IEEE Open Journal of Signal Processing

Received 21 February 2023; revised 21 April 2023; accepted 22 April 2023. Date of publication 16 May 2023; date of current version 6 June 2023. The review of this article was arranged by Associate Editor H. Vicky Zhao.

Digital Object Identifier 10.1109/OJSP.2023.3275052

Self-Supervised Spontaneous Latent-Based Facial Expression Sequence Generation

CHUIN HONG YAP, MOI HOON YAP (Senior Member, IEEE), ADRIAN K. DAVISON, AND RYAN CUNNINGHAM

Department of Computing and Mathematics, Manchester Metropolitan University, M15 6BH Manchester, U.K.

CORRESPONDING AUTHOR: CHUIN HONG YAP (e-mail: chuin.h.yap@stu.mmu.ac.uk).

This work was supported by Manchester Metropolitan University VC PhD Studentship.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by (Name of Review Board or Committee) (IF PROVIDED under Application No. xx, and performed in line with the (Name of Specific Declaration).

ABSTRACT In this article, we investigate the spontaneity issue in facial expression sequence generation. Current leading methods in the field are commonly reliant on manually adjusted conditional variables to direct the model to generate a specific class of expression. We propose a neural network-based method which uses Gaussian noise to model spontaneity in the generation process, removing the need for manual control of conditional generation variables. Our model takes two sequential images as input, with additive noise, and produces the next image in the sequence. We trained two types of models: single-expression, and mixed-expression. With single-expression, unique facial movements of certain emotion class can be generated; with mixed expressions, fully spontaneous expression sequence generation can be achieved. We compared our method to current leading generation methods on a variety of publicly available datasets. Initial qualitative results show our method produces visually more realistic expressions and facial action unit (AU) trajectories; initial quantitative results using image quality metrics (SSIM and NIQE) show the quality of our generated images is higher. Our approach and results are novel in the field of facial expression generation, with potential wider applications to other sequence generation tasks.

INDEX TERMS Affective computing, artificial neural networks, self-supervised learning.

I. INTRODUCTION

Video sequence generation and anticipation is one of the crucial components for estimating likelihood of future events. Potential applications include self-driving cars [13], [21], human fall prediction [30], [42] and data augmentation for sequence-based datasets. Currently available methods require high computational cost and multiple GPUs to train.

In facial expressions generation, most existing works focuses on single image generation [18], [34], [46]. Inspired by Pumarola et al. [23], researchers began to generate facial expression sequences by using a variable to control certain part of the face [10] or facial expression intensity [7]. However, as proven by our empirical experiments, facial expression movements are non-linear. Using a linear variable is an over-simplification on how facial movements works in real life. Linearity does not account for realism of the generated facial expression. We trained our model using a non-guided approach and without using labels. This can mitigate AU labelling errors that is dependent on other methods. Bypassing the need for ground truth labels can prevent the network from registering biases of other methods.

We propose a new approach in training an autoencoder. Our method generates image sequences with temporal correlation on aligned images. Our method is capable of spontaneous generation of facial expressions without any labels across datasets with image sequences (MUG [1] and MMI [31]) and a neutral face image (FFHQ) [17], as illustrated in Fig. 1.

To the best of our knowledge, our method is the only approach that uses two images as input to generate unlimited number of image sequences recursively. We are able to achieve this by leveraging the Markov property of our recursive sequence generation method. Despite the lack of external

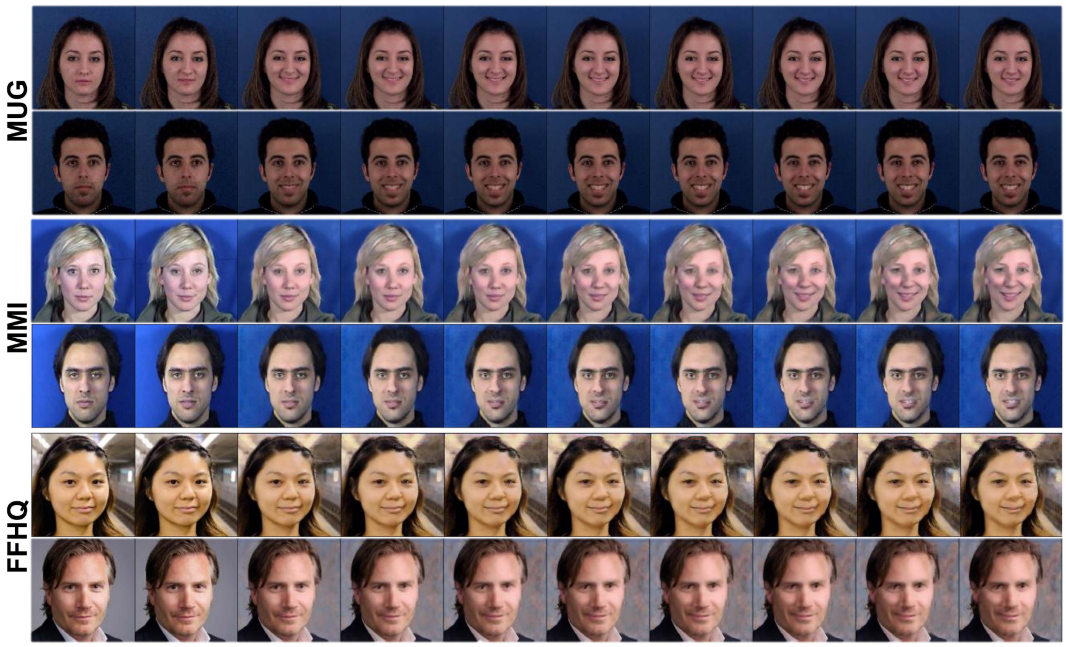

FIGURE 1. We propose an end-to-end facial expressions generation method, with no reference frames or labels to guide the generation process. Our method is able to generate facial expression across different datasets (MUG, MMI and FFHQ dataset) with subjects of different gender and ethnicity. FFHQ is an image dataset, all input frames are from a single neutral face. This demonstrates the ability of our model in generating facial expression sequences on unseen neutral faces.

input guidance, we show that realistic facial expression sequences can be generated end-to-end without facial alignment or facial landmarks detection in a self-supervised manner.

II. RELATED WORK

A. AUTOENCODER

An autoencoder is an algorithm that consists of two main components: an encoder that downsamples the input into a latent representation with lower dimensionality, and a decoder that upsamples the latent representation into output. Majority of sequence generation using autoencoder is on tasks such as language processing or trend lines. Images or video sequence generation using pure autoencoder [25] is not common as generated images/videos are blurry due to information loss of the encoder downsampling. Auto-regressor [11] is a model that regresses on previous values of the time series. Autoencoder can be used to predict the next sequence using a similar manner by regresses on the previous latent representations. A well-known example is the decoder of transformer [32] that function in a similar manner.

There are some attempts in generating facial expression using autoencoders [12], [20], [47], however, the resulting outputs are very low resolution (36×36 pixels) or requires optical flow as an additional input feature. All of the autoencoder based methods require face alignment and cropping.

B. MARKOV CHAIN

Markov chain is a mathematical system that predicts the next state based only on the current state. It is often called "memoryless" stochastic process. It is only based on current observation and previous sequence are not taken into account. The idea is through many iteration of looping of Markov Chain, there will be a convergence to certain value without the need of previous information which is demonstrated by learning through Contrastive Divergence [15]. As long as the chain and inputs obeys Markov property, over time, the output will generalise to a certain value. A few examples of Markov based model are RBM [15], [26] and Deep Belief Network [16].

C. FACIAL EXPRESSIONS GENERATION

Face alignment and face cropping are two pre-processing stages for facial expressions generation tasks. Commonly, the face of an image are extracted as implemented by various approaches [7], [23], [27]. There are other attempts [10], [39] that crop region of interests (i.e., eyes and mouth region) for generation. Even though this pre-processing step reduces computational cost by excluding unwanted image regions, it does not include the context of the faces and has drawback of losing crucial facial information.

Reference frames are often used to guide the facial generation process [44], [45]. These methods implement style transfer on available facial expression datasets to transfer the facial movements to a neutral face, guided by the reference frames. While this approach increases the diversity of the faces, it does not generate novel facial movements.

Another popular approach to generate facial expression is using facial action units as labels [19], [23], [33], [38]. Facial action unit (AU) is the label associated with facial muscle movement, introduced by Ekman in Facial Action Coding System [9]. These approaches use a variable to adjust the

intensity of the targeted expression. GANimation [23] is a generative model guided by AU annotations. This model is able to map facial movements onto a face based on the AUs trained. This method requires an external module to extract AUs from an image sequence for the network to function. Cascade EF-GAN [38] implements network branches named "attention driven local focuses" on top of the AU guided generator. This attention based branches improve identity preservation according to the author. It is widely accepted that AUs can be coded from a face to analyse facial expression. However, using AUs to generate facial expression is not the same. The act of AU extraction introduces a layer of errors and some valuable information might be lost. Imaginator [35] is a conditional GAN that combines motion and content using transposed 3D convolutions to capture spatio-temporal relationships. This network architecture of this method is not dynamic as it can only produce videos with a fixed resolution of 64×64 pixels and a fixed frame number of 32 frames.

Approaches that uses action control variable [10] to manipulate facial expression on image containing face is often conducted using linear action variables. However, these methods are not realistic where the onset and offset of facial expression are not linear.

There are other attempts [5], [6] that control targeted face properties by changing the latent representation. Controlling the latent representation can be useful in generating novel images. These methods has limitation of generating artifacts in between some intensity or overlapping of latent variables.

D. OTHER GENERATION METHODS

There are also abundance of huge models that use extensive computing power for video based generation [24], [43]. These models are too computationally expensive to be implemented by regular users. GAN-based domain adaptation generation method [40] that generate videos based on a guiding video also attempted. However, this approach is fully guided by the input video and is not capable of generating spontaneous new movements.

III. METHODS

A. PROPOSED METHOD

Facial expression sequence generation is defined as generating the next consecutive facial expression sequences. The algorithm is task to predict the next facial expression sequence based a few observed facial expression sequences. Our method considers the previous and current facial expression frames, and attempts to generate the next sequence recursively. To determine the next frame of the facial expression sequence, we use self-supervision that does not require frame-by-frame guided action variables.

Our model is a sequential-based autoencoder, which leverage the latent representation extracted by the encoder to generate subsequent image sequence. Our method was inspired by conditional restricted boltzmann machine [29], where the generation of the next sequence is done by an undirected model with binary latent variables connected to visible variables. Instead of using simple binary latent variable, we obtained latent representations during each downsampling of encoder and passes the latent representations to the decoder.

The network architecture is shown in Fig. 2. Typically, autoencoder-based generation produces blurry images. To overcome this issue, we extract each levels of latent representation as the encoder down-samples the images. These latent representations are then fed to the decoder, which results in detailed images generation. The latent representations contain high- and low-level information that can help in generating images with higher complexity. We found that this step is required in facial expression generation as it is a high complexity task.

Temporal information is captured using our recursive way of training network, where the output frame was fed into our network (as current frame) for the subsequent generation, as illustrated in Fig. 3. By comparing more than one frame sequence, the model is able to figure out the temporal correlation between the sequence by looking a few steps forward into the future frame.

The loss function used in this model are as follows:

$$
\begin{aligned}
P_1 &= D(E(F_c), E(F_1 + N_1), E(F_2 + N_2)) \\
P_2 &= D(E(F_c), E(F_2 + N_2), E(P_1 + N_1)) \\
P_{n>2} &= D(E(F_c), E(P_{n-2} + N_{n-2}), E(P_{n-1} + N_{n-1})) \\
loss &= \frac{1}{n} \sum_{i=1}^{i=n} (P_i - F_{i+2})^2 \qquad (1)
\end{aligned}
$$

where F is the frame originated from the dataset, N is random Gaussian noise, P is the generated frame, n is the number of subsequent frames used (for training only, our experiment uses $n = 3$), i is the recursive loop number. During training, we initialise using F_1 and F_2, sampled from any part of the video (with at least 3 subsequent frames) for training. F_c is a constant neutral face that remains the same throughout the generation process. The generated frame then replaces the current frame with the latter becoming the previous frame as shown in Fig. 3. Gaussian noise added in each loop primarily increases the variability of the generated frame.

Each generation loop is a Markov chain. Even though the recursive loop does not contain any memory, Markov property enables model convergence for sequential generation. (2) shows that the convergence of sequence is possible with sufficient repetition. Each recursive generation is a Markov chain. With the addition of noise, the model is still able to converge as Markov property converges the generation to a certain value.

$$
M(x_{n+1}|x_1, x_2, \ldots, x_n) = M(x_{n+1}|x_n) \qquad (2)
$$

where M is a function with Markov property, x is the input and n is the number of repetition.

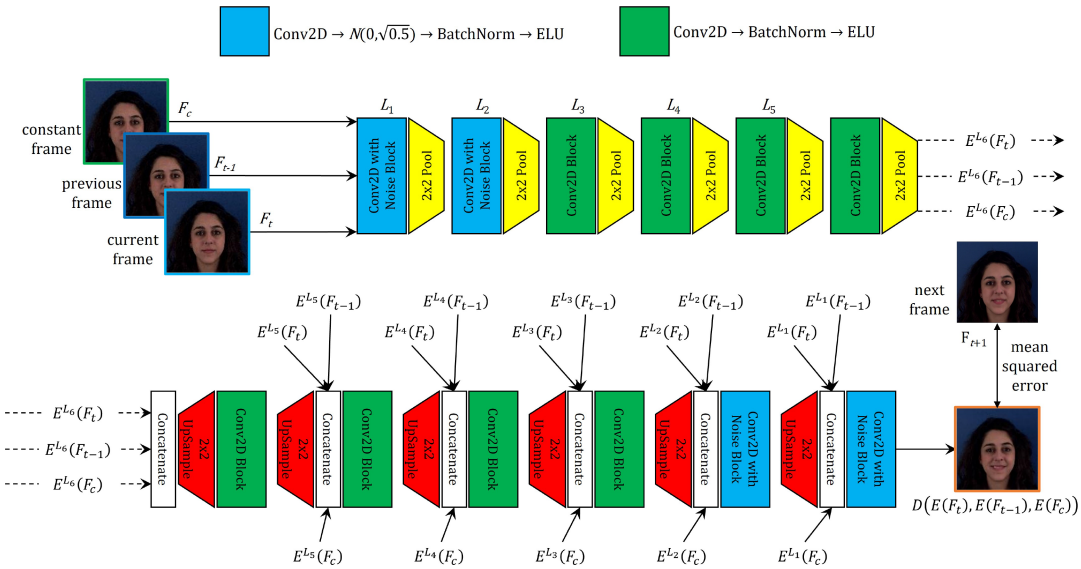

FIGURE 2. Our network architecture that takes three input frames (F_c, F_{t-1}, F_t) and predicts the next frame using auto-regression. Auto-regression of this model is performed by training our network recursively (refer to Fig. 3). The top part is the encoder while the bottom part is the decoder. The latent representation, L, extracted by each downsampling are fed into the decoder layers. Gaussian noise is used in the shallow layers.

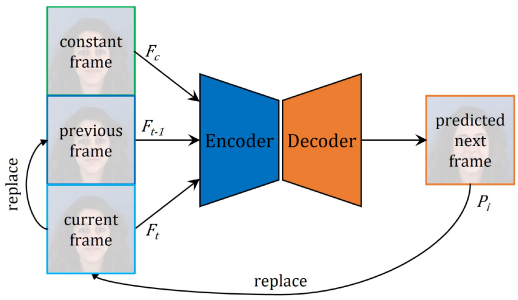

FIGURE 3. Recursive generation loop. This loop feeds the output as input for the next iteration. This arrangement allows the generation of the next sequence by replacing the output frame as current frame and current frame as previous frame in the next iteration.

Assuming x is a probability of data point, θ is model parameter and f is a certain function. The probability of X is defined as $p(x,\theta)$ as in (3).

$$p(x,\theta) = \frac{1}{Z(\theta)} f(x,\theta) \tag{3}$$

The partition function of $f(x,\theta)$, Z is defined as in (4).

$$Z(\theta) = \int f(x,\theta)dx \tag{4}$$

The model parameters can be learned by minimising the negative log-likelihood of probability, $p(x,\theta)$. This negative log-likelihood is also known as the energy function, $E(x,\theta)$ in (5).

$$\begin{aligned} E(x,\theta) &= -\log\left(\frac{1}{Z(\theta)} f(x,\theta)\right) \\ &= \log\ Z(\theta) - \frac{1}{K}\sum_{i=1}^{K} \log\ f(x_i,\theta) \end{aligned} \tag{5}$$

The derivative of contrastive divergence [15] (adapted from Woodford [37]) contains only the information of current and one of any previous state, as shown in (6). In our implementation, we use one previous step. This is the main advantage of this method whereby the energy function of the sequence can be minimised using only the current and one previous state.

$$\frac{\partial E(x,\theta)}{\partial\theta} = \left\langle \frac{\partial \log(f(x,\theta))}{\partial\theta} \right\rangle_{x_0} - \left\langle \frac{\partial \log(f(x,\theta))}{\partial\theta} \right\rangle_{x_1} \tag{6}$$

Hence, by calculating the derivative of contrastive divergence, the model parameters can be learn as in (7) where t is the time step.

$$\theta_{t+1} = \theta_t + \eta\left(\left\langle \frac{\partial \log(f(x,\theta))}{\partial\theta} \right\rangle_{x_0} - \left\langle \frac{\partial \log(f(x,\theta))}{\partial\theta} \right\rangle_{x_1}\right) \tag{7}$$

where η is the learning rate. The model parameter can be found using only one previous step and current step of observation until the derivative of contrastive divergence is converged (equal to zero). Theoretically, the model will converge after n cycles without the need of any previous memory or recurrent mechanism. We incorporated Gaussian noise on the weights of the shallow layers of our model. The addition of noise addresses overfitting and adds variability to our output data.

B. DATASETS

We use the MUG facial expression dataset (MUG) [1] as our training set. The image sequences in this dataset consist of frontal faces. The image sequences were labelled with six basic emotions (anger, disgust, fear, happiness, sadness and surprise). Neutral face sequences of each subjects were also captured. The frame rate of this dataset is 19 fps, with each image sequence ranging from 50 to 160 images.

FIGURE 4. Single expression sequence generation frame-by-frame across 6 emotion classes (anger, disgust, fear happiness, sadness, surprise). Our network is able to generate targeted emotion class based on the training data. The identity of the subject remained the same throughout the generation iterations. Note: First and second images are the input frames from MUG dataset, all the subsequent frames are generated.

We conduct facial expression generation experiments on MMI Facial Expression Database (MMI) [31] and Flickr-Faces-HQ dataset (FFHQ) [17]. MMI consists of posed facial expression sequences. FFHQ is an image dataset of human face images, which intended to benchmark the performance of generative model. We use FFHQ to test our model's ability in generating expression from static images.

C. EVALUATION METRICS

AU Comparison Facial action unit (AU) [8] is an objective measure to describe human facial movements. We compare normalised AU intensity of actual and generated facial expressions from onset to offset to check the realism of generated expression.

Image Quality Assessment Structural-similarity index (SSIM) [36] is a full reference metric (where images are compared pixel by pixel) that measures similarity between two images. Natural Image Quality Evaluator (NIQE) [22] is a no-reference metric. It is based on the construction of a quality aware collection of statistical features using space domain natural scene statistic model.

IV. EXPERIMENTS AND RESULTS

A. IMPLEMENTATION DETAILS

Our method is implemented using Tensorflow. ADAM optimiser with β_1 of 0.9 and β_2 of 0.999 was used. The learning rate is set to 0.001. The dropout rate is 0.2 and the standard deviation of Gaussian noise is set to 0.5. The network architecture is shown in Fig. 2. For training, we use 5 consecutive frames (2 initial frames for initialisation and 3 subsequent frames to capture spatiotemporal information). One constant neutral frame is fed to the network in every iteration for identity preservation. Mean squared error (MSE) is also computed

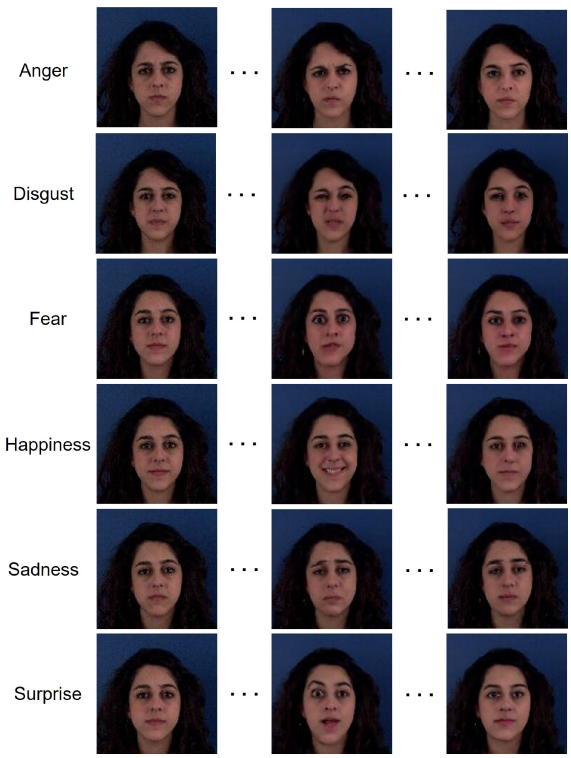

FIGURE 5. Single expression sequence generation of different emotional phases (onset, apex, offset) across 6 emotion classes (anger, disgust, fear happiness, sadness, surprise). Our method is capable of completing the entire sequence of facial expressions. Note: First image (onset frame) is the input frames from MUG dataset.

for each subsequent frame and generated frame. For evaluation, we use 2 consecutive frames and one constant frame for initialisation. For the number of frames generated we can

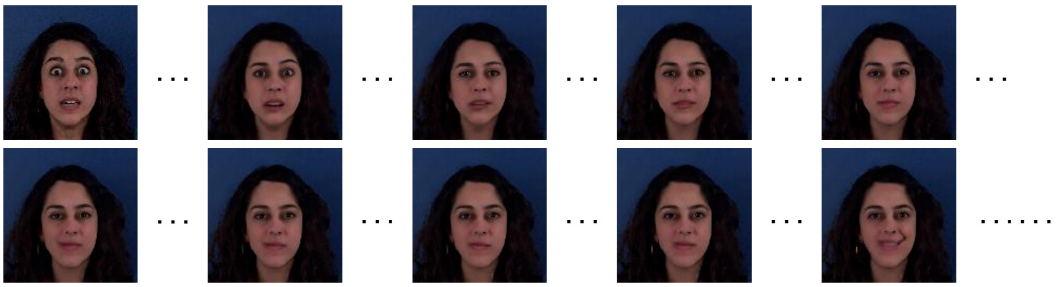

FIGURE 6. Spontaneous multi-expression sequence generation. Initialised using only 2 frames (fear expression). This version is trained using all six expressions. The generated sequences consisted of fear, followed by slight mouth movement (left side) and a broad grin.

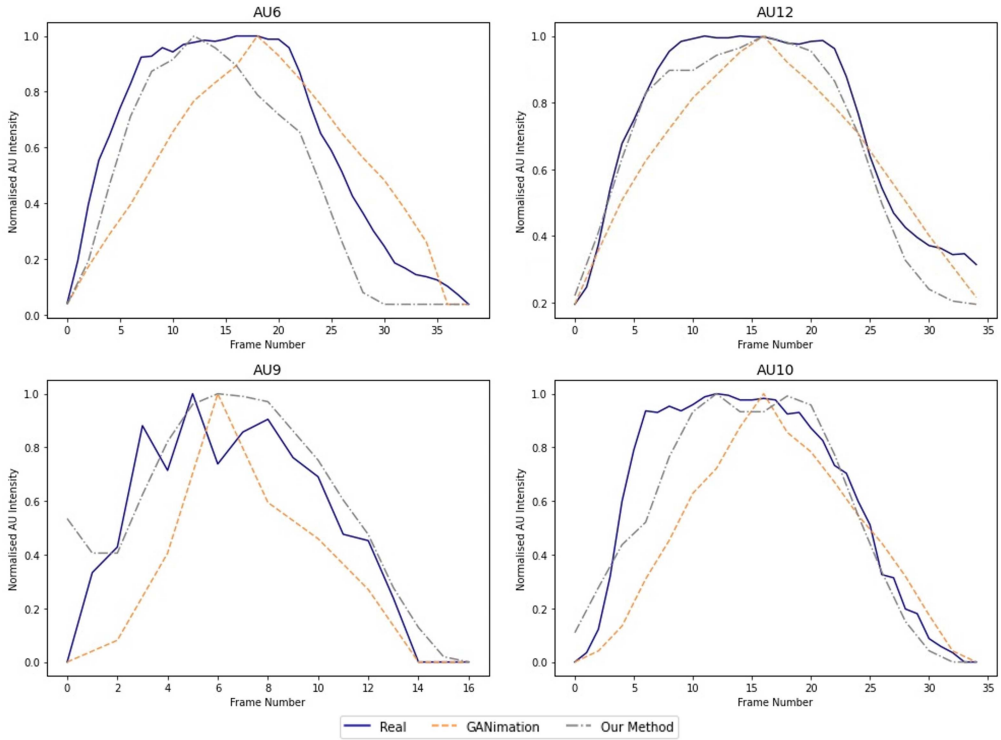

FIGURE 7. AUs comparison of real and generated sequences measured using OpenFace. Our method closely resembles AU of real facial expression with non-linear onsets and offsets. However, GANimation, a linear interpolation based facial expression generation method, shows constant AU intensity increment or decrement. GANimation generates expression on a per image basis and is fully guided by intensity input which differs from real sequence as shown. Note: AU6 and AU12 are from "Happiness" while AU9 and AU10 are from "Disgust".

defined using a parameter, n_pred, which is set to 50, it can also be adjusted by the user.

All our networks are trained without any labels. We use single category of facial expression (for Single Expression Sequence Generation) and all six basic emotional classes (for Spontaneous Multi-Expression Sequence Generation).

B. FACIAL EXPRESSION GENERATION

Single Expression Sequence Generation Our model is able to generate realistic sequence (as shown in Fig. 4) from a static frame and complete the expression sequence that involves onset, apex, and offset phases (as shown in Fig. 5). The identity of the participant is preserved throughout the generation. Our model is also able to generate across dataset with faces of different gender and ethnicity as shown in Fig. 1. Each generated sequence are temporally correlated and when the expression reaches its apex, it is able to return back to the offset phase (neutral).

Spontaneous Multi-Expression Sequence Generation This network is trained with all six basic emotional classes. In Fig. 6, our model is able to generate multiple emotion classes of facial expression with two initial frames ("Fear" class). This shows the ability of our model to generate realistic sequence across different expression. This demonstrates that our network understands the realism of facial expression based on the training data.

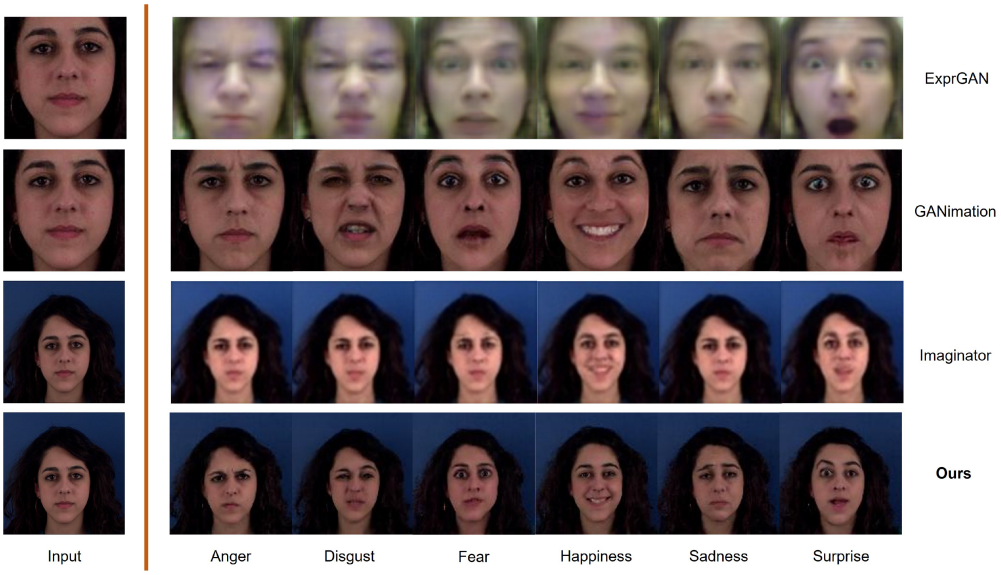

FIGURE 8. **Comparison of single facial expression generation on MUG dataset. Six basic emotions are tested. ExprGAN completely changes the identity of the participants. Note that GANimation failed to generate "Surprise" class. ExprGAN and GANimation require facial alignment and uses a linear variable to tune the facial expression intensity. Imaginator can generate data without facial alignment, however, "Anger", "Fear" and "Sadness" expression generated are visually similar to "Disgust" class. Our method is able to generate realistic expression based on each emotion class without any labels or guidance.**

C. AU COMPARISON

We compare the AU extracted by OpenFace [3] for real and generated sequences. In Fig. 7, the results show our method resembles real facial expression when compared to GANimation. However, as demonstrated in Fig. 7, facial expression of real sequence is non-linear. Hence, GANimation lacks spontaneity, which is a common issue in guided facial expression generation. Our approach bypasses the need of labels and the use of linear variable. Our results are more realistic as the network learns directly from the raw data and potential biases from another model (for AU extraction) can be avoided.

D. COMPARISON OF SIX BASIC EMOTIONS

We compare the appearance of the apex of the facial expression with ExprGAN, GANimation, and Imaginator [35]. Both of ExprGAN and GANimation requires facial alignment and face crop that may remove certain facial part (especially the chin). They also uses linear variable to augment the facial expression intensity. Although ExprGAN claims their model is able to preserve the identity of the input face, but based on the model uploaded by the author, the results is contradictory. Whilst GANimation is able to preserve the identify, it generates "Surprise" as "Anger". Imaginator can generate data without facial alignment and face crop. Nonetheless, the facial expressions depicting "Anger", "Fear", and "Sadness" are visually reminiscent of the facial expression associated with the "Disgust" category. We show that we are able to generate facial expressions without the need of facial alignment or face crop. Our generated results are also novel as our approach allows the network to decide the facial expression that represent each emotion which is a stark contrast to the common approach that tune a linear variable to generate facial movements.

E. ABLATION STUDIES

Removing the constant frame The effect of generation without using a constant frame can be seen in Fig. 9. The identity of the face changes slightly. The facial expression quality also reduces drastically as the generation goes on. This shows the constant frame is essential for retaining the complex facial features.

Changing input order We investigate the effect of input frames sequence by swapping the order of input 1 and input 2. From Fig. 10, we observe that the output results are not the same. Hence, we conclude that the input sequence matters. Our model has no issue with completing and regenerating more facial expression sequence in both cases.

F. IMAGE QUALITY ASSESSMENT

The results are shown in Fig. 11. Our method outperforms all other methods in Natural Image Quality Evaluator (NIQE) while exhibit similar performance in Structural-similarity index (SSIM) with GANimation and Imaginator. These image quality analysis are performed over full video sequences. For reference-based assessment (i.e. SSIM), GANimation, Imaginator and our model has similar average performance. For no-reference based assessment (i.e. NIQE), our model performs the best. This shows the sequences generated by our model has the highest image quality which is spontaneous and novel to the input frame (as this metric is not based on reference position).

G. COMPUTATIONAL COST

The computational cost of the model is calculated using floating point operations (FLOPs). Our model has a complexity of 18.01 GFLOPs. Our model complexity are close to image based model which sits between Inceptionv3 [28] (11

FIGURE 9. Ablation study: remove constant neutral frame. By removing the constant frame, the face identity changes and image become unrecognisable in latter stages of generation.

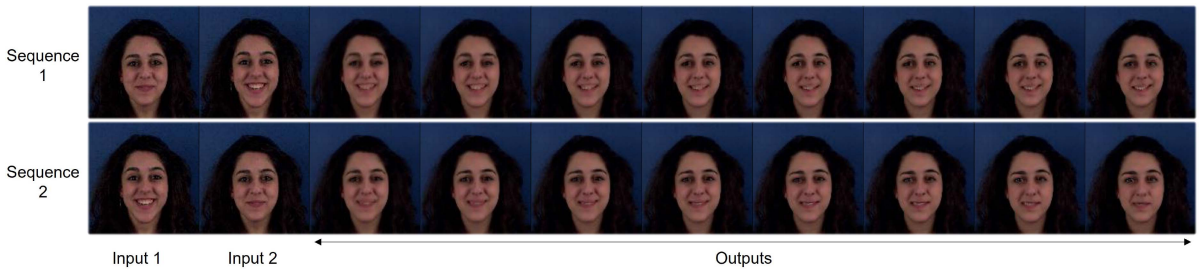

FIGURE 10. Ablation study: Changing input order. The effect of input order compared by swapping input 1 and input 2. The input order of the model matters as the generations are different.

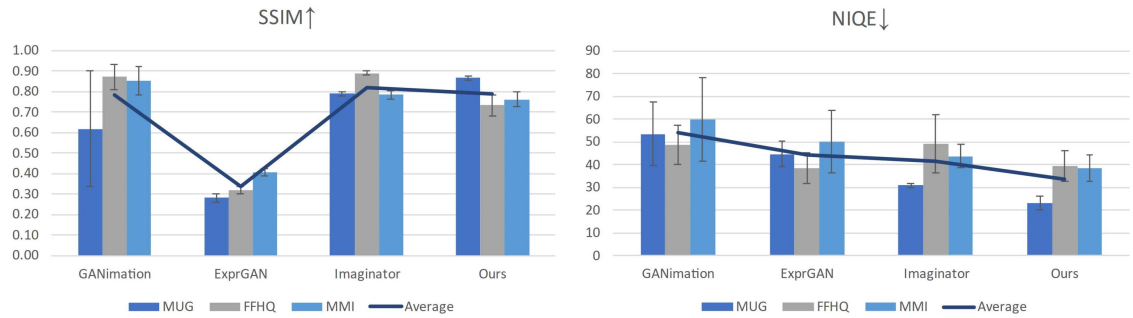

FIGURE 11. Image Quality Assessment using MUG, FFHQ and MMI as input. (Left) Comparison of SSIM, the higher the better. Our model has similar average performance across 3 datasets with GANimation and Imaginator. (Right) Comparison of NIQE, the lower the better. Our model has the best average performance across 3 datasets. SSIM and NIQE are explained in Section III-C.

GFLOPs) and ResNet [14] (21 GFLOPs). As a comparison, video based models e.g. S3D [41] has a complexity of 71.38 GFLOPs, I3D [4] uses 107.89 GFLOPs and Vision Video Transformer (ViViT) [2] has over 4000 GFLOPs.

V. CONCLUSION

We presented a novel latent representation based model for facial expression sequences (of different emotion class) generation. Our model does not rely on any external guidance or labels to generate spontaneous facial expression sequence. It is able to recursively generate and complete the entire sequence of facial expression with only two sequential input frames and a neutral frame. We demonstrate that our model trained with single-expression can generate unique facial movements and model trained with mixed expressions is able to generate fully spontaneous expression sequences. We show that our generated sequence closely resembles real facial expression using AUs comparison and image quality of our generated images is higher compared to other facial expression generation methods. We will run our network architecture on other different sequence or video dataset and investigate further on metrics to quantify the spontaneity of expression generated.

REFERENCES

[1] N. Aifanti, C. Papachristou, and A. Delopoulos, "The MUG facial expression database," in *Proc. IEEE 11th Int. Workshop Image Anal. Multimedia Interactive Serv.*, 2010, pp. 1–4.
[2] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6836–6846.
[3] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *Proc. IEEE 13th Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 59–66.
[4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.
[5] Y.-C. Chen, X. Xu, Z. Tian, and J. Jia, "Homomorphic latent space interpolation for unpaired image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2408–2416.
[6] Y. Deng, J. Yang, D. Chen, F. Wen, and X. Tong, "Disentangled and controllable face image generation via 3D imitative-contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5154–5163.
[7] H. Ding, K. Sricharan, and R. Chellappa, "ExprGAN: Facial expression editing with controllable expression intensity," in *Proc. AAAI Conf. Artif. Intell.*, 2018.
[8] P. Ekman and W. V. Friesen, *Facial Action Coding System: Investigator's Guide*. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.
[9] R. Ekman, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System*. London, U.K.: Oxford Univ. Press, 1997.

[10] L. Fan, W. Huang, C. Gan, J. Huang, and B. Gong, "Controllable image-to-video translation: A case study on facial expression generation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 3510–3517.

[11] K. Gregor, I. Danihelka, A. Mnih, C. Blundell, and D. Wierstra, "Deep autoregressive networks," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1242–1250.

[12] J. Han, M. R. Min, L. Han, L. E. Li, and X. Zhang, "Disentangled recurrent wasserstein autoencoder," in *Proc. Int. Conf. Learn. Representations*, 2020.

[13] P. A. Hancock, I. Nourbakhsh, and J. Stewart, "On the future of transportation in an era of automated and autonomous vehicles," *Proc. Nat. Acad. Sci.*, vol. 116, no. 16, pp. 7684–7691, 2019.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[15] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, Aug. 2002.

[16] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[17] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4401–4410.

[18] J. Kossaifi, L. Tran, Y. Panagakis, and M. Pantic, "GAGAN: Geometry-aware generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 878–887.

[19] J. Ling, H. Xue, L. Song, S. Yang, R. Xie, and X. Gu, "Toward fine-grained facial expression manipulation," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 37–53.

[20] Y. Liu, X. Hou, J. Chen, C. Yang, G. Su, and W. Dou, "Facial expression recognition and generation using sparse autoencoder," in *Proc. IEEE Int. Conf. Smart Comput.*, 2014, pp. 125–130.

[21] Q. Memon, M. Ahmed, S. Ali, A. R. Memon, and W. Shah, "Self-driving and driver relaxing vehicle," in *Proc. IEEE 2nd Int. Conf. Robot. Artif. Intell.*, 2016, pp. 170–174.

[22] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[23] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 818–833.

[24] R. Rakhimov, D. Volkhonskiy, A. Artemov, D. Zorin, and E. Burnaev, "Latent video transformer," in *Proc. 16th Int. Joint Conf. Comput. Vis. Imag. Comput. Graph. Theory Appl.*, 2021, pp. 101–112.

[25] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3D faces using convolutional mesh autoencoders," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 704–720.

[26] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," Boulder Dept. of Comput. Sci., Colorado Univ., Boulder, CO, USA, Tech. Rep. CU-CS-321-86, 1986.

[27] K. Songsri-in and S. Zafeiriou, "Face video generation from a single image and landmarks," in *Proc. IEEE 15th Int. Conf. Autom. Face Gesture Recognit.*, 2020, pp. 69–76.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[29] G. W. Taylor and G. E. Hinton, "Factored conditional restricted boltzmann machines for modeling motion style," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 1025–1032.

[30] L. Tong, Q. Song, Y. Ge, and M. Liu, "HMM-based human fall detection and prediction method using tri-axial accelerometer," *IEEE Sensors J.*, vol. 13, no. 5, pp. 1849–1856, May 2013.

[31] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: An addition to the MMI facial expression database," in *Proc. 3rd Intern. Workshop Emotion Satell. LREC: Corpora Res. Emotion Affect*, Paris, France, 2010, Art. no. 65.

[32] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.

[33] F. Wang, S. Xiang, T. Liu, and Y. Fu, "Attention based facial expression manipulation," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2021, pp. 1–6.

[34] X. Wang, X. Wang, and Y. Ni, "Unsupervised domain adaptation for facial expression recognition using generative adversarial networks," *Comput. Intell. Neurosci.*, 2018, Art. no. 7208794.

[35] Y. Wang, P. Bilinski, F. Bremond, and A. Dantcheva, "Imaginator: Conditional spatio-temporal GAN for video generation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 1160–1169.

[36] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[37] O. Woodford, "Notes on contrastive divergence," *Dept. of Eng. Sci.*, Univ. Oxford, London, U.K., Tech Rep., 2006.

[38] R. Wu, G. Zhang, S. Lu, and T. Chen, "Cascade EF-GAN: Progressive facial expression editing with local focuses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5021–5030.

[39] Y. Xia, W. Zheng, Y. Wang, H. Yu, J. Dong, and F.-Y. Wang, "Local and global perception generative adversarial network for facial expression synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1443–1452, Mar. 2022.

[40] S. Xiang, Y. Fu, G. You, and T. Liu, "Unsupervised domain adaptation through synthesis for person re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2020, pp. 1–6.

[41] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 305–321.

[42] Q. Xu, G. Huang, M. Yu, and Y. Guo, "Fall prediction based on key points of human bones," *Phys. A: Stat. Mechanics Appl.*, vol. 540, 2020, Art. no. 123205.

[43] W. Yan, Y. Zhang, P. Abbeel, and A. Srinivas, "Videogpt: Video generation using VQ-VAE and transformers," 2021, *arXiv:2104.10157*.

[44] C. Yang and S.-N. Lim, "Unconstrained facial expression transfer using style-based generator," 2019, *arXiv:1912.06253*.

[45] C. H. Yap, R. Cunningham, A. K. Davison, and M. H. Yap, "Synthesising facial macro-and micro-expressions using reference guided style transfer," *J. Imag.*, vol. 7, no. 8, pp. 142, 2021.

[46] Y. Zhang, R. Liu, Y. Pan, D. Wu, Y. Zhu, and Z. Bai, "GI-AEE: GAN inversion based attentive expression embedding network for facial expression editing," in *Proc. IEEE Int. Conf. Image Process.*, 2021, pp. 2453–2457.

[47] Y. Zhou and B. E. Shi, "Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder," in *Proc. IEEE 7th Int. Conf. Affect. Comput. Intell. Interact.*, 2017, pp. 370–376.