3D Facial Expression Generator Based on Transformer VAE

Kaifeng Zou, Boyang Yu, Hyewon Seo

3D FACIAL EXPRESSION GENERATOR BASED ON TRANSFORMER VAE

Kaifeng Zou   *Boyang Yu*   *Hyewon Seo*

ICube Laboratory, CNRS–University of Strasbourg, France

ABSTRACT

We present a generative model for the 3D facial expression mesh sequences, from onset to the termination of a desired expression. We tailor a Transformer VAE architecture: The encoder compresses a sequence of facial landmarks into an expression-aware regularized latent space, while the decoder generates a new sequence from the sampled latent variable, conditioned on a desired expression. After a landmark-guided mesh deformation, a given 3D neutral face is driven to an animated mesh sequence with the expected expression. The generated sequences are consistent, of quality, and exhibit a good level of diversity, improving over state-of-the-art methods. We validate our model by conducting extensive experiments on two representative datasets. The supplementary video and code are available on a GitHub page (`https://github.com/ZOUKaifeng/FacialExpressionGeneration`).

Index Terms— transformer, variational autoencoder, facial expression generation

1. INTRODUCTION

The prevailing shape capture technology has paved the way for data-driven approaches of facial animation. A common strategy is to use multi-view systems that can capture full shape geometry and appearance either in real-time or offline, depending on the target application and the desired quality of the data. Although the results thus obtained are often impressively realistic, they are limited to capturing existing shapes, necessitating generative models or retargeting methods to obtain new, imaginary faces and/or varying expression styles.

Recent deep learning techniques have become dominant solutions in facial modeling tasks, and have achieved impressive performance. In particular, facial expression synthesis in 2D image domain [1] has been boosted by the generative deep learning models like Generative Adversarial Networks (GANs) [2] and Variational Autoencoders (VAEs) [3]. Although these methods are capable of synthesizing realistic expression poses, most of them address the generation of static expression poses, and that too in 2D. 2D-to-3D facial reconstruction approaches partly overcome these limitations, by performing per-frame 3D reconstruction from video input. However, such a frame-based approach neglects the temporal aspect of animation, which is a vital element of the sequence data. Some of the recent generative models leverage auto-regressive models, such as LSTMs [4], GRUs[5], or Transformers[6], for the temporal encoding of animation sequences. However, with a few exceptions[7], they mostly focus on the human body motion [8, 9].

In this paper, we address the challenging problem of 3D dynamic facial expression generation. The specific aim is to generate a sequence of appearance-preserving facial expressions, conditioned on a categorical expression. We deploy a conditional version of Transformer-based encoder-decoder architecture, trained with the VAE losses to learn the distribution of the facial expressions. It is inspired by the recent success of Transformer in the sequence generation in different application scenarios[10, 8] and the conditional generation of VAE and its variants[11].

After training, our model can generate temporally consistent sequences of 3D facial landmarks, conditioned on a desired expression label. Similarly to most existing works[7, 12], we train our generative model with many sequences of landmarks sampled on 3D face meshes. To obtain the full mesh deformation, we then adopt the sparse to dense decoder (S2D-Dec) proposed in [7] to apply the geometric deformation encoded in landmarks to a given facial mesh at its neutral pose.

2. APPROACH

Our method works in two stages. Firstly, a Transformer VAE is trained to perform the conditional generation of landmarks sequences. Then the S2D-Dec estimates the vertex-wise displacements of a neutral face mesh for each frame of the landmark set in the generated sequence, in a frame-by-frame manner. The overview of our method is shown in Fig.1.

2.1. Expression representation

Most of the few available 3D facial expression datasets [13, 14] come in the form of sequences of dense triangular meshes, each containing thousands of vertices. It is tedious and takes too long to train a generative model directly using all the ver-

This work was partly funded by the ANR19-CE23-0020 Human4D project of the French National Research Agency (ANR). Kaifeng Zou has been supported by ANR19-CE45-0015 TOPACS project.

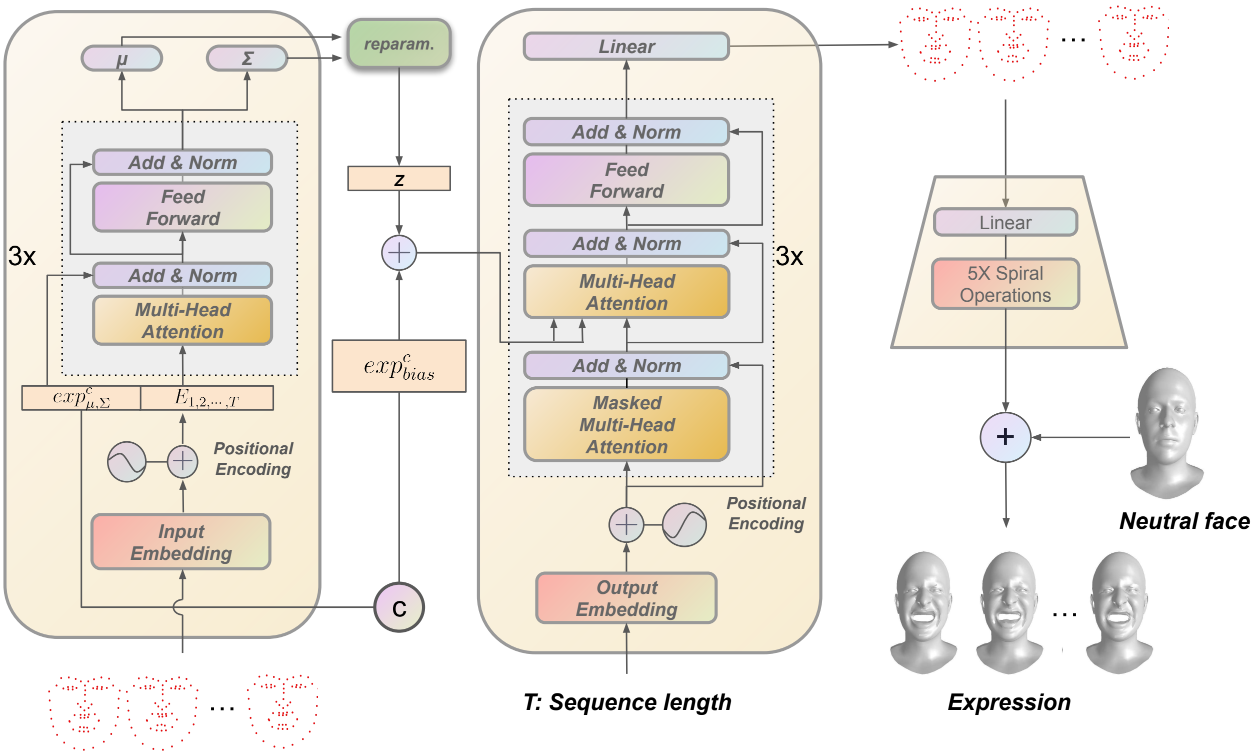

Fig. 1: Overview of our models.

tices, so we use a set of predefined facial landmarks instead. It is a viable and compact way to represent facial expressions by using feature points or landmarks on the face outline and on the contours of eyes, nose, and mouth, etc.

Once the sequence of landmarks is in place, one can use them to deform a neutral face mesh either by applying some geometric method or by using a learned model, both in a per-frame manner. Here we use the GCN-based S2D-Dec to predict per-vertex displacements of the facial mesh given a landmark set. The deformation of the input neutral face mesh is obtained by adding the displacements to its vertices (see Fig.1). Note that the magnitude of the expression can be changed by scaling a coefficient to the displacements.

2.2. Transformer VAE

The key idea of VAE is to regularize the distribution in the latent space of an autoencoder. As in the vanilla VAE [15], we define a weak prior $p(z)$ over z as a standard distribution $\mathcal{N}(0, I)$. As we denote a sequence of facial landmarks by $x = [L_1, L_2, ..., L_T]$ and its expression label by c, the generative process can be written as:

$$p_\theta(x, c, z) = p_\theta(x|c, z)p(c)p(z). \tag{1}$$

We assume that $p(c) = 1/K$ is a categorical distribution, with K being the number of expression classes.

The posterior $p(z|x, c)$ is also modeled as $\mathcal{N}(0, I)$. Since it is intractable[3], a variational approximation $q_\phi(z|x, c)$ is learned by an encoder neural network, with ϕ standing for the network's parameters. Supposing that z and c are independent, it has been shown that the objective optimization is achieved by maximizing the evidence lower bound (ELBO)[15]:

$$E_{z\sim q_\phi(z|x,c)}\left[\log \frac{p_\theta(x, c, z)}{q_\phi(z \mid x, c)}\right]. \tag{2}$$

Our objective derived from Eq.2 is:

$$\begin{aligned}\mathcal{L}(\theta, \phi; x, c) = & -\lambda KL(q_\phi(z|x, c), p(z))+ \\ & E_{q_\phi(z|x,c)}(log(p_\theta(x|c, z)),\end{aligned} \tag{3}$$

where KL is the Kullback–Leibler divergence that measures the difference between the latent distribution and the normal distribution. The minimization of KL enforces the latent distribution to follow the standard distribution. The second term can be interpreted as a reconstruction loss in the implementation, with $p(x)$ modeled as a Gaussian distribution. As in β-VAE [16], we have a hyperparameter λ to balance the reconstruction and the stochastical property of the latent space.

While using VAE to generate sequences of landmarks, the temporal information is encoded into the latent space using Transformer. This allows the model to learn the underlying

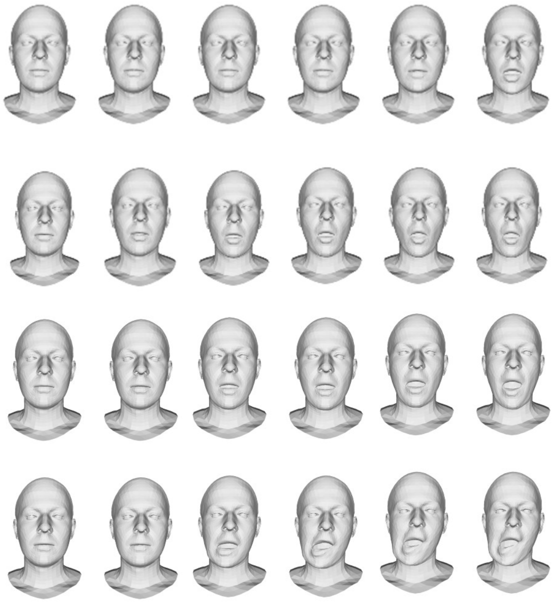

Fig. 2: Conditional generation with the labels 'mouth open' (first two rows) and 'mouth side' (last two rows). Two sequences are shown per each class to demonstrate the generation diversity of our model. More examples are shown in our project page.

patterns and relationships between different landmarks over time. As shown in Fig.1, our generative model inherits its global architecture from the classic variational autoencoder [15] with stacks of Transformer components inside. While based on the vanilla Transformer [6], we further incorporate the expression condition information into the encoding of parameters(μ, Σ) for the distribution $q_\phi(z|x,c)$. More precisely, for each facial expression label c, there are two expression label features denoted as exp_μ^c and exp_Σ^c. Rather than directly feeding c to the encoder, its associated expression label features (exp_μ^c and exp_Σ^c) are prepended with the input sequence embedding before feeding into the encoder. Such use of label features has been introduced previously in other domains for pooling purposes[10], especially for time pooling in [8].

The first two frames of our encoder outputs (corresponding to the two expression label features of the input) serve as the latent distribution parameters: mean μ and variance Σ. The rest of the outputs are ignored, which acts as time pooling.

A latent vector z is sampled from the estimated distribution using the reparameterization trick[3]. It is then shifted by a per-label learnable expression feature exp_{bias}^c and fed into the decoder as 'key' and 'value' to calculate the cross attention. The decoder then generates the sequence of landmarks corresponding to the expression label added in the latent space. The decoder also takes time information as a query, consisting of an all-zero vector of size T, the output frame length. In our work, we set it identically to the input sequence length.

3. EXPERIMENTS

3.1. Training details

The proposed method was tested on two 4D face datasets, with some preprocessing to adapt each for our model. We provide further details below.

CoMA dataset [14] is a commonly used 3D facial expression dataset in face modeling tasks [17, 18]. It consists of more than a hundred 3D animated head models captured from twelve subjects, each performing 12 facial expressions. An expression data contains a sequence of triangular meshes of 5023 vertices, undergoing some deformation elicited by an expression. 68 landmarks sampled on the full mesh have been used for training. Then, a similar method to [7] has been used to extract sub-sequences from each sequence to align them temporally in the semantic sense, e.g. from neutral to apex (the most expressive state). The selected subsequences have different frame lengths, for which we have applied subsampling and linear interpolation to obtain a uniform length.

BU-4DFE dataset[13], another widely used dataset in facial expression synthesis, contains a total of 606 sequences (101 subjects, each performing 6 basic emotional expressions). A sequence of 83 landmarks manually selected on 3D facial scans is made available. With some exceptions, all sequences had been more or less aligned temporally, so we only performed frame length uniformization to 100 frames.

Implementation details. The encoder and the decoder of our Transformer VAE follow the conventional architecture of the original Transformer[6], with 4 heads in each attention module and an embedding dimension of 256. We use a linear layer followed by 5 layers of spiral operation for the graph convolution in S2D-Dec. Each model has been trained separately, both using Adam [19] optimizer. The learning rate has been set to a constant $1e-4$ for S2D-Dec, whereas there is a warm-up step[20] for Transformer VAE.

Cross validation. To evaluate the model's ability to handle unseen data, we adopted the nested cross-validation strategy. The dataset has been splitted into 5 folders, with an even distribution of labels in each folder. Then, we train five models separately, each using three folders as training dataset while sparing one as validation set and another as test set. The mean and variance of the results from all five models are taken as the final result.

Table 1: FID scores and classification accuracy of different models.

	CoMA		BU-4DFE	
Model	Accuracy	FID	Accuracy	FID
Groud truth	$83.78\% \pm 4.23\%$	2.77 ± 0.62	$99.51\% \pm 0.65\%$	6.02 ± 1.05
CondLSTM [12]	$8.33\% \pm 0.09\%$	92.36 ± 6.36	$16.69\% \pm 0.05\%$	101.02 ± 9.29
Action2Motion [9]	$52.36\% \pm 8.58\%$	29.44 ± 4.98	$80.83\% \pm 3.75\%$	19.6 ± 4.52
Motion3DGAN [7]	69.44%	19.01	-	-
Ours	$81.40\% \pm 2.47\%$	7.11 ± 1.24	$99.13\% \pm 1.07\%$	14.56 ± 1.92

3.2. Experimental results

We can generate diverse expressions with our decoder by sampling the latent representation from the Gaussian distribution. It is confirmed by the good level of diversity shown in the generated sequences using a same class label, as demonstrated in Fig.2.

We compare our Transformer VAE with several contenders in terms of conditional generation, by evaluating their classification accuracy and the FID score. Three other methods are chosen for the comparison: Motion3DGAN [7], Action2Motion [9], and CondLSTM. The latter is an adapted model of [12] that takes the facial shape and the expression label together as input. The results of Motion3DGAN [7] are directly taken from their paper. In order to make a fair comparison, we use the same classifier as theirs: one LSTM layer followed by a fully connected layer.

The quantitative results are shown in Table 1. The FID of ground truth is calculated between the test dataset and the training dataset. It reveals that CondLSTM fails to handle multi-class sequence generation, as evidenced by the poor accuracy ($< 9\%$). We observe that the Action2Motion model requires a large amount of data to achieve optimal performance. With only a small dataset, consisting of a few hundred examples in contrast to the several thousand in their original motion project, the model seems to have learned the expression behavior only partially. For instance, their generated face does not always return to its neutral pose when tested on the BU-4DFE dataset.

Although our classifier takes the same architecture, it yields a higher accuracy than the one used in Motion3DGAN during the evaluation. This may be partially due to the differences in data preprocessing. Nevertheless, we note that the generation accuracy of our model has been obtained on a set of abundant size: (288 samples, each conditioned on 12 labels, amounting to a total of 3456) whereas Motion3DGAN[7] is tested only on 144 sequences. Meanwhile, our model shows outstanding performance on the BU-4DFE dataset, which confirms the competitiveness of our model. In addition, our model shows stable performance in handling different test sets, with the smallest variances among all the methods.

3.3. Ablation study

We ablate several components of the Transformer VAE to emphasize the contribution of each unit. First, we compare the Transformer with another popular recurrent neural network, GRU [5]. A GRU-VAE inheriting its architecture from that of Transformer VAE has been used, with both the encoder and the decoder consisting of 3 GRU layers.

Another design element worth investigating is the label feature. We replaced the label feature setting by a label variable while keeping the basic architecture of Transformer VAE, i.e. the label value has been directly placed in the latent space. The quantitative results on the experiments we conducted with BU-4DFE dataset are shown in Table 2. We can observe that both the Transformer and the label feature setting are indispensable for the improved performance.

Table 2: Ablation study for Transformer VAE.

Model	Accuracy	FID
GRU-VAE	$71.07\% \pm 5.74\%$	28.70 ± 8.31
w/o label feature	$93.49\% \pm 3.87\%$	17.28 ± 3.65
Transformer VAE	$99.13\% \pm 1.07\%$	14.56 ± 1.92

4. CONCLUSION

We have presented a generative model for the synthesis of 3D dynamic facial expressions, The dynamics of facial expressions has been successfully learned by using an expression-sensitive latent representation, from which we can randomly sample instances with an expression category to synthesize diverse expression sequences. The proposed method can produce realistic face meshes of diverse types of expression on different subjects, outperforming SOTA models both qualitatively and quantitatively.

5. REFERENCES

[1] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer, “Ganimation: Anatomically-aware facial animation from a

single image," in *Proc. European conference on computer vision*, 2018, pp. 818–833.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[3] Diederik P. Kingma and Max Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2014.

[4] Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 11 1997.

[5] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *SSST@EMNLP*, 2014.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[7] Naima Otberdout, Claudio Ferrari, Mohamed Daoudi, Stefano Berretti, and Alberto Bimbo, "Sparse to Dense Dynamic 3D Facial Expression Generation," in *Conference on Computer Vision and Pattern Recognition*, New Orleans, United States, June 2022.

[8] Mathis Petrovich, Michael J. Black, and Gül Varol, "Action-conditioned 3D human motion synthesis with transformer VAE," in *IEEE International Conference on Computer Vision (ICCV)*, October 2021, pp. 10985–10995.

[9] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng, "Action2motion: Conditioned generation of 3d human motions," in *Proc. ACM Multimedia*, 2020, pp. 2021–2029.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] Kaifeng Zou, Sylvain Faisan, Fabrice Heitz, and Sebastien Valette, "Joint disentanglement of labels and their features with vae," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 1341–1345.

[12] Hyewon Seo and Guoliang Luo, "Generating 3d facial expressions with recurrent neural networks," in *Intelligent Scene Modeling and Human-Computer Interaction*, pp. 181–196. Springer International Publishing, 2021.

[13] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu, "A high-resolution spontaneous 3d dynamic facial expression database," in *IEEE workshops on automatic face and gesture recognition*, 2013, pp. 1–6.

[14] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black, "Generating 3d faces using convolutional mesh autoencoders," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 704–720.

[15] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling, "Semi-supervised learning with deep generative models," in *Advances in neural information processing systems*, 2014, pp. 3581–3589.

[16] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," 2016.

[17] Giorgos Bouritsas, Sergiy Bokhnyak, Stylianos Ploumpis, Michael Bronstein, and Stefanos Zafeiriou, "Neural 3d morphable models: Spiral convolutional networks for 3d shape representation learning and generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7213–7222.

[18] Zi-Hang Jiang, Qianyi Wu, Keyu Chen, and Juyong Zhang, "Disentangled representation learning for 3d face shape," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11957–11966.

[19] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] Martin Popel and Ondřej Bojar, "Training tips for the transformer model," *arXiv preprint arXiv:1804.00247*, 2018.