HAL open science

Autoregressive GAN for Semantic Unconditional Head Motion Generation

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

To cite this version:

Louis Airale, Xavier Alameda-Pineda, Stéphane Lathuilière, Dominique Vaufreydaz. Autoregressive GAN for Semantic Unconditional Head Motion Generation. ACM Transactions on Multimedia Computing, Communications and Applications, 2024, pp.1-11. 10.1145/3635154 . hal-03833759v3

HAL Id: hal-03833759

https://inria.hal.science/hal-03833759v3

Submitted on 19 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

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

In this work, we address the task of unconditional head motion generation to animate still human faces in a low-dimensional semantic space from a single reference pose. Different from traditional audio-conditioned talking head generation that seldom puts emphasis on realistic head motions, we devise a GAN-based architecture that learns to synthesize rich head motion sequences over long duration while maintaining low error accumulation levels. In particular, 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 experimentally demonstrate the relevance of the proposed method and show its superiority compared to models that attained 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 [33, 31, 21, 42, 48]. Different from face reenactment where a driving video clip is provided, in talking head generation the 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 the image space, a predominant line of research generates dynamics in a lower dimensional space [36]. Those representations comprise supervised facial landmarks [7, 49], 3D mesh [13] or unsupervised keypoints [39, 40], and following the designation of *high level semantics* used in [36], we refer to this space 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. In the lack of a driving audio signal, it is yet crucial for the synthesis model to produce natural and diverse head motions. This is relevant in applications where no audio signal is available, e.g. when animating background characters in a scene or a video game. In this unconditional generation setting, the focus shifts from audio-visual synchrony toward long-term consistency of unguided generation, which is known to be particularly challenging [37]. Tackling this problem can also be beneficial for audio-conditioned talking head synthesis, as it fosters fine handling of head dynamics. In this work, we address the task of unconditional head motion sequence generation, i.e. synthesizing head pose and facial expression given a single reference pose and no audio driving signal. We do so using a low dimensional *semantic* representation, namely 2D facial landmarks, that facilitates the manipulation of head dynamics and can easily be mapped back to the image domain [44, 45, 46, 27].

Continuous sequence prediction problems such as head motion generation have previously been addressed by producing first order derivatives, e.g. instantaneous velocities instead of positions, for different tasks such as human trajectory prediction [17] or human pose generation [24, 23]. One advantage of generating residual quantities is that it can allow using shallower networks [26]. Second, it is conveniently modeled by autoregressive generators which provide an inductive bias for cumulative sum operations [28]. However, autoregressive models may accumulate error, or alternatively end up generating average values over time when trained with a mean squared error loss [26], which

advocates for the use of other loss functions. We hereby introduce an adversarial framework to tackle head motion generation as an autoregressive velocity prediction problem, which to the best of our knowledge has never been done before in talking head generation. To that end, we carefully designed our discriminator network by taking advantage of the specificity of head motion data. Head motion dynamics are structured data composed of temporal patterns that evolve over varied timescales. Previous works have addressed structured data generation with discriminator networks operating on receptive fields of different sizes [41, 22] or on local windows, enabling a better representation of high-frequency components [19]. We build on this knowledge and use a multi-scale, window-based discriminator, but noticeably implement it in a single network, allowing us to flexibly incorporate any new resolution. Last, to mitigate mode collapse we follow Lin et al. [25] 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 architecture, labeled Semantic Unconditional Head Motion or SUHMo, allows for long-term head motion synthesis, and experiments confirm its proficiency against a diversity of models and baselines.[1]

The contributions of this research work are:

- An autoregressive GAN framework for unconditional head motion generation in the 2d-landmarks domain, able to mitigate error accumulation over long sequences, even extending the duration of training sequences,
- A training methodology that can be generalized over diverse architectures, for which we detail two implementations based on LSTM and Transformers,
- Extensive experiments showing that the proposed SUHMo method surpasses competitive methods from closely related tasks on two benchmark datasets.

2 Related work

2.1 Talking head generation

Talking head generation aims at syncing driving audio with head motions, and has seen tremendous recent progress [7, 47, 38, 49, 40, 30]. Although early identified as a key component for faithful face animation [16], 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 [43, 48, 20]. 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 [8, 42, 49], although the range of achieved motions remains limited. Recently, Wang et al. [39] presented natural-looking results with head pose and face expression produced in a sparse keypoints manifold by two separately trained modules, and further extended their work in Wang et al. [40]. In comparison, our model generates all semantic data in a single module, learning possible correlations between pose and expression, and uses an autoregressive formulation to enforce temporal consistency.

2.2 Deep continuous autoregressive models

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 [22], human trajectory prediction in a crowd [17], or human motion prediction [26, 24, 23, 2]. Surprisingly, the talking face generation literature is much sparser on this subject, Fan et al. [13] presenting one of the few autoregressive talking head generation architectures, but they do not attempt to generate head motions. Different from previous works, we leverage the potential of autoregressive models to produce smooth and realistic head motions.

2.3 Multi-scale generative adversarial networks

Uncovering multiple patterns with GANs was first addressed in Isola et al. [19] where the authors introduced a discriminator network taking image patches as input to enhance high spatial frequency components. In Wang et al. [41], 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 [24, 22]. An interesting aspect of the latter discriminator architectures is that they combine multi-scale with window-based evaluations in a 1D equivalent of PatchGAN [19], and benefit from the advantages of processing short windows, such as a lighter architecture and faster inference. Our

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

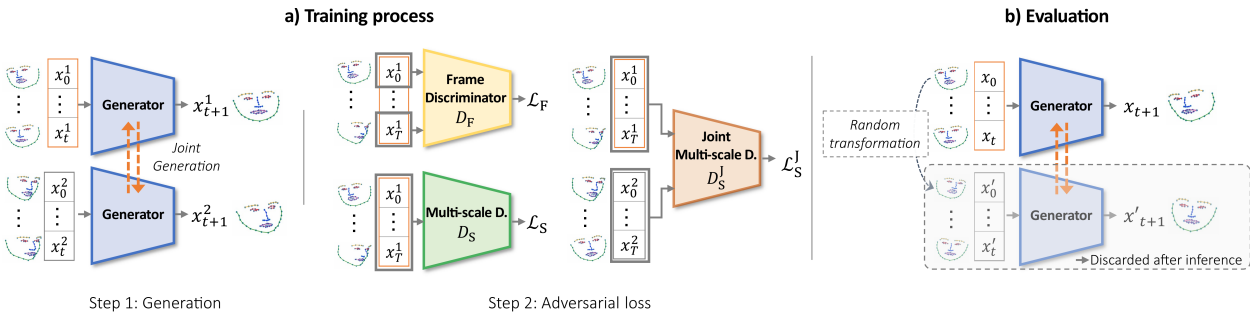

Figure 1: Overview of SUHMo training process. The autoregressive generator produces pairs of outputs, that are evaluated by three discriminator networks. At test time, the second sample is replaced by a transformed version of the reference pose.

window-based multi-scale discriminator follows that of Airale et al. [1], which has the additional advantage of being very flexible regarding the evaluation scales, for a fixed number of parameters.

2.4 Mode collapse mitigation

Mode collapse reduction methods in GANs have comprised efforts towards better optimization procedures [3], generation space regularization [6], or forcing the network to account for the noise vector [9], among a rich body of literature. Lin et al. [25] 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 two 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 unconditional head motion generation task and the key components of our learning framework. Given a set of facial landmarks x_0 representing a face in an initial pose, we seek to generate a sequence $x_{1:T} = (x_1, \ldots, x_T)$ of arbitrary length T such that the probability distributions of the generated and the ground truth data, p_{G} and p_{data}, match:

$$p_{\text{G}}(x_{0:T}) = p_{\text{data}}(x_{0:T}), \quad \forall x_{0:T} \tag{1}$$

We hereafter describe our adversarial architecture to address this problem, an overview of which can be found in Figure 1. Its main components include the autoregressive generator, described in Section 3.1, and the multi-scale sequence discriminator, presented in Section 3.2. As an attempt to mitigate the potential negative impact of mode collapse, we design our architecture to learn to generate and discriminate joint probability distributions, as explained in Section 3.3. The overall loss function is presented in Section 3.4. Finally, in Section 3.5 we propose two implementations of our method to stress its generalizibility.

3.1 Autoregressive velocity generation

We implement our generator network G as an autoregressive function of past landmark positions, that at each time steps provides the instantaneous velocity:

$$x_t = x_{t-1} + G(x_{0:t-1}) \tag{2}$$

Working with velocities ensures smooth transitions between subsequent time steps but also enables simpler model architectures [26] and provides a convenient way to take advantage of the inherent potential of autoregressive models to represent cumulative sums [28]. On the other hand, autoregressive models tend to accumulate errors over time and special care must be taken in the training process to mitigate it, thus allowing for practical applications. The following sections detail the architecture of our discriminator and the learning strategy that enable long sequence generation.

3.2 Window-based multi-scale discriminator

We use a multi-scale, window-based discriminator network architecture to train the model to generate temporal patterns unfolding over different timescales. To relieve the burden of training one network per input scale, we follow Airale et al. [1] who achieved this objective using a recurrent network (RNN), considerably simplifying the

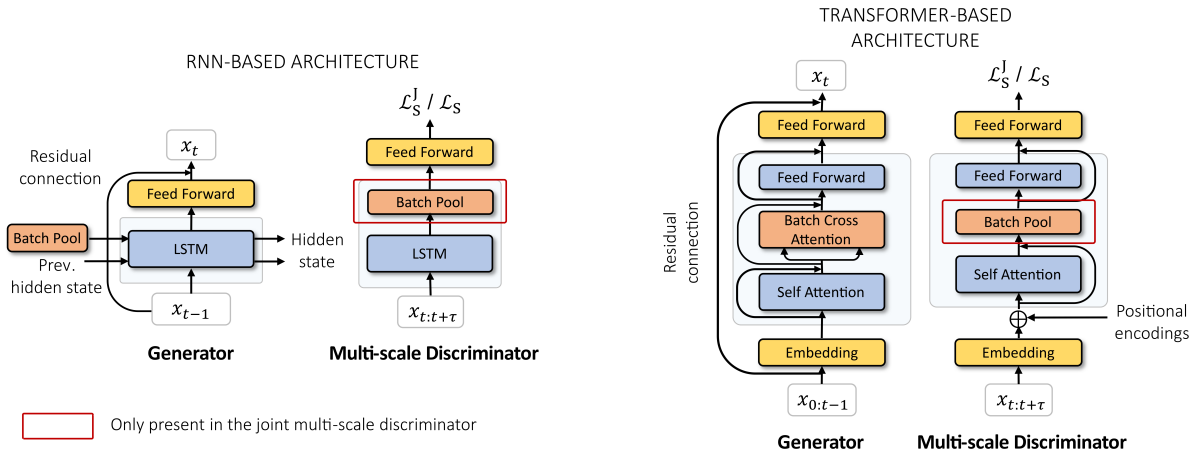

Figure 2: The two architecture variants of the proposed SUHMo method.

discriminator architecture. Here we give a more formal definition of the window-based multi-scale discriminator that is not restricted to RNN variants. First, let $D_{\text{M}} : (x_{t:t+\tau}, \theta) \in \mathbb{R}^{\tau \times d} \times \mathbb{R}^{d_\theta} \mapsto D_{\text{M}}(x_{t:t+\tau}; \theta) \in \mathbb{R}$ be a discriminator function parameterized by θ that operates on sequences of d-dimension vectors of arbitrary length τ. This definition includes RNNs, Transformers [35], and more generally any function enabling pooling in the time axis or processing time steps separately. We then define the window-based multi-scale discriminator D_{S} on sequences $x_{0:T}$ of length T as an expectation over evaluations of D_{M} on temporal slices of $x_{0:T}$:

$$D_{\text{S}}(x_{0:T}; \theta) = \mathbb{E}_{\tau,t}[D_{\text{M}}(x_{t:t+\tau}; \theta)], \quad t \geq 0, t + \tau \leq T \tag{3}$$

where τ and t are the duration, or equivalently the scale, and starting index of the window. In practice both t and τ are sampled from discrete uniform distributions. The advantage of this framework is that it gives a flexible way to adjust the scales by choosing other distributions on τ.

3.3 Learning to generate and discriminate joint probability distributions

To mitigate the mode collapse problem, we consider both the generation and discrimination of joint sample distributions. Let the objective, with generic data points x^1 and x^2, write (superscript J for joint ground truth / generated distributions):

$$\mathbb{E}_{x^1, x^2 \sim p^{\text{J}}_{\text{data}}}[\log D(x^1, x^2)] + \mathbb{E}_{x^1, x^2 \sim p^{\text{J}}_{\text{G}}}[\log(1 - D(x^1, x^2))] \tag{4}$$

This has to be minimized (resp. maximized) w.r.t. the parameters of the generator G (resp. the discriminator D). In the case of independent and identically distributed data and enough network capacity, the joint generated distribution converges to the product of the marginal data distributions [15]:

$$p^{\text{J}}_{\text{G}}(x^1, x^2) = p^{\text{J}}_{\text{data}}(x^1, x^2) = p_{\text{data}}(x^1) p_{\text{data}}(x^2) \tag{5}$$

If G produces samples independently, then p^{J}_{G} readily factorizes. This is the setting of [25], which proved useful to reduce mode collapse. However, if x and y are produced together, then G simply *learns* to factorize. Both cases lead to the equality of marginal distributions $p_{\text{G}} = p_{\text{data}}$, hence the optimization objective of Goodfellow et al. [15] is unaffected. In the real case scenario of limited network capacity, p^{J}_{G} does not factorize, and hence we argue that 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.

At test time, a single initial pose is typically provided. Since the model expects a pair of samples, one strategy consists in providing a transformed version of the reference pose as a second input. To that end we used random flip, rescaling and translation. This approach gives a practical way of injecting stochasticity in the generation process (see Section 4.3).

3.4 Training SUMHo

Following the discussion in 3.2 and 3.3, we propose to use two window-based multi-scale discriminators on the generated sequences. The joint discriminator D^{J}_{S} operates on sample pairs, while a second network, D_{S}, takes single

sequences as input and explicitly enforces the convergence of the marginal distributions p_{G} and p_{data}. Finally, to complement the sequential losses, we employ a frame discriminator D_{F} to measure the realism of each time step of the produced sequences (see Figure 1).

The loss function is the sum of the three corresponding adversarial losses (joint sequential, sequential, and frame-wise), plus a mean squared error loss $\mathcal{L}_2^{reco}$ that we scale to remain negligible after the first training epochs:

$$\mathcal{L} = \underbrace{(\mathcal{L}_{\text{S}}^{\text{J}} + \mathcal{L}_{\text{S}} + \mathcal{L}_{\text{F}})}_{\text{Adversarial loss}} + \lambda \mathcal{L}_2^{reco} \tag{6}$$

3.5 Implementation

So far the discussion has not assumed any precise functional form for either the generator or the discriminator network. Here we propose two implementations of the SUHMo method, based on LSTM and Transformers. The motivation is to highlight that the provided methodological tools can be relevant beyond a single architecture, as we further discuss in Section 4. An overview of both proposed variants can be found in Figure 2. 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 is then repeated to preserve the input batch size.

In the LSTM-based generator, the hidden state h_t goes through a batch-pool layer, yielding a pooled vector p_t that is concatenated with the next input to the LSTM. A multi-layer perceptron is used on h_t to output the landmark positions. The joint discriminator D_{S}^{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. This layer is labelled batch-cross attention. We do not use positional encoding as we observed no change in performance, while omitting it allows the generation of longer test sequences. As for the discriminator networks, a batch-pool layer replaces the batch-cross attention in D_{S}^{J} as it only needs to provide a single score per pair. A learnable class token, prepended to the input sequence, is used to give the final score, as it has been customary for Transformers [12].

4 Experiments

4.1 Experimental details

We used 1-layer LSTM with hidden size 1024 for all networks in the RNN variant of our method, and a single 1-head self attention block for the Transformer networks. In the latter architecture, embedding layers produce 1024 dimensional vectors for the generator and 128 dimensional vectors for the discriminators, i.e. the balance between G and D is mainly controlled by the embedding dimension. Models were trained on sequences of 40 time steps, and up to 5 observed frames were given as input to the LSTM to stabilize training. At inference time a single reference frame is provided, and we explore predicting sequences of two different durations, namely 40 and 80 time steps, or respectively $1.6s$ and $3.2s$.

The hinge version of the GAN loss was used in all experiments, and we set λ in equation 6 to 10^{-2}. Networks were trained with Adam optimizers with β_1 and β_2 parameters set to 0.5 and 0.999, and with generator and discriminator learning rates set to 2×10^{-5} and 1×10^{-5} respectively. Importantly, a step learning rate decay of a factor 10 was applied once performance started to stall, corresponding to roughly 60k iterations for a batch size of 120 ($\sim$ 3000 epochs for CONFER and 1000 epochs for our VoxCeleb2 subset). Training took on average two days on a single Titan RTX GPU.

We investigated concatenating velocities or instantaneous accelerations to landmark positions as input to the generator or the discriminators, expecting that it might help penalizing 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.

Experiments were conducted on two audio-visual datasets with upper-body frontal views of different speakers. **CONFER** [14] contains 72 video clips of TV debates between two persons, each about 1 minute long. We pre-processed the data preserving head translations and selected 5 clips as test data featuring persons unseen at training. Second, we trained on a randomly selected subset from **VoxCeleb2** [10], leaving 674 video clips corresponding to 10 unseen identities as test set. In both datasets the video frame rate is 25 fps.

Figure 3: Sample sequences from CONFER and VoxCeleb2 datasets and the associated *motion maps*. Samples featuring little movement produce a very sharp motion map (example c). The other samples give a good illustration of the differences in dataset preprocessing: head translation is suppressed in VoxCeleb2 sequences that only contain rotations, hence the quasi-static position of noise-tip landmarks in d, e and f. On the contrary, both translation and rotation movements are visible in the motion maps of sample a and b.

Table 1: Model comparison on the head motion generation task from a single reference frame. FVD and *t*-FID are linked to a fixed sequence length that is reported as subscript. Here all metrics are computed over the 40 last predicted time steps.

Dataset	CONFER [14]						VoxCeleb2 [10]					
Out length (frames)	40			80			40			80		
Method	FVD_{40}	FID	$t\text{-}FID_{40}$	FVD_{40}	FID	$t\text{-}FID_{40}$	FVD_{40}	FID	$t\text{-}FID_{40}$	FVD_{40}	FID	$t\text{-}FID_{40}$
HiT-DVAE [4]	368±19	6±0.4	130±7	764±35	50±2	157±12	686±37	**1±0.1**	167±4	644±27	**2±0.1**	164±6
ACTOR [29]	480±12	8±0.3	147±3	667±20	9±0.8	163±5	357±55	4±0.5	78±9	431±26	5±2	145±21
Δ-based	318±115	21±3	67±10	357±104	24±3	77±18	386±32	48±6	89±4	518±48	60±30	112±31
L_2-only	480±42	10±3	133±2	777±54	21±3	159±6	530±20	2±0.2	158±6	684±23	8±0.8	172±9
SUHMo - RNN	**162±31**	**3±0.2**	**61±8**	**147±45**	8±2	**48±11**	**76±8**	3±0.7	**21±3**	**135±33**	9±5	**31±7**
SUHMo - Trans.	175±46	4±0.7	67±12	169±33	**7±1**	52±4	134±33	3±0.8	42±10	141±31	9±3	55±16

In order to draw robust conclusions despite the inherent variability associated with GAN training, each GAN model was trained three times, such that the results reported in all tables contain both mean values and standard deviations.

4.2 Metrics

The Fréchet Inception Distance (FID) [18] and Fréchet Video Distance (FVD) [34] are used to measure the distance of the generated samples to the ground truth data distribution. While FID gives a score of static face realism, FVD measures the smoothness of the dynamics. A preliminary rasterization step is applied on landmarks to cast them in the image domain for the inceptionV3 [32] and I3D [5] networks. We also complement the FVD with a second dynamical metric based on a FID measure on *motion maps*, that we use to represent sequences on a single image. To do so, we compute an exponential moving average centered on the last time frame, thus enforcing a visual correlation between pixel intensity and time step index. The resulting metric, that is relevant in particular to discriminate sequences with little movement, is coined *t-FID* (t standing for time). Examples of data samples and their corresponding motion maps are illustrated in Figure 3.

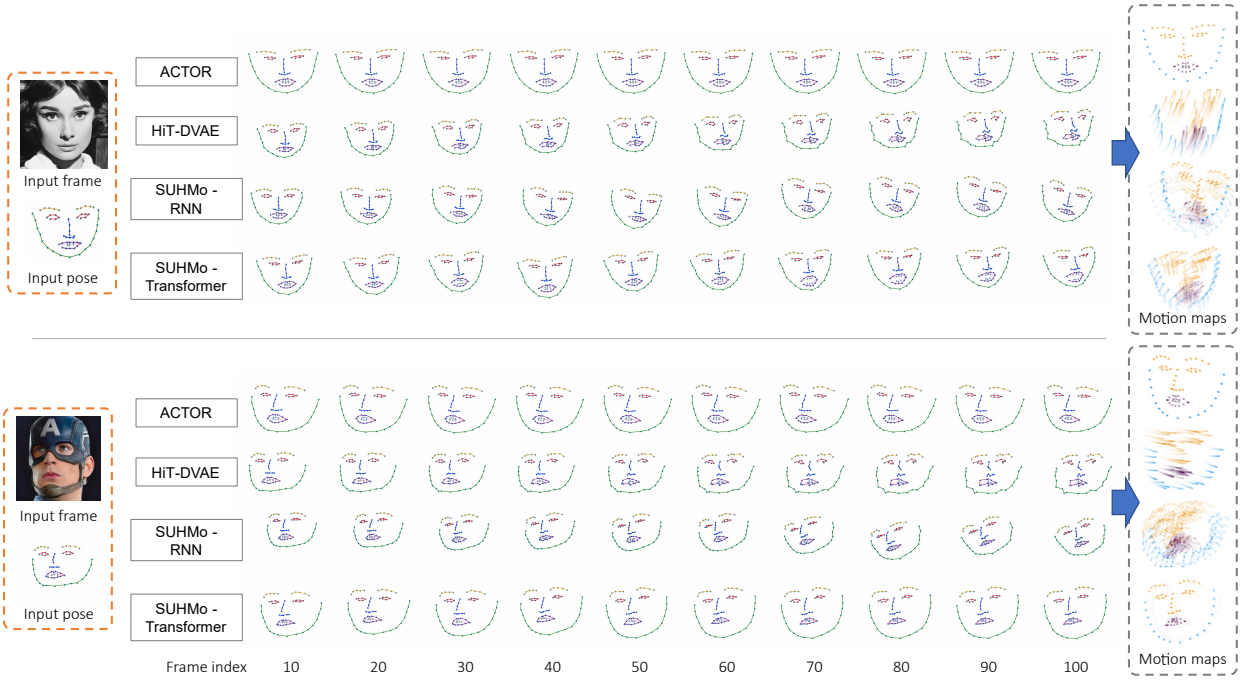

Figure 4: Qualitative evaluation of results from different models on in-the-wild images, and for sequence generation of one hundred frames. Models are trained on CONFER dataset.

4.3 Models comparison

Quantitative comparison The performances of SUHMo were compared with two state-of-the-art architectures for human pose prediction, **HiT-DVAE** [4] and **ACTOR** [29]. This task consists in predicted future positions of body joints given a short observed sequence or an action label and is therefore very close to unconditional head motion generation. One notable difference arises from the training data, which typically contains samples of a predefined set of actions and is therefore explicitly multimodal, contrary to talking head datasets. To adapt the previous models to our setting, a minimal amount of changes was therefore required. In particular, we replaced action conditioning in ACTOR by the observed initial frame. Although we could not directly work with talking head generation methods that require an audio signal for motion prediction, we took inspiration from common practices to build two additional baselines. The **Δ-based** model reproduces the SUHMo-RNN method, but similarly to Zhou et al. [49] and Das et al. [11] 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. Note that this corresponds to one time and twice the training sequence duration. 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 (see also next paragraph and Figure 4). 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. Interestingly, SUHMo exhibits very little drift as time stretches and dynamics metrics remain very low, contrary for instance to HiT-DVAE. We note however that this is an extreme setting for the use of HiT-DVAE in terms of generation over observed length ratio which is typically of the order of 3 to 5 in Bie et al. [4], whereas here it exceeds 40.

Qualitative evaluation An illustration of the results of different models on two in-the-wild images is represented in Figure 4, along with the associated motion maps. It is clear from the observation of motion maps that ACTOR produces very little movement. HiT-DVAE sequences are likewise almost static, and start drifting after 40 time steps. SUHMo sequences remain sharp aften 100 time steps, suggesting a very limited error accumulation. These results suggest that despite many similarities in the addressed problems, current human pose prediction models cannot be readily trained on head motion data without suffering a degradation in performance.

Table 2: FVD scores over different sub-sequence lengths, with and without multi-scale window-based discriminator (CONFER). Subscripts indicate the length associated with the metric.

Method	FVD_{10}	FVD_{20}	FVD_{40}
SUHMo-RNN	**28±8**	**35±4**	162±31
w/o multi-scale discriminator	35±8	42±8	**157±21**
SUHMo-Transformer	**34±9**	**40±12**	**175±46**
w/o multi-scale discriminator	57±6	60±12	236±58

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

SUHMo variant	RNN			Transformer		
Ablation	FVD_{40}	FID	t-FID_{40}	FVD_{40}	FID	t-FID_{40}
Full	**162±31**	**3±0.2**	**61±8**	175±46	**4±0.7**	67±12
One-sample D	226±76	3±1	71±16	**162±36**	7±1	**65±11**
One-sample G	222±23	5±0.7	58±7	237±58	8±2	74±10

An interesting feature of SUHMo is that the joint generation allows to produce diverse outputs given the same reference pose. We illustrate this in Figure 5. This is important for many applications that require the ability to generate different outcomes. These results also show that our training strategy is effective to prevent mode collapse.

4.4 Ablation study

Multi-scale discriminator To assess the ability of SUHMo to produce realistic patterns over diverse time scales we measure the FVD on motion chunks of 10, 20, and 40 frames, and compare it with a model trained without the window-based multi-scale discriminator (Table 2). Both models were trained to generate sequences of 40 frames and therefore perform on par on this duration. The benefit of the window-based multi-scale approach however clearly appears on shorter timescales, indicating a finer modeling of high frequency patterns.

Joint generation and discrimination We tried removing the pair mixing in the generator and the discriminator at turns (Table 3). Models trained with a standard marginal discriminator ("One-sample D") fall behind in terms of FVD and FID, respectively for the RNN and the Transformer model. Surprisingly, suppressing the joint generation ("One-sample G") has an even more detrimental effect, visible on FVD and FID for both models. In addition to its previously known benefits in mode collapse reduction, we observe that working with pairs of samples also helps improving the overall quality of the generated motion sequences in the unconditional generation setting.

5 Conclusion

In this paper we presented SUHMo, an unconditional head motion generation method able to animate a human face over long sequences from a single initial frame in a semantic space. Our method is based on the autoregressive generation of incremental displacements, or instantaneous velocities, of pairs of samples, and it is trained using a window-based multi-scale discriminator. We showed that our methodological contributions can accommodate several implementations, consistently outperforming state-of-the-art human pose generation methods and head motion prediction baselines in terms of dynamics quality and pose realism. In a future work we plan to extend our method and notably assess if it can improve the fidelity of head motion in an audio-conditioned talking head generation setting, which remains an open problem.

Acknowledgements

This work was supported by the IDEX project MIDGEN from Univ. Grenoble Alpes, and by the European Commission with the project H2020 SPRING (Socially Pertinent Robots in Gerontological Healthcare, GA #871245).

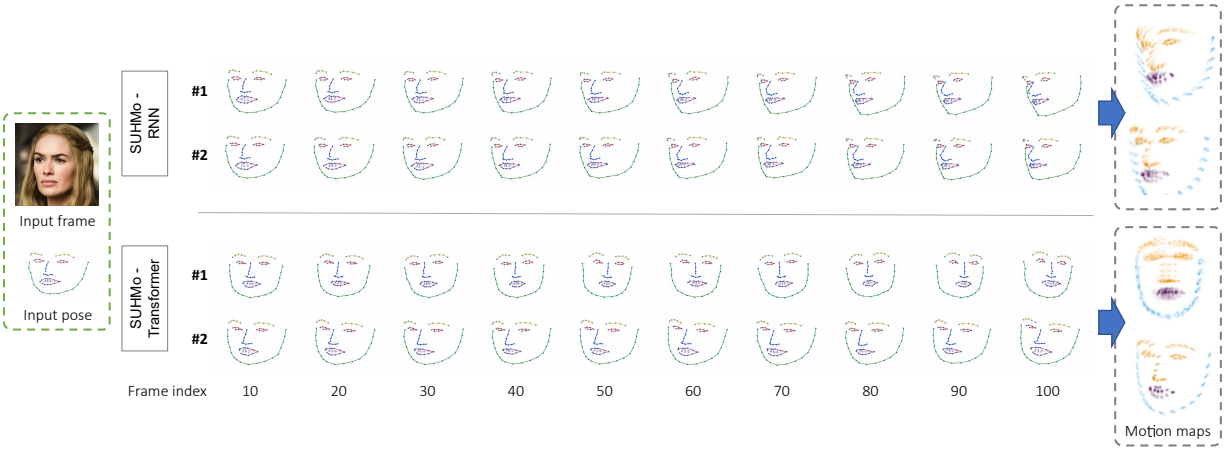

Figure 5: Illustrative examples of diverse results given the same reference pose, for both variants of SUHMo method. Models are trained on VoxCeleb2 dataset.

References

[1] L. Airale, D. Vaufreydaz, and X. Alameda-Pineda. Socialinteractiongan: Multi-person interaction sequence generation. *IEEE Transactions on Affective Computing*, 2022.

[2] S. Aliakbarian, F. Saleh, L. Petersson, S. Gould, and M. Salzmann. Contextually plausible and diverse 3d human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11333–11342, 2021.

[3] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017.

[4] 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.

[5] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

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

[7] L. Chen, R. K. Maddox, Z. Duan, and C. Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7832–7841, 2019.

[8] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, and C. Xu. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision*, pages 35–51. Springer, 2020.

[9] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

[10] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

[11] D. Das, S. Biswas, S. Sinha, and B. Bhowmick. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *European conference on computer vision*, pages 408–424. Springer, 2020.

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

[13] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18780, 2022.

[14] C. Georgakis, Y. Panagakis, S. Zafeiriou, and M. Pantic. The conflict escalation resolution (confer) database. *Image and Vision Computing*, 65:37–48, 2017.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014.

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

[17] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2255–2264, 2018.

[18] 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. *Advances in Neural Information Processing Systems (NeurIPS)*, 30: 6626–6637, 2017.

[19] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[20] X. Ji, H. Zhou, K. Wang, Q. Wu, W. Wu, F. Xu, and X. Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. *arXiv preprint arXiv:2205.15278*, 2022.

[21] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017.

[22] 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. *Advances in neural information processing systems*, 32, 2019.

[23] J. N. Kundu, M. Gor, and R. V. Babu. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8553–8560, 2019.

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

[25] Z. Lin, A. Khetan, G. Fanti, and S. Oh. Pacgan: The power of two samples in generative adversarial networks. *Advances in neural information processing systems*, 31, 2018.

[26] J. Martinez, M. J. Black, and J. Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900, 2017.

[27] M. Meshry, S. Suri, L. S. Davis, and A. Shrivastava. Learned spatial representations for few-shot talking-head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13829–13838, 2021.

[28] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, and Y. Bengio. Chunked autoregressive GAN for conditional waveform synthesis. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=v3aeIsY_vVX`.

[29] M. Petrovich, M. J. Black, and G. Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021.

[30] M. Stypułkowski, K. Vougioukas, S. He, M. Zięba, S. Petridis, and M. Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. *arXiv preprint arXiv:2301.03396*, 2023.

[31] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[33] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4):1–11, 2017.

[34] 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.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[36] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee. Learning to generate long-term future via hierarchical prediction. In *international conference on machine learning*, pages 3560–3569. PMLR, 2017.

[37] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 613–621, 2016.

[38] K. Vougioukas, S. Petridis, and M. Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128(5):1398–1413, 2020.

[39] S. Wang, L. Li, Y. Ding, C. Fan, and X. Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. In *IJCAI*, 2021.

[40] S. Wang, L. Li, Y. Ding, and X. Yu. One-shot talking face generation from single-speaker audio-visual correlation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2531–2539, 2022.

[41] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

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

[43] L. Yu, J. Yu, M. Li, and Q. Ling. Multimodal inputs driven talking face generation with spatial–temporal dependency. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(1):203–216, 2020.

[44] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9459–9468, 2019.

[45] E. Zakharov, A. Ivakhnenko, A. Shysheya, and V. Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference on Computer Vision*, pages 524–540. Springer, 2020.

[46] R. Zhao, T. Wu, and G. Guo. Sparse to dense motion transfer for face image animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1991–2000, 2021.

[47] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9299–9306, 2019.

[48] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021.

[49] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020.