Autoregressive GAN for Semantic Unconditional Head Motion Generation

Louis Airale, Xavier Alameda-Pineda, Stéphane Lathuilière, Dominique Vaufreydaz

AUTOREGRESSIVE GAN FOR SEMANTIC UNCONDITIONAL HEAD MOTION GENERATION

AUTHOR VERSION

Louis Airale[1,2], **Xavier Alameda-Pineda**[2], **Stéphane Lathuilière**[3], **Dominique Vaufreydaz**[1]
[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[2] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
[3] LTCI, Télécom Paris, Institut polytechnique de Paris, France

ABSTRACT

We address the task of unconditional head motion generation to animate still human faces in a low-dimensional semantic space. Deviating from talking head generation conditioned on audio that seldom puts emphasis on realistic head motions, we devise a GAN-based architecture that allows obtaining rich head motion sequences while avoiding known caveats associated with GANs. Namely, the autoregressive generation of incremental outputs ensures smooth trajectories, while a multi-scale discriminator on input pairs drives generation toward better handling of high and low frequency signals and less mode collapse. We demonstrate experimentally the relevance of the proposed architecture and compare with models that showed state-of-the-art performances on similar tasks.

Keywords GAN, Head motion, Face landmarks

1 Introduction

Talking head generation refers to the task of animating a human face generally using a single reference image, an audio clip, and possible additional conditioning signals such as emotional state or exemplar pose dynamics [1, 2]. Differently from face reenactment where a driving video clip is provided, here head pose, facial animation and lip synchronization need to be inferred from other modalities. To tackle the difficulty of handling both facial dynamics and photorealism directly in image space, a predominant line of research generates dynamics in a lower dimensional space [3]. Those representations comprise supervised facial landmarks [4], 3D mesh [5] or unsupervised keypoints [6], and in the present work we refer generally to them as the *semantic space*. Although several works achieved compelling results in lip-syncing and realistic rendering, generating natural head motions has, until recently, consistently received less attention. We shift focus and propose to tackle the task of unconditional head motion sequence generation (i.e. head pose and facial expression) given only a reference pose. This task is relevant in applications where no audio signal is available, e.g. when animating background characters in a scene or a video game. In addition, it is reasonable to assume that a proficient unconditional head motion generator would constitute a solid baseline for modeling realistic head motion in a standard talking head generation setting. Similar to [7, 5], we address face generation in a semantic space, in our case 2D facial landmarks.

Head motion generation is a continuous sequence prediction problem. For this task, it is a viable choice to produce first order derivatives, or instantaneous velocities [8], allowing for simpler model architectures [9]. This strategy benefits from autoregressive generation, which provides an inductive bias for modeling cumulative sum operations [10]. However, autoregressive models trained with a mean squared error loss tend to predict mean values as time stretches [9], advocating for the use of other loss functions. We hereby introduce an adversarial loss to tackle the unconditional head motion generation problem. Head motion dynamics are structured data composed of temporal patterns that evolve over varied timescales. Previous works have addressed structured data generation with the help of discriminator networks operating on receptive fields of different sizes [11, 12] or on local windows, enabling a better representation of high frequency components [13]. We build on this knowledge and use a multi-scale, window-based discriminator, but noticeably implement it in a single network, allowing to flexibly incorporate any new resolution. Last, to mitigate mode collapse we follow [14] and provide input pairs to the discriminator network, but also produce samples together in the generator. While this does not change the optimization objective, it brings a significant performance boost for a limited additional overhead. The proposed GAN model, labelled SUHMo (for Semantic Unconditional Head Motion), allows for long-term head motion synthesis, and experiments confirm its proficiency against a diversity of models and baselines.[1]

[1] Source code and results can be found at: https://github.com/LouisBearing/UnconditionalHeadMotion.

2 Related work

Talking head generation aims at syncing driving audio with head motions, and has seen tremendous recent progress [4, 6]. Although early identified as a key component for faithful face animation [7], the prediction of head pose and facial expression beyond lip region has been noticeably less investigated, in favor of the use of a driving head motion sequence [2]. As it is a one-to-many mapping, learning to generate head motion from audio is challenging, and the usual mean squared error loss typically produces static average poses. Successful attempts at handling head poses include [15, 1, 4], although the range of achieved motions remains limited. Recently, [6] presented natural-looking results with head pose and face expression produced in a sparse keypoints manifold by two separately trained modules. Compared to these works, our model generates all semantic data in a single module, learning possible correlations between pose and expression.

Deep continuous autoregressive models are ubiquitous in sequence modeling, as they enable strong temporal consistency thanks to the explicit relation between consecutive outputs. In the context of deep continuous sequence prediction, autoregressive models proved powerful in as diverse domains as waveform synthesis [12], human trajectory prediction in a crowd [8], or human motion prediction [9]. Surprisingly, the talking face generation literature is much sparser on this subject, [5] presenting one of the few autoregressive talking head generation architectures.

Uncovering multiple patterns with GANs was first addressed by [13] that introduced a discriminator network taking generated image patches as input to enhance high spatial frequency components. In [11], an output image pyramid is processed by several discriminators that operate on decreased resolutions and larger receptive fields, driving the generator network to produce realistic patterns at different scales. The multi-scale discriminator has then been extended to sequence generation tasks [16, 12]. An interesting aspect of the latter discriminator architectures is that they combine multi-scale with window-based evaluations in a 1D equivalent of [13], and benefit from the side advantages it offers, such as lighter architecture and faster inference.

Mode collapse reduction methods in GANs have comprised efforts towards better optimization procedures [17], generation space regularization [18], or forcing the network to account for the noise vector [19], among a rich body of literature. [14] proposed an intuitive way of driving the generator to produce diverse outputs by feeding the discriminator with several input samples. We extend this framework by generating these inputs *together*, which provided better results while leaving the optimization objective unchanged.

3 Autoregressive unconditional head motion generation

In this section, we formally define the task we address and the key components of our learning framework. We work with 2D facial landmarks since this low-dimensional representation allows for fast training and inference and can be mapped back to the image domain [20]. Given a set of points $\boldsymbol{x}_0$ representing a face in an initial pose, we seek to generate a sequence $\boldsymbol{x}_{1:T} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_T)$ of arbitrary length T such that the probability distributions of the generated and the ground truth data, p_{G} and p_{D}, match:

$$p_{\mathrm{G}}(\boldsymbol{x}_{0:T}) = p_{\mathrm{D}}(\boldsymbol{x}_{0:T}), \quad \forall \boldsymbol{x}_{0:T}. \tag{1}$$

We achieve this through autoregressive generation of instantaneous velocities. Formally, the proposed generator G outputs one frame at a time based on previous frames:

$$\boldsymbol{x}_t = G(\boldsymbol{x}_{0:t-1}) = \boldsymbol{x}_{t-1} + \hat{G}(\boldsymbol{x}_{0:t-1}). \tag{2}$$

3.1 Unified multi-scale discriminator

We use a multi-scale, window-based discriminator to train the model to generate temporal patterns unfolding over different timescales. Similarly to [21], we rely on a single network, considerably simplifying the discriminator architecture, yet we extend their work with a principled approach independent from the exact form of the network. Concretely, the full discriminator D_{S} on sequences is defined as an expectation over multi-scale assessments of a network D_{M} on temporal slices:

$$D_{\mathrm{S}}(\boldsymbol{x}_{0:T}) = \mathbb{E}_{\tau,t}[D_{\mathrm{M}}(\boldsymbol{x}_{t:t+\tau})], \tag{3}$$

where τ and t are the duration and starting index of the slice. In practice both t and τ are sampled from discrete uniform distributions. The advantage of this framework is that a single network D_{M} operates on multiple time scales.

3.2 Joint generation and discrimination

To improve the performances and mitigate the mode collapse problem, we consider both the generation and discrimination of joint sample distributions. Let the objective, with generic data points $\boldsymbol{x}$ and $\boldsymbol{y}$, write (superscript J for joint distributions):

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p^{\mathrm{J}}_{\mathrm{D}}}[\log D(\boldsymbol{x},\boldsymbol{y})] + \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p^{\mathrm{J}}_{\mathrm{G}}}[\log(1 - D(\boldsymbol{x},\boldsymbol{y}))] \tag{4}$$

This has to be minimised (resp. maximised) w.r.t. the parameters of the generator G (resp. the discriminator D). In the case of an i.i.d. data distribution and enough networks capacity, the joint generated distribution converges to the product of the marginal data distributions [22]:

$$p^{\mathrm{J}}_{\mathrm{G}}(\boldsymbol{x},\boldsymbol{y}) = p^{\mathrm{J}}_{\mathrm{D}}(\boldsymbol{x},\boldsymbol{y}) = p_{\mathrm{D}}(\boldsymbol{x})p_{\mathrm{D}}(\boldsymbol{y}). \tag{5}$$

If G produces samples independently, then $p^{\mathrm{J}}_{\mathrm{G}}$ readily factorizes. This is the setting of [14], which proved useful to reduce mode collapse. However if x and y are produced together, then G simply *learns* to factorize. Both cases

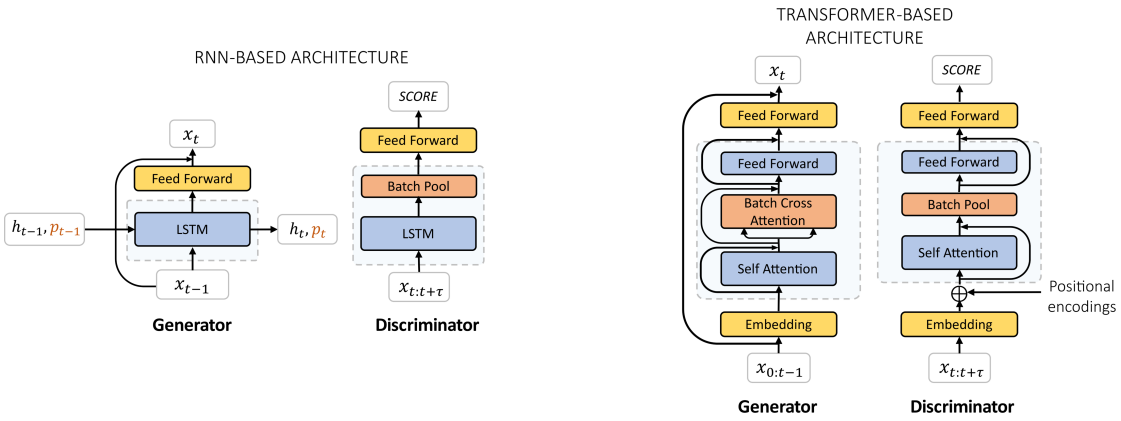

Figure 1: The two architecture variants of the proposed SUHMo model. We use layer normalization in the Transformer and in the feed forward layer of the LSTM generator, but we omit it for the sake of clarity. See the text for a more detailed description.

lead to the equality of marginal distributions $p_{\mathrm{G}} = p_{\mathrm{D}}$, hence the optimization objective of [22] is unaffected. We argue that in the real case scenario of limited network capacity, if the generation is prone to mode collapse then the overall optimization can benefit from this joint generation process. In such cases, it is an easy task for D to identify generated pairs by comparing the two samples, hence driving G to leverage its two inputs to increase the generation diversity.

3.3 Architecture

Training loss. The overall discriminator consists of three networks. In addition to a joint multi-scale discriminator $D_{\mathrm{S}}^{\mathrm{J}}$ on sequence pairs, we add a marginal multi-scale discriminator D_{S} on single sequences to explicitly enforce the convergence of the marginal distributions, and a frame discriminator D_{F} on each frame of ground truth and generated samples. The loss function corresponds to the sum of the three associated losses, plus a supervised mean squared error term that we scale such that the adversarial losses remain dominant.

Implementation. We implement two variants of our model, based on LSTM and Transformers [23]. An overview of both architectures can be found in Figure 1. To account for pairs of inputs, we define a batch-pool operator that acts as a max pooling layer of kernel size 2 along the batch dimension; with the difference that the result can be repeated for the two samples to preserve the input batch size.

In the LSTM-based generator, the hidden state $\boldsymbol{h}_t$ goes through a batch-pool layer, yielding a pooled vector $\boldsymbol{p}_t$ that is concatenated with the next input to the LSTM. A multi-layer perceptron is used on $\boldsymbol{h}_t$ to output the landmark positions. The joint discriminator $D_{\mathrm{S}}^{\mathrm{J}}$ is composed of a LSTM, a batch-pool layer and a feed forward network; the marginal discriminator D_{S} is similar but without the batch-pool layer.

In the Transformer generator, pair mixing is done in a multi-head attention layer that takes input pairs stacked in the batch dimension as queries, and the same pairs in reversed order as keys and values. This way, each sample in a pair can attend to the history of the other sample. We label this layer batch-cross attention. We do not use positional encoding as no performance drop was observed, while this allows the generation of longer test sequences. As for the discriminator networks, a batch-pool layer replaces the batch-cross attention in $D_{\mathrm{S}}^{\mathrm{J}}$ as it only needs to provide a single score per pair. A learnable class token, prepended to the input sequence, is processed through a last feed forward network to give the final score [24].

Model inputs. It is possible to concatenate velocities or instantaneous accelerations to landmark positions as input to the generator or the discriminators. This might for instance help penalize static sequences produced by G. In practice, we use positions and velocities as inputs to the generator and all three quantities in the discriminator networks. At test time, since the model expects pairs of samples but only one reference frame is given, we augment the input, e.g. with random flip, rescaling or translation, and concatenate it to the reference pose. This also gives a practical way of injecting stochasticity into the generation process.

4 Experiments

Datasets. We conduct experiments on two video datasets with upper-body frontal views of different speakers. CONFER [25] contains 72 video clips of TV debates between two persons, each about 1 minute long. We preprocessed the data preserving head translations and selected 5 clips as test data featuring persons unseen at training. Second, we extracted and averaged over random subsets from VoxCeleb2 [26], leaving 674 video clips as test set. Models were trained to predict sequences of 40 time steps, given 1 to 5 initial frames.

Table 1: Model comparison on the head motion generation task. Sequences of 40 and 80 time steps are generated from a single input frame (subscripts indicate length). In both cases, metrics are computed over the 40 last predicted time steps.

	CONFER						VoxCeleb2					
Model	FVD_{40}	FID_{40}	$t\text{-}FID_{40}$	FVD_{80}	FID_{80}	$t\text{-}FID_{80}$	FVD_{40}	FID_{40}	$t\text{-}FID_{40}$	FVD_{80}	FID_{80}	$t\text{-}FID_{80}$
HiT-DVAE	$368^{\pm 19}$	$6^{\pm 0.4}$	$130^{\pm 7}$	$764^{\pm 35}$	$50^{\pm 2}$	$157^{\pm 12}$	$686^{\pm 37}$	$\mathbf{1^{\pm 0.1}}$	$167^{\pm 4}$	$644^{\pm 27}$	$\mathbf{2^{\pm 0.1}}$	$164^{\pm 6}$
ACTOR	$480^{\pm 12}$	$8^{\pm 0.3}$	$147^{\pm 3}$	$667^{\pm 20}$	$9^{\pm 0.8}$	$163^{\pm 5}$	$357^{\pm 55}$	$4^{\pm 0.5}$	$78^{\pm 9}$	$431^{\pm 26}$	$5^{\pm 2}$	$145^{\pm 21}$
Δ-based	$318^{\pm 115}$	$21^{\pm 3}$	$67^{\pm 10}$	$357^{\pm 104}$	$24^{\pm 3}$	$77^{\pm 18}$	$386^{\pm 32}$	$48^{\pm 6}$	$89^{\pm 4}$	$518^{\pm 48}$	$60^{\pm 30}$	$112^{\pm 31}$
L_2-only	$480^{\pm 42}$	$10^{\pm 3}$	$133^{\pm 2}$	$777^{\pm 54}$	$21^{\pm 3}$	$159^{\pm 6}$	$530^{\pm 20}$	$2^{\pm 0.2}$	$158^{\pm 6}$	$684^{\pm 23}$	$8^{\pm 0.8}$	$172^{\pm 9}$
SUHMo - RNN	$\mathbf{162^{\pm 31}}$	$\mathbf{3^{\pm 0.2}}$	$\mathbf{61^{\pm 8}}$	$\mathbf{147^{\pm 45}}$	$8^{\pm 2}$	$\mathbf{48^{\pm 11}}$	$\mathbf{76^{\pm 8}}$	$3^{\pm 0.7}$	$\mathbf{21^{\pm 3}}$	$\mathbf{135^{\pm 33}}$	$9^{\pm 5}$	$\mathbf{31^{\pm 7}}$
SUHMo - Transformer	$175^{\pm 46}$	$4^{\pm 0.7}$	$67^{\pm 12}$	$169^{\pm 33}$	$\mathbf{7^{\pm 1}}$	$52^{\pm 4}$	$134^{\pm 33}$	$3^{\pm 0.8}$	$42^{\pm 10}$	$141^{\pm 31}$	$9^{\pm 3}$	$55^{\pm 16}$

Metrics. Landmark positions are rasterized into image sequences, on which we measure Fréchet video (FVD) [27] and inception distances (FID) [28]. While FID gives a score of static face realism, FVD measures the smoothness of the dynamics. We also introduce a second dynamical metric based on a FID measure on sequences where we squeeze the time dimension. To do so, we compute an exponential moving average centered on the last frame, thus enforcing a visual correlation between pixel intensity and time step index. The resulting score, denoted t-FID (t standing for time), proved relevant to discriminate sequences with little movement.

Models comparison. The performances of SUHMo were compared with two state-of-the-art architectures for human pose prediction, **HiT-DVAE** ([29]) and **ACTOR** ([30]). A minimal amount of changes was required to adapt these models to our setting. In particular, action conditioning was replaced by the observed initial frame(s). In addition, we implemented two baselines taken from the talking head generation literature. The **Δ-based** model, similarly to [4], produces displacements from a fixed set of reference points, in this case the initial landmark positions. **L_2-only** follows a common trend in head motion prediction and relies on a single mean squared error loss. We evaluate the above models and our two architecture variants on both CONFER and VoxCeleb2, on sequences of duration 40 and 80 frames. Results are reported in table 1. SUHMo consistently outperforms all other architectures in terms of dynamics quality. HiT-DVAE and ACTOR attain lower FID values on VoxCeleb2, suggesting slightly sharper faces, but this is at the cost of producing quasi-static sequences, hence the poor FVD and t-FID scores. The same is true for models trained with a L_2 reconstruction loss, advocating for the use of an adversarial loss to ensure realistic dynamics. The Δ-based variant produces dynamics of uneven quality, as per the high standard deviations, and the realism of produced faces falls significantly behind, as suggests the higher FID values.

Multi-scale discriminator ablation. To measure the ability of SUHMo to produce realistic patterns over diverse time scales we compute the FVD on 10, 20, and 40-frames chunks and compare with a single-scale discriminator on full sequences, see table 2. Models were trained to generate 40-frames sequences and perform more or less on par on this duration. The benefit of the window-based multi-scale approach however clearly appears on shorter timescales, indicating finer modeling of high frequency patterns.

Table 2: FVD scores over different sub-sequences lengths, with and without multi-scale window-based D (CONFER).

Chunk length (frames)	10	20	40
Full (RNN)	$\mathbf{28^{\pm 8}}$	$\mathbf{35^{\pm 4}}$	$162^{\pm 31}$
w/o window-based D	$35^{\pm 8}$	$42^{\pm 8}$	$\mathbf{157^{\pm 21}}$
Full (Transformer)	$\mathbf{34^{\pm 9}}$	$\mathbf{40^{\pm 12}}$	$\mathbf{175^{\pm 46}}$
w/o window-based D	$57^{\pm 6}$	$60^{\pm 12}$	$236^{\pm 58}$

Table 3: Two-samples strategy ablation results on CONFER.

Sampling	RNN			Transformer		
	FVD_{40}	FID_{40}	$t\text{-}FID_{40}$	FVD_{40}	FID_{40}	$t\text{-}FID_{40}$
Full model	$\mathbf{162^{\pm 31}}$	$\mathbf{3^{\pm 0.2}}$	$\mathbf{61^{\pm 8}}$	$175^{\pm 46}$	$\mathbf{4^{\pm 0.7}}$	$67^{\pm 12}$
1 sample D	$226^{\pm 76}$	$3^{\pm 1}$	$71^{\pm 16}$	$\mathbf{162^{\pm 36}}$	$7^{\pm 1}$	$\mathbf{65^{\pm 11}}$
1 sample G	$222^{\pm 23}$	$5^{\pm 0.7}$	$58^{\pm 7}$	$237^{\pm 58}$	$8^{\pm 2}$	$74^{\pm 10}$

Double samples strategy ablation. We tried removing the pair mixing in the generator and the discriminator at turns (table 3). Although the one-sample discriminator matches the full model dynamics scores for the Transformer, enabling pair mixing in both networks achieves overall significantly better performances, supporting the proposed approach.

5 Conclusion

In this paper we presented SUHMo, an architecture for semantic unconditional head motion generation, able to animate facial landmarks over long sequences given a single initial frame. The proposed framework can be implemented in different ways, and outperforms state-of-the-art pose generation methods and head motion prediction baselines regarding the quality of produced motions. In a future work we plan to evaluate the relevance of this framework in a regular setting of audio-conditioned talking head generation.

Acknowledgement

This research was partially funded in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02), and by the SPRING project supported by the European Commission under the Horizon 2020 framework program for Research and Innovation (H2020-ICT-2019-2, GA #871245).

References

[1] R. Yi, Z. Ye, J. Zhang, H. Bao, and Y. Liu, "Audio-driven talking face video generation with learning-based personalized head pose," *arXiv preprint arXiv:2002.10137*, 2020.

[2] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *CVPR*, 2021.

[3] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee, "Learning to generate long-term future via hierarchical prediction," in *ICML*, 2017.

[4] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "Makelttalk: speaker-aware talking-head animation," *TOG*, vol. 39, 2020.

[5] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "Faceformer: Speech-driven 3d facial animation with transformers," in *CVPR*, 2022.

[6] S. Wang, L. Li, Y. Ding, and X. Yu, "One-shot talking face generation from single-speaker audio-visual correlation learning," in *AAAI*, 2022.

[7] D. Greenwood, I. Matthews, and S. Laycock, "Joint learning of facial expression and head pose from speech," in *Interspeech*, 2018.

[8] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *CVPR*, 2018.

[9] J. Martinez, M. J Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *CVPR*, 2017.

[10] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, and Y. Bengio, "Chunked autoregressive GAN for conditional waveform synthesis," in *ICLR*, 2022.

[11] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *CVPR*, 2018.

[12] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *NeurIPS*, 2019.

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[14] Z. Lin, A. Khetan, G. Fanti, and S. Oh, "Pacgan: The power of two samples in generative adversarial networks," *NeurIPS*, 2018.

[15] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, and C. Xu, "Talking-head generation with rhythmic head motion," in *ECCV*, 2020.

[16] X. Lin and M. R Amer, "Human motion modeling using dvgans," *arXiv preprint arXiv:1804.10652*, 2018.

[17] M. Arjovsky, S. Chintala, and . Bottou, "Wasserstein generative adversarial networks," in *ICML*, 2017.

[18] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li, "Mode regularized generative adversarial networks," *arXiv preprint arXiv:1612.02136*, 2016.

[19] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *NeurIPS*, 2016.

[20] E. Zakharov, A. Ivakhnenko, A. Shysheya, and V. Lempitsky, "Fast bi-layer neural synthesis of one-shot realistic head avatars," in *ECCV*, 2020.

[21] L. Airale, D. Vaufreydaz, and X. Alameda-Pineda, "Socialinteractiongan: Multi-person interaction sequence generation," *IEEE TaffC*, 2022.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, 2017.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019.

[25] C. Georgakis, Y. Panagakis, S. Zafeiriou, and M. Pantic, "The conflict escalation resolution (confer) database," *Image and Vision Computing*, vol. 65, 2017.

[26] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[27] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv preprint arXiv:1812.01717*, 2018.

[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *NeurIPS*, 2017.

[29] X. Bie, W. Guo, S. Leglaive, L. Girin, F. Moreno-Noguer, and X. Alameda-Pineda, "Hit-dvae: Human motion generation via hierarchical transformer dynamical vae," *arXiv preprint arXiv:2204.01565*, 2022.

[30] M. Petrovich, M. J Black, and G. Varol, "Action-conditioned 3d human motion synthesis with transformer vae," in *ICCV*, 2021.