I, and $F^i_{vgg}(\widehat{Y})$ are the identity features of the i^{th} frame of $\widehat{Y}$.

We furthermore use $\mathscr{L}_{id}$ on 4 frames of the video instead of 32 frames. We count on spatio-temporal consistency of the 3D convolutions to generalize the identity preservation over the rest of the video.

3.3. Training algorithm

Algorithm 1 highlights the training process of the proposed FEV-GAN model. The training dataset (I, L, Y) is fed into Algorithm 1. First, the model parameters are initialized. The parameters of E_{Id} are loaded from the pre-trained VGG-FACE model, while the rest of the parameters are initialized from a Gaussian distribution **(1;2)**. The outer for loops are used to learn from the data for a given number of iterations **(3;4)**. The input image I_i is fed into the encoders E_{Id} **(5)** and E_s **(6)**, that encode it to identity F_{Id} and spatial F_s features, respectively. These features and the

Table 1
Quantitative comparison results of FEV-GAN and baseline models using PSNR, SSIM, ACD, and ACD-I metrics.

	Trained on MUG				Trained on Oulu-Casia			
Model	PSNR	SSIM	ACD	ACD-I	PSNR	SSIM	ACD	ACD-I
VGAN (Vondrick et al., 2016)	16.32	0.41	0.14	1.55	15.09	0.61	0.27	1.37
C-VGAN (Vondrick et al., 2016)	22.30	0.83	0.09	0.73	15.98	0.61	0.25	1.26
MoCoGAN (Tulyakov et al., 2018)	18.16	0.58	0.15	0.90	-,	-,	-,	-,
ImaGINator (Wang et al., 2020b)	20.29	0.85	0.08	0.29	22.98	0.84	0.07	0.16
(Otberdout et al., 2019)	25.9	0.90	-,	-,	24.44	0.89	-,	-,
The proposed FEV-GAN	**27.10**	**0.91**	**0.09**	**0.23**	**25.61**	**0.89**	**0.12**	**0.19**

given label L_i are then used by the decoder G_{dec} to generate a fake video $\widehat{Y}_i$ **(7)**. Next, the discriminator network D estimates the probability that a video is sampled from the dataset rather than the generator **(8;9)**. The generated videos and the discriminator estimations are used to calculate the losses **(10;12;13;14)**. Finally, the back propagation method and Adam optimizer are used to train the FEV-GAN model **(11;15)**.

4. Experiment

To evaluate our model we performed an extensive experimental validation. In this following section, the experimental setup of our learning is detailed. Then, the model is evaluated quantitatively and qualitatively and compared to VGAN, MoCoGAN, ImaGINator and (Otberdout et al., 2019). Finally, an ablation study is presented to observe the influence of each component of our model.

4.1. Dataset

The evaluation of our method is performed on:

the MUG Facial Expression database (Multimedia Understanding Group Facial Expression database) (Aifanti et al., 2010): contains videos of 86 people, performing seven facial expressions: "happiness", "sadness", "surprise", "anger", "disgust", "fear" and "neutral". Videos in the database start and end with neutral expressions and display the peak of the expression in the middle. Each video consists of 50 to 160 RGB frames of 896 × 896 resolution. The data of 52 subjects is available to authorized Internet users, 25 subjects are available on request and the remaining 9 subjects are available only in the MUG laboratory. In this work we used the public data of 52 subjects. We only used the first half of the six basic expressions videos, which starts with the neutral expression and ends with the expression peak.

Oulu-CASIA NIR&VIS facial expression database (Zhao et al., 2011): consists of 480 videos of 80 people, between 23 to 58 years old, performing six facial expressions: "happiness", "sadness", "surprise", "anger", "disgust" and "fear". Each video consists of 9 to 72 RGB frames of 320 × 240 resolution, that begins with a neutral expression and end with the apex of the corresponding expression. The whole database is available to authorized Internet users.

4.2. Implementation details

Before using the data, we first split the data in a subject independent manner into two sets, 80% of the data for the learning and 20% for the testing phases. We then cropped the face area and removed the background using OpenFace (Baltrusaitis et al., 2018), normalized all videos to 32 frames using the framework proposed in Karcher (1977), and scaled each frame to 64 × 64 pixels. After the pre-processing phase, we ended up with 32 × 64 × 64 × 3 videos of different expressions and subjects with black background.

The weights of the networks E_S, G_{dec} and D are initially sampled from a Gaussian distribution of the mean zero and the standard deviation 0.01. The weights of the network E_{Id} are initialized with the weights of the pre-trained VGG-FACE (Parkhi et al., 2015) and frozen in the training phase. Gradient descent is used for 400 epochs, in order to solve Eq. (2). Binary cross entropy loss is used for the adversarial loss, and L_1 norm is used for both reconstruction and identity losses. Additionally, Adam Optimizer (Kingma and Ba, 2014) is used to train the model with a learning rate of 0.0002 and a Momentum of 0.5. The pixels values of the input image and videos are scaled to the interval $[-1.1]$. Each layer of G_{dec} is followed by the activation function ReLU (Agarap, 2018) and batch normalization (Ioffe and Szegedy, 2015), except for the output, which use tanh. LeakyReLU (Xu et al., 2015) and batch normalization are used in the discriminator except for the input layer. The implementation of the network is done with the TensorFlow Framework based on the implementation of Vondrick et al. (2016). The training on MUG data takes approximately 70 h on an Nvidia GeForce GTX 1650 GPU (4Gb of memory). The training on Oulu-CASIA NIR&VIS takes approximately 40 h on an Nvidia Titan V GPU (12GB of memory).

4.3. Evaluation metrics and baselines

To deeply evaluate our model quantitatively, we use several metrics:

1) PSNR: (Peak Signal-to-Noise Ratio) measures the pixel-level similarity between the generated videos and their ground truth.

2) SSIM: (Structural Similarity Index Measure) represents the structural similarity between real and reconstructed videos.

3) ACD (Tulyakov et al., 2018): (Average Content Distance) measures content consistency in generated videos, based on the average of all pairwise L_2 distances between facial features of every two consecutive frames in a generated video. However, the ACD only represent identity consistency in the video, lacking information on the identity preservation of the input image.

4) ACD-I (Zhao et al., 2018): an ACD extension that measures the identity preservation of the input face in the generated video. It calculates the average of L_2 distances between the facial features of the frames of the video and the input image. To extract the facial feature vectors, we use OpenFace (Amos et al., 2016), which is a deep model trained for facial recognition that can outperform human performance.

Higher SSIM and PSNR scores indicate better generated videos quality, lower ACD scores indicate similar faces in consecutive generated video frames, and lower ACD-I values indicate higher similarity between faces in the input images and the generated videos.

Regarding the state-of-the-art models used for comparison, we used the public codes of VGAN and ImaGINator provided by the authors with some minor changes. Since we deal with facial expression, we trained two versions of VGAN, the original version and a conditional version. In the conditional VGAN we used an encoder that transforms the input image to a latent code, the latent code is then concatenated with the labels and fed to the generator, that generates videos of the target expressions of the same input face. For MoCoGAN, we utilized the results given in Wang et al. (2020b). As for Otberdout et al. (2019), we used the results in the original paper.

4.4. Experimental results & analysis

Quantitative results - to perform our quantitative analysis, we first generate 106 videos of 18 different subjects from the testing subset performing the six basic facial expressions.

First, we demonstrate that our model FEV-GAN offers better

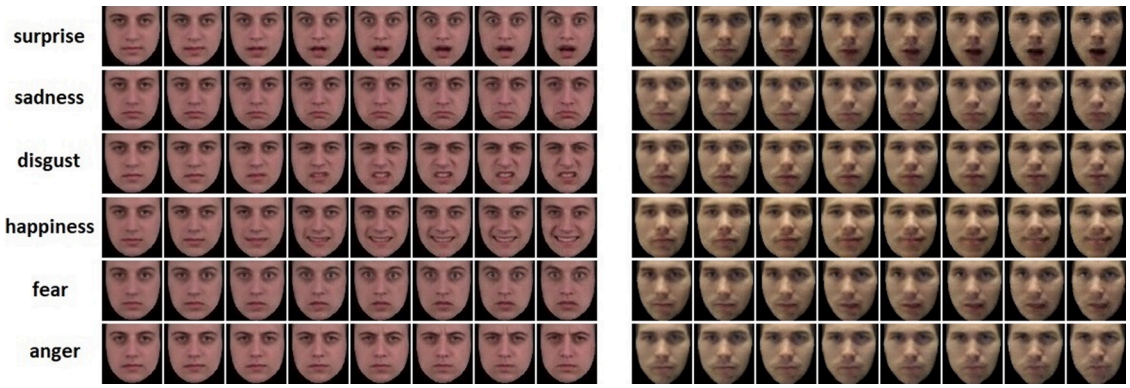

Fig. 2. The generation of facial expression videos on the MUG database (left) and Oulu-Casia (right). The image sequences show the six basic facial expressions of the same subject on the test dataset. The presented images are sampled every two frames. More examples of different identities are given in supplementary material.

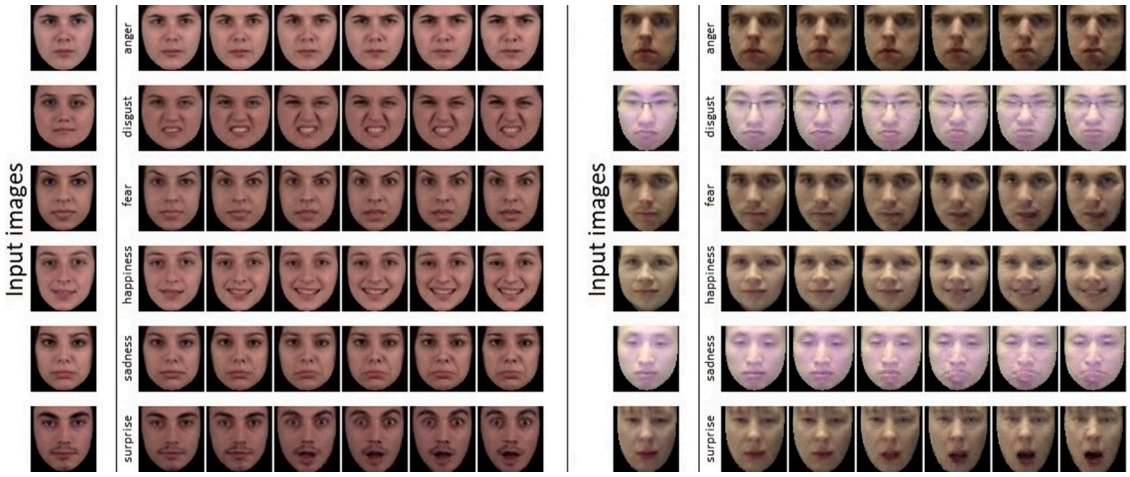

Fig. 3. Examples of videos generated by our model of the six basic facial expression performed by a person given in the input image.

Fig. 4. Qualitative comparison of sequences generated by FEV-GAN model and by the state-of-the-art models on the MUG database. The sequences generated by our model (a, c, e, g, i), by VGAN (b), C-VGAN (d) and by ImaGINator (h) are randomly selected from the test results. The sequence generated by MoCoGAN (f) and by Otberdout et al. (2019) (j) are taken from the original papers. All the images are sampled with the time step of 4.

reconstruction capabilities than all baselines using **PSNR** and **SSIM**. Table 1 indicates that our model generates better quality videos with less noise, and preserves the general structure of the input through all the video. Then, we analyse the content consistency using **ACD** metric. Our model achieves similar content consistency to models with spatio-temporal convolutions (VGAN and ImaGINator) that are known to have high content consistency, while it surpasses MoCoGAN that uses Image-to-Image translation. As for identity preservation, our proposed model largely surpasses VGAN, C-VGAN and MoCoGAN, and is competitive with ImaGINator.

We note that VGAN does not preserve the identity. Its score is used for the purpose of representing the metric values where identity is not preserved. The ACD and ACD-I comparison does not include the model used in Otberdout et al. (2019) for the reason that the authors have used

Table 2
Subjective comparison of FEV-GAN and baseline models. The reported results are the mean of the raters preferences.

Models	Rater preference(%)
FEV-GAN/C-VGAN	91.44% / 08.56%
FEV-GAN/MoCoGAN	78.09% / 21.91%
FEV-GAN/ImaGINator	71.07% / 28.93%
FEV-GAN/(Otberdout et al., 2019)	68.05% / 31.95%

a different approach to calculate the metrics.

Qualitative results - Fig. 2 presents examples of generated videos of the six facial expressions of the same identity. Fig. 3 demonstrates examples of generated videos of the six facial expressions of different given identities. Both figures show results of our proposed model trained on MUG dataset (left) and Oulu-Casia dataset (right). The generated videos are taken randomly from the test results. We recall that the data used for learning and testing the model is subject independent. The synthesized videos display generally natural, smooth and continuous facial variations that can be controlled according to the input label. They also preserve the features of the input images, such as the identity, the beard and the glasses. Our model shows slightly better results on MUG data than on Oulu-Casia data. This is due to the fact that Oulu-Casia contains fewer data instances with a variety of skin colors, lighting conditions and with accessories (glasses) on the face, which increases the difficulty of pattern learning for the model. More examples of videos generated by our model are given in the Fig. 7 in Appendix.

In Fig. 4, a comparison between our results and the state-of-the-art on the MUG database is conducted. The figure illustrates videos generated by our model, VGAN, C-VGAN, MoCoGAN, ImaGINator and (Otberdout et al., 2019). On the left side of the image, we show the input images utilized to generate the videos, in the middle, we show the videos generated by our model (a, c, e, g, i), and on the right side we show the videos generated by the baselines (b, d, f, h, j). In each comparison, each line, the videos generated by our model and the baseline contain the same identity and expression. In the first comparison (a,b), VGAN shows a natural smooth expression, but a total loss of the identity. In (c,d), C-VGAN generally preserves the input identity but with detail loss (like skin color), noise and artifacts. In (e,f), MoCoGAN offers sufficient identity preservation, but the identity is already used in the training set. MoCoGAN also suffers from the unnatural expression and distortions in the generated images, we suspect this causes the high ACD-I value even when the identity is preserved better than C-VGAN. In (g,h), ImaGINator preserves the input identity, but changes the skins color and displays an unnatural expression. In (i,j), the baseline shows sufficient identity preservation and structure consistency, but displays artifacts near the mouth and nose area. The figure demonstrates that our model generally surpasses the baselines in terms of identity preservation and quality of the generated videos. It also demonstrate that our model maintains the same content-consistency and expression naturalness as spatio-temporal models. More qualitative comparison with these methods and others are conducted in Fig. 8 in Appendix.

In addition, we performed a subjective test, in which we aim to obtain opinions of human raters on videos generated by our method in comparison to the state-of-the-art. We asked 17 volunteers to compare videos generated by our method and by the baselines. All the volunteers are PhD students and researchers in the deep learning field, from different university laboratories. The volunteers were asked more than 30 questions, where, each time we offered them two videos, generated by our model and by one of the baselines, without revealing the source of the videos. The question they were asked is to choose the best one of two given videos based on the average of expression naturalness, identity preservation and video quality. As shown in Table 2, our method largely outperforms the cited models. The raters show strong preference for our method over C-VGAN (91.44% VS 08.56%), MoCoGAN (78.09% VS 21.91%), ImaGINator (71.07% VS 28.93%), which corresponds to the quantitative results. The raters also commented that **1.** in comparison to C-VGAN, we show almost the same expression naturalness and content consistency, while surpassing it in identity preservation and quality of the generated videos. **2.** In comparison to MoCoGAN, we surpass it in all criteria. **3.** As for ImaGINator, we have the same level of identity preservation, but we offer better quality and more natural expressions. **4.** The raters also mentioned that the closest to our model is the one proposed in Otberdout et al. (2019). They have stated that the baseline slightly outperforms ours in identity preservation, while we generate videos with less noise and artifacts.

The results of the quantitative and the qualitative analysis prove that our model maintain the same content consistency and expressions naturalness as other spatio-temporal convolution models, while offering better quality and identity preservation.

Ablation study - In this section, we focus on demonstrating the importance of the techniques used to build our model. We showcase the effect of the double encoder method and the effect of the pre-trained facial feature extractor. This is conducted by performing the ablation study on the MUG database. We use the same evaluation metrics we used previously, on multiple versions of our model, where we cancel the target component and observe its effect. We first train two new other versions of our model. In the first one, we discard the spatial encoder E_S, while in the second one we remove the identity features encoder E_{Id}. Both networks are trained and evaluated using the same dataset, parameters, losses and for the same number of epochs, as the full network. Fig. 5 shows sequences generated by our full network, by the network w/o E_S, and by the network w/o E_{Id}. We deduce that our full network generates videos of the input identity performing the target expression with high quality and better facial details. As revealed in the sequences of Fig. 5.**b**, the network w/o E_S generates videos of lower quality and worse identity preservation, and it does not sufficiently preserve

Table 3
Performance comparison of our model without E_{Id} and E_S

Model	PSNR	SSIM	ACD	ACD-I
FEV-GAN w/o E_S	25.44	0.90	0.11	0.34
FEV-GAN w/o E_{Id}	19.71	0.69	0.13	1.22
Full FEV-GAN	**27.10**	**0.91**	**0.09**	**0.23**

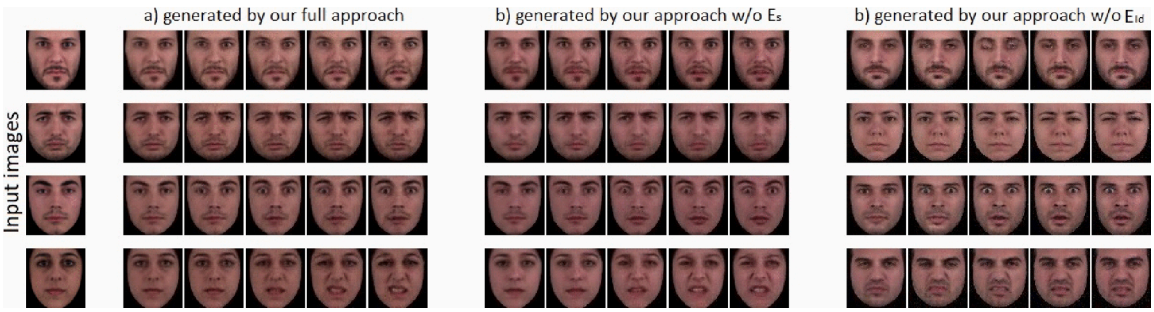

Fig. 5. Qualitative comparison of sequences generated by our full FEV-GAN model (a), by FEV-GAN w/o E_S (b), and by FEV-GAN w/o E_{Id} (c) on the MUG database.

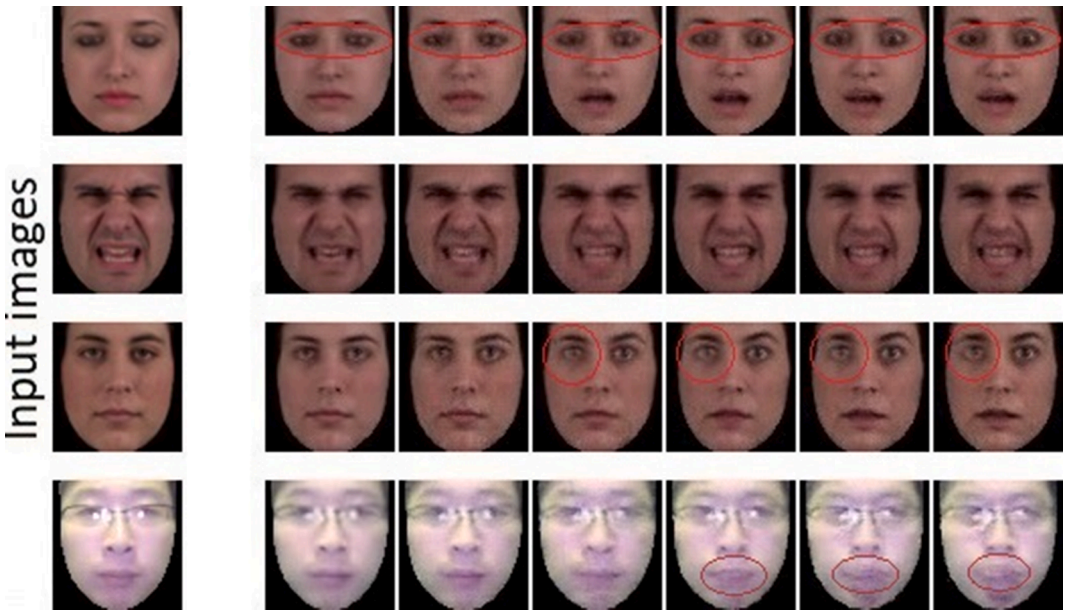

Fig. 6. Examples of flawed video generation by our model.

important details like the skin color, facial hair, and minor details like the surrounding areas of the mouth, nose and eyes. As for Fig. 5.**c**, the videos generated by network w/o E_{Id} show more natural expressions but they also show more noise and distortions and they totally lack the identity preservation,.

The tests shown in Table 3 reveal that the full version of the network gives much better results than the modified versions in terms of quality, content consistency and identity preservation. This can be explained by the existence of the two encoders architecture. E_{Id} capability to guarantee strongly identity related features makes it easier for E_S to learn extracting other features that provide more details and information, giving the decoder the capability to learn generating better quality videos.

4.5. Discussion and limitations

From the quantitative and qualitative comparisons, we conclude that our model largely outperforms the benchmarks C-VGAN and MoCoGAN in all criteria. As for the comparison to ImaGINator, we show similar content consistency, however ImaGINator results display some identity detail loss and unnatural expressions with clear distortions in the mouth and eyes area Fig. 4.**f** Fig. 8.**j**. For (Otberdout et al., 2019), It generates videos with high identity and structure preservation, but it also shows artifacts in the mouth and eyes area. We argue that our proposed model generates generally better quality videos of natural expressions of the input identity, with minimum noise and distortions.

However, there are some inaccuracies that flaw our FEV-GAN model. Fig. 6 illustrates some examples of imprecise video generation. For example, the model learns to synthesis the teeth region when not given in the source image, but fails at generating the eyes when they are not clearly displayed in the input image (1^{st} example of Fig. 6). Also, the model is trained with neutral expression input images. If given a non-neutral expression, the generated video does not display the transition from neutral to the target expression (2^{nd} example of Fig. 6). In addition, the model occasionally constructs videos with some minor distortions in the eyes or mouth regions (3^{rd} and 4^{th} example in Fig. 6).

5. Conclusion and perspectives

In this paper, we present a novel Conditional GAN, namely FEV-GAN, for effectively generating the six basic facial expressions videos, given an input single neutral image and a target facial expression category. Specifically, we address the low quality and identity preservation issues encountered by facial expressions generation models that utilize fractionally strided spatio-temporal convolutions.

Based on our state-of-the-art study, these issues are related to the difficulty of the task of generating dynamic facial expressions. Our FEV-GAN model remedies these issues by utilizing two distinct encoders E_{Id} (identity-encoder) and E_S (spatial-encoder) that extract respectively the identity features F_{Id} and spatial features F_S. These features are given as input to the decoder G_{dec} in order to better preserve the identity and generate high quality facial expression videos.

We have deeply evaluated our method on two benchmark databases, The MUG facial expression database and Oulu-CASIA NIR&VIS facial expression database, quantitatively by using different metrics (PSNR, SSIM, ACD, ACD-I), and qualitatively by using expert human eye rating. The results of these tests confirm our claims and show that our method significantly surpasses the state-of-the-art baselines in dynamic facial expression generation.

To further this research, we plan to test the model with other state-of-the-art facial recognition encoders, such as FaceNet (Schroff et al., 2015), OpenFace (Baltrusaitis et al., 2018), and VGG-FACE2 (Cao et al., 2017). We also project to adapt our model to utilize facial landmarks sequences and/or facial action units instead of one-hot vector as a target category label. This will offer a large variation of possible expressions and more control over the generated expression.Another interesting perspective is to extend this approach to 3D facial expression video and 2D/3D human action generation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

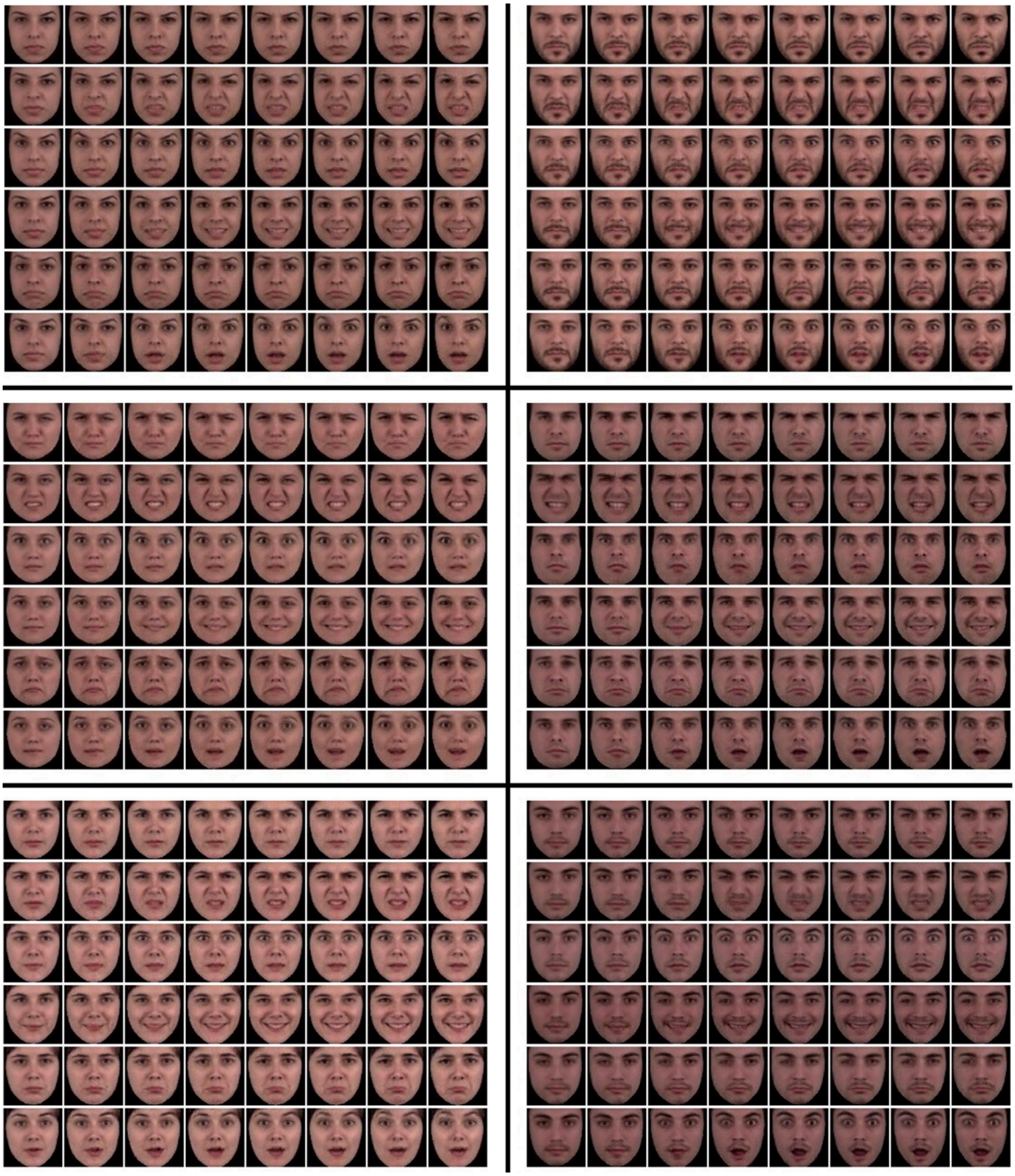

Fig. 7. examples of sequences generated by our proposed model. The image sequences in each box show the six basic facial expressions performed by the same subject on the MUG test dataset.

We notice MoCoGAN (b) shows large artifacts and generates videos where the intensity of the expression does not increase continuously. The video starts with neutral expression and suddenly turns to the target expression. G3AN (d) shows the transition between the expression, but the transition frames are noisy and contain artifacts. RV-GAN (f) video displays better quality and expression smoothness. However, these methods lack the identity preservation, which is very important when dealing with facial expression generation. C-VGAN (h) shows smooth facial expression and preserves the main structure of the input identity but loses much important details. ImaGINator (j) sufficiently preserves the identity but the expressions are unnatural and the generated frames are distorted. (Otberdout et al., 2019) (l) and VDSM(Vowels et al., 2021) (n) show natural smooth facial expressions, and sufficient identity preservation. Nonetheless, VDSM images are blurry and (Otberdout et al., 2019) images display artifacts near the mouth area. Our generated samples present smoother and continuous facial expression evolution, sufficient identity details preservation, and they show less noise and artifacts compared to the stat-of-the-art models.

Fig. 8. Qualitative comparison of sequences generated by FEV-GAN model and by the state-of-the-art models on the MUG database. The sequences generated by our model (a, c, e, g, i, k, l), by C-VGAN (h) and by ImaGINator (j) are randomly selected from the test results. The sequence generated by MoCoGAN (b), (Otberdout et al., 2019) (h), G3AN (d), RV-GAN (f), and by VDSM (n) are taken from the original papers. The absence of input identity indicate that the baseline does not preserve the input identity.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.iswa.2022.200139

References

Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *CoRR, abs/1803.08375*.

Aifanti, N., Papachristou, C., & Delopoulos, A. (2010). The mug facial expression database. *11th international workshop on image analysis for multimedia interactive services wiamis 10* (pp. 1–4). IEEE.

Akram, A., & Khan, N. (2021). Us-gan: On the importance of ultimate skip connection for facial expression synthesis. *arXiv preprint arXiv:2112.13002*.

Al Chanti, D., & Caplier, A. (2018). Deep learning for spatio-temporal modeling of dynamic spontaneous emotions. *IEEE Transactions on Affective Computing, 12*(2), 363–376.

Ali, A., Zhu, Y., & Zakarya, M. (2021). Exploiting dynamic spatio-temporal correlations for citywide traffic flow prediction using attention based neural networks. *Information Sciences, 577*, 852–870.

Ali, A., Zhu, Y., & Zakarya, M. (2022). Exploiting dynamic spatio-temporal graph convolutional neural networks for citywide traffic flows prediction. *Neural networks, 145*, 233–247.

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

Bai, W., Quan, C., & Luo, Z.-W. (2022). Data-driven dimensional expression generation via encapsulated variational auto-encoders. *Cognitive Computation*, 1–13.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. *2018 13th ieee international conference on automatic face & gesture recognition (fg 2018)* (pp. 59–66). IEEE.

Beier, T., & Neely, S. (1992). Feature-based image metamorphosis. *ACM SIGGRAPH computer graphics, 26*(2), 35–42.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2017). Vggface2: A dataset for recognising faces across pose and age. *CoRR, abs/1710.08092*.

Chen, L., Cui, G., Liu, C., Li, Z., Kou, Z., Xu, Y., & Xu, C. (2020). Talking-head generation with rhythmic head motion. *European conference on computer vision* (pp. 35–51). Springer.

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. *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 8789–8797).

Ding, H., Sricharan, K., & Chellappa, R. (2017). Exprgan: Facial expression editing with controllable expression intensity. *CoRR, abs/1709.03842*.

Fan, L., Huang, W., Gan, C., Huang, J., & Gong, B. (2019). Controllable image-to-video translation: A case study on facial expression generation, *vol. 33. Proceedings of the aaai conference on artificial intelligence* (pp. 3510–3517).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems* (pp. 2672–2680).

Guo, Y., Chen, K., Liang, S., Liu, Y.-J., Bao, H., & Zhang, J. (2021). Ad-nerf: Audio driven neural radiance fields for talking head synthesis. *Proceedings of the ieee/cvf international conference on computer vision* (pp. 5784–5794).

Gupta, S., Keshari, A., & Das, S. (2022). Rv-gan: Recurrent gan for unconditional video generation. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 2024–2033).

Ha, S., Kersner, M., Kim, B., Seo, S., & Kim, D. (2020). Marionette: Few-shot face reenactment preserving identity of unseen targets, *vol. 34. Proceedings of the aaai conference on artificial intelligence* (pp. 10893–10900).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR, abs/1502.03167*.

Jang, Y., Kim, G., & Song, Y. (2018). Video prediction with appearance and motion conditions. *International conference on machine learning* (pp. 2225–2234). PMLR.

Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics, 30*(5), 509–541.

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

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

Kollias, D., Cheng, S., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision, 128*(5), 1455–1484.

Li, D., Qi, W., & Sun, S. (2021). Facial landmarks and expression label guided photorealistic facial expression synthesis. *IEEE Access, 9*, 56292–56300.

Liang, B., Pan, Y., Guo, Z., Zhou, H., Hong, Z., Han, X., Han, J., Liu, J., Ding, E., & Wang, J. (2022). Expressive talking head generation with granular audio-visual control. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 3387–3396).

Makhzani, A., Shlens, J., Jaitly, N., & Goodfellow, I. J. (2015). Adversarial autoencoders. *CoRR, abs/1511.05644*.

Otberdout, N., Daoudi, M., Kacem, A., Ballihi, L., & Berretti, S. (2019). Dynamic facial expression generation on hilbert hypersphere with conditional wasserstein generative adversarial nets. *CoRR, abs/1907.10087*.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. *Proceedings of the british machine vision conference (bmvc)* (pp. 41.1–41.12). BMVA Press. https://doi.org/10.5244/C.29.41. https://dx.doi.org/10.5244/C.29.41

Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., & Salesin, D. H. (2006). Synthesizing realistic facial expressions from photographs. *Acm siggraph 2006 courses* (pp. 19–es).

Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., & Moreno-Noguer, F. (2018). Ganimation: Anatomically-aware facial animation from a single image. *Proceedings of the european conference on computer vision (eccv)* (pp. 818–833).

Qiao, F., Yao, N., Jiao, Z., Li, Z., Chen, H., & Wang, H. (2018). Geometry-contrastive generative adversarial network for facial expression synthesis. *CoRR, abs/1802.01822*.

Raouzaiou, A., Tsapatsoulis, N., Karpouzis, K., & Kollias, S. (2002). Parameterized facial expression synthesis based on mpeg-4. *EURASIP Journal on Advances in Signal Processing, 2002*(10), 521048.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology, 39*(6), 1161.

Sabzevari, M., Toosizadeh, S., Quchani, S. R., & Abrishami, V. (2010). A fast and accurate facial expression synthesis system for color face images using face graph and deep belief network, *vol. 2. 2010 international conference on electronics and information engineering* (pp. V2–354). IEEE.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. *CoRR, abs/1503.03832*.

Sinha, S., Biswas, S., Yadav, R., & Bhowmick, B. (2022). Emotion-controllable generalized talking face generation. *arXiv preprint arXiv:2205.01155*.

Song, L., Lu, Z., He, R., Sun, Z., & Tan, T. (2018). Geometry guided adversarial facial expression synthesis. *Proceedings of the 26th acm international conference on multimedia* (pp. 627–635).

Tang, J., Shao, Z., & Ma, L. (2021). Eggan: Learning latent space for fine-grained expression manipulation. *IEEE MultiMedia, 28*(3), 42–51.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 6450–6459).

Tu, X., Zou, Y., Zhao, J., Ai, W., Dong, J., Yao, Y., Wang, Z., Guo, G., Li, Z., Liu, W., et al. (2021). Image-to-video generation via 3d facial dynamics. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(4), 1805–1819.

Tulyakov, S., Liu, M.-Y., Yang, X., & Kautz, J. (2018). Mocogan: Decomposing motion and content for video generation. *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1526–1535).

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

Vowels, M. J., Camgoz, N. C., & Bowden, R. (2021). Vdsm: Unsupervised video disentanglement with state-space modeling and deep mixtures of experts. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8176–8186).

Wang, S., Li, L., Ding, Y., Fan, C., & Yu, X. (2021). Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*.

Wang, S., Li, L., Ding, Y., & Yu, X. (2022). One-shot talking face generation from single-speaker audio-visual correlation learning, *vol. 36*. *Proceedings of the aaai conference on artificial intelligence* (pp. 2531–2539).

Wang, W., Alameda-Pineda, X., Xu, D., Fua, P., Ricci, E., & Sebe, N. (2018a). Every smile is unique: Landmark-guided diverse smile generation. *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 7083–7092).

Wang, X., Li, W., Mu, G., Huang, D., & Wang, Y. (2018b). Facial expression synthesis by u-net conditional generative adversarial networks. *Proceedings of the 2018 acm on international conference on multimedia retrieval* (pp. 283–290).

Wang, Y., Bilinski, P., Bremond, F., & Dantcheva, A. (2020a). G3an: Disentangling appearance and motion for video generation. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 5264–5273).

Wang, Y., Bilinski, P., Bremond, F., & Dantcheva, A. (2020b). Imaginator: Conditional spatio-temporal gan for video generation. *The ieee winter conference on applications of computer vision* (pp. 1160–1169).

Wu, H., Jia, J., Wang, H., Dou, Y., Duan, C., & Deng, Q. (2021). Imitating arbitrary talking style for realistic audio-driven talking face synthesis. *Proceedings of the 29th acm international conference on multimedia* (pp. 1478–1486).

Wu, R., Zhang, G., Lu, S., & Chen, T. (2020). Cascade ef-gan: Progressive facial expression editing with local focuses. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 5021–5030).

Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *CoRR, abs/1505.00853*.

Zeiler, M. D., Taylor, G. W., Sigal, L., Matthews, I., & Fergus, R. (2011). Facial expression transfer with input-output temporal restricted boltzmann machines. *Advances in neural information processing systems* (pp. 1629–1637).

Zhao, G., Huang, X., Taini, M., Li, S. Z., & PietikäInen, M. (2011). Facial expression recognition from near-infrared videos. *Image and Vision Computing, 29*(9), 607–619.

Zhao, L., Peng, X., Tian, Y., Kapadia, M., & Metaxas, D. (2018). Learning to forecast and refine residual motion for image-to-video generation. *Proceedings of the european conference on computer vision (eccv)* (pp. 387–403).

Zhou, Y., & Shi, B. E. (2017). Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. *2017 seventh international conference on affective computing and intelligent interaction (acii)* (pp. 370–376). IEEE.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the ieee international conference on computer vision* (pp. 2223–2232).