Autoregressive GAN for Semantic Unconditional Head Motion Generation

LOUIS AIRALE, Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, France
XAVIER ALAMEDA-PINEDA, Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, France
STÉPHANE LATHUILIÈRE, LTCI, Télécom Paris, Institut polytechnique de Paris, France
DOMINIQUE VAUFREYDAZ, Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, France

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.

CCS Concepts: • **Computing methodologies** → **Reconstruction**.

Additional Key Words and Phrases: GAN, Head motion, Face landmarks

1 INTRODUCTION

This work deals with learning the dynamics of a human face from a single initial pose. It relates to the task of talking head generation, where the head and lip motion generation is conditioned on an audio clip, plus possible additional signals such as emotional state or exemplar pose dynamics [21, 34, 36, 47, 53]. 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 ensuring long-term dynamics quality, which is known to be particularly challenging in the absence of a conditioning signal [42]. Another advantage of tackling this problem is that the focus put on the fine handling of head dynamics may benefit audio-conditioned talking head synthesis, where generating natural head motions has, until recently, consistently received less attention. Following the common practice in talking head synthesis, we produce the dynamics in a low-dimensional subspace [37, 41]. This is usually done to alleviate the difficulty of handling both facial dynamics and photorealism directly in the image space. These low-dimensional representations comprise supervised facial landmarks [8, 31, 54], 3D mesh [13] or unsupervised keypoints [44, 45], and following the designation of *high level semantics* used in Villegas et al. [41], we refer to this space as the *semantic space*.

Authors' addresses: Louis Airale, louis.airale@univ-grenoble-alpes.fr, Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France, 38000; Xavier Alameda-Pineda, xavier.alameda-pineda@inria.fr, Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, France, 38000; Stéphane Lathuilière, stephane.lathuiliere@telecom-paris.fr, LTCI, Télécom Paris, Institut polytechnique de Paris, France; Dominique Vaufreydaz, Dominique.Vaufreydaz@univ-grenoble-alpes.fr, Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France, 38000.

ACM 1551-6857/2023/12-ART
https://doi.org/10.1145/3635154

In this work, we address the task of unconditional head motion sequence generation, i.e. synthesizing head pose and facial expression from a single reference frame, in the 2D facial landmarks semantic space. Several models were notably proposed to map landmarks to real-world images, making this representation relevant in practice [29, 49–51].

Continuous sequences such as facial landmark coordinates x_t can be conveniently represented as a cumulative sum of incremental displacements v_t, or instantaneous velocities, starting from the observed initial position x_0:

$$x_t = x_{t-1} + v_t = x_0 + \sum_{\tau \leq t} v_\tau. \tag{1}$$

This approach has been followed successfully, for instance, by Lin and Amer [26] or Kundu et al. [24] for human pose generation and by Gupta et al. [17] for trajectory prediction. As shown by Martinez et al. [28], this formulation allows the use of shallower neural network architectures. Another feature of such cumulative sums is that they can be properly described by autoregressive models [30]. In a most general definition, an autoregressive function G produces coordinates x_t one by one given the input positions x_0 and all previously generated positions:

$$x_t = G(x_0, x_{1:t-1}) \tag{2}$$

In practice, conditional independence property assumptions can be made to reduce the necessity to model all previous time steps and allow for the use of a large diversity of network architectures on a fixed history length. Although they can produce sequences of arbitrary length, autoregressive models may however accumulate error, or alternatively end up generating average values over time when trained with a mean squared error loss [28]. This 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 for head motion prediction, be it speech-conditioned or not. To that end, we extend the window-based multi-scale discriminator network we introduced in Airale et al. [1] for discrete token generation to continuous head dynamics prediction, which are intricate data composed of temporal patterns that evolve over varied timescales. Previous works have addressed the generation of such data with discriminator networks operating on receptive fields of different sizes [23, 46] or on local windows, enabling a better representation of high-frequency components [19]. On the contrary, our discriminator implements a multi-scale window-based architecture in a single network, which allows it to operate at any temporal resolution. Last, in the light of Lin et al. [27], we provide pairs of samples to the discriminator network as a mode collapse mitigation technique, but also produce sample pairs in the generator. As we show, this approach does not change the optimization objective but 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 that even extend 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,
- An experimental validation of the effectiveness of the proposed SUHMo method on this novel task, where we compare to prominent models on the closely related task of human pose prediction and other strong baselines on two benchmark datasets.

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

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 [8, 33, 43, 45, 52, 54]. 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 [20, 48, 53]. 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 [7, 47, 54], although the range of achieved motions remains limited. Recently, Wang et al. [44] 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. [45]. 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 [23], human trajectory prediction in a crowd [17], or human motion prediction [2, 24, 26, 28]. 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 by 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. [46], 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 [23, 26]. 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 (see Figure 1 for a schematic representation of these discriminators). Our window-based multi-scale discriminator extends that of our preceding work [1] to more diverse architectures and to continuous input. This formulation has the 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. [27] 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

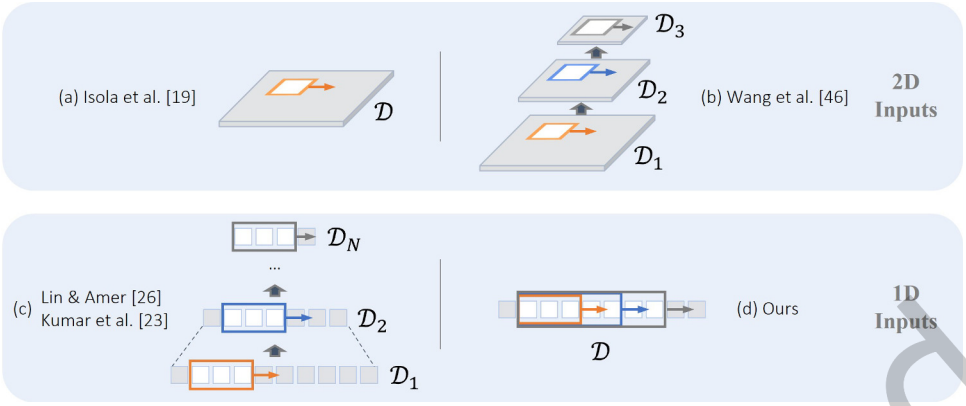

Fig. 1. One and two-dimension multi-scale & window-based discriminator architectures. (a) The purely window-based PatchGAN discriminator [19]. (b) Extension to a 3-scale architecture in Wang et al. [46]. (c) The 1D multi-scale PatchGAN structure used in DVGANs [26] and MelGAN [23] discriminators. (d) The proposed multi-scale window-based discriminator has a unique set of parameters and takes sequences of any size as input, giving a free hand to select the scales.

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:

$$p_{\mathrm{G}}(x_{0:T}) = p_{\mathrm{data}}(x_{0:T}), \quad \forall x_{0:T} \tag{3}$$

We hereafter describe our adversarial architecture to address this problem, an overview of which can be found in Figure 2. 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:

$$x_t = x_{t-1} + G(x_{0:t-1}) \tag{4}$$

Working with velocities ensures smooth transitions between subsequent time steps but also enables simpler model architectures [28] and provides a convenient way to take advantage of the inherent potential of autoregressive models to represent cumulative sums [30]. 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 build on a model we previously introduced [1] which considerably simplifies the discriminator architecture. Here we give a more formal definition of the window-based multi-scale discriminator that is not restricted to RNN

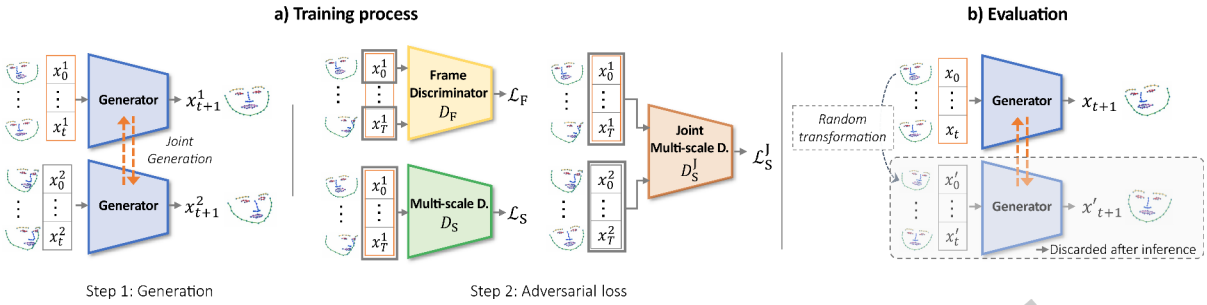

Fig. 2. Overview of SUHMo training process. The autoregressive generator produces pairs of outputs, that are evaluated by three discriminator networks. At test time, the second sample is replaced by a transformed version of the reference pose.

variants and discrete inut data. First, let $D_{\mathrm{M}} : (x_{t:t+\tau}, \theta) \in \mathbb{R}^{\tau \times d} \times \mathbb{R}^{d_\theta} \mapsto D_{\mathrm{M}}(x_{t:t+\tau}; \theta) \in \mathbb{R}$ be a discriminator function parameterized by θ that operates on sequences of d-dimension vectors of arbitrary length τ. This definition includes RNNs, Transformers [40], 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 $x_{0:T}$:

$$D_{\mathrm{S}}(x_{0:T}; \theta) = \mathbb{E}_{\tau,t}[D_{\mathrm{M}}(x_{t:t+\tau}; \theta)], \quad t \geq 0, t + \tau \leq T \tag{5}$$

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 various distributions on τ. Our discriminator is represented in Figure 1, along with previous multi-scale and window-based architectures.

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):

$$\mathbb{E}_{x^1,x^2 \sim p^{\mathrm{J}}_{\mathrm{data}}}[\log D(x^1, x^2)] + \mathbb{E}_{x^1,x^2 \sim p^{\mathrm{J}}_{\mathrm{G}}}[\log(1 - D(x^1, x^2))] \tag{6}$$

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]:

$$p^{\mathrm{J}}_{\mathrm{G}}(x^1, x^2) = p^{\mathrm{J}}_{\mathrm{data}}(x^1, x^2) = p_{\mathrm{data}}(x^1) p_{\mathrm{data}}(x^2) \tag{7}$$

If G produces samples independently, then $p^{\mathrm{J}}_{\mathrm{G}}$ readily factorizes. This is the setting of [27], 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 $p_{\mathrm{G}} = p_{\mathrm{data}}$, hence the optimization objective of Goodfellow et al. [15] is unaffected. In the real case scenario of limited network capacity, $p^{\mathrm{J}}_{\mathrm{G}}$ 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

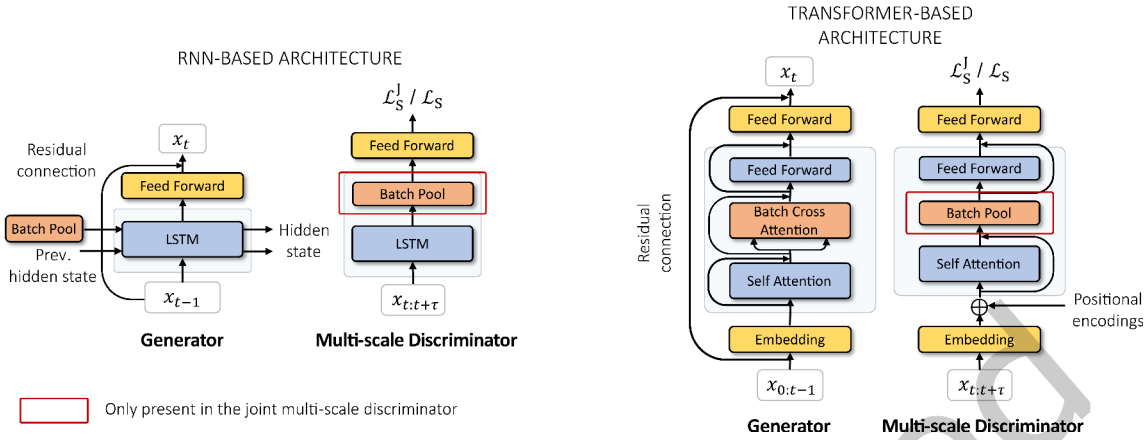

Fig. 3. The two architecture variants of the proposed SUHMo method.

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 $D_{\mathrm{S}}^{\mathrm{J}}$ 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 2). The generator adversarial losses writes:

$$\mathcal{L}_{\mathrm{S}}^{\mathrm{J}} = -\mathbb{E}_{x_{0:T}^1 \sim p_{\mathrm{G}}, x_{0:T}^2 \sim p_{\mathrm{G}}} \left[D_{\mathrm{S}}^{\mathrm{J}}(x_{0:T}^1, x_{0:T}^2) \right], \tag{8}$$

$$\mathcal{L}_{\mathrm{S}} = -\mathbb{E}_{x_{0:T} \sim p_{\mathrm{G}}} \left[D_{\mathrm{S}}(x_{0:T}) \right], \tag{9}$$

$$\mathcal{L}_{\mathrm{F}} = -\mathbb{E}_{x_{0:T} \sim p_{\mathrm{G}}} \left[\frac{1}{T} \sum_{t \geq 1} D_{\mathrm{F}}(x_t) \right]. \tag{10}$$

The overall loss function is the sum of these three losses plus a mean squared error term $\mathcal{L}_2^{reco}$ that we scale to remain negligible after the first training epochs:

$$\mathcal{L} = \underbrace{(\mathcal{L}_{\mathrm{S}}^{\mathrm{J}} + \mathcal{L}_{\mathrm{S}} + \mathcal{L}_{\mathrm{F}})}_{\text{Adversarial loss}} + \lambda \mathcal{L}_2^{reco} \tag{11}$$

We employ the geometric GAN formulation of Lim and Ye [25] for the discriminator loss functions:

$$\mathcal{L}_{D_*} = \mathbb{E}_{x \sim p_{\mathrm{G}}} \left[\max(0, 1 + D_*(x)) \right] + \mathbb{E}_{x \sim p_{\mathrm{data}}} \left[\max(0, 1 - D_*(x)) \right], \tag{12}$$

where D_* is replaced respectively by D_{F}, D_{S} and $D_{\mathrm{S}}^{\mathrm{J}}$ and sequences are sampled as in equations 8 to 10.

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 3. To account for pairs of inputs, we define a batch-pool (BP) 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:

$$x^o = \mathrm{BP}(x^i), \tag{13}$$

$$x^o_{2n-1,c,d} = x^o_{2n,c,d} = \max(x^i_{2n-1,c,d}, x^i_{2n,c,d}), \tag{14}$$

where the subscripts represent the batch, channel and dimension indices. In the LSTM-based generator, the hidden state h_t goes through a BP 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 $D^{\mathrm{J}}_{\mathrm{s}}$ is composed of a LSTM, a BP layer and a feed forward network; the marginal discriminator D_{s} is similar but without the BP layer (see Figure 3, left).

In the Transformer generator (Figure 3, right), pair mixing is done in a multi-head attention (MHA) layer by inverting the batch indices of paired samples in the key and value vectors. This way, each sample in a pair can attend to the history of the other sample. This layer is labelled batch-cross attention (BXA):

$$\mathrm{BXA}(q, k, v) = \mathrm{MHA}(q, k^r, v^r), \tag{15}$$

$$k^r_{2n} = k_{2n-1}, \quad k^r_{2n-1} = k_{2n} \quad \forall \quad 1 \le n \le N/2, \tag{16}$$

$$v^r_{2n} = v_{2n-1}, \quad v^r_{2n-1} = v_{2n} \quad \forall \quad 1 \le n \le N/2, \tag{17}$$

with N the batch size, and $q = k = v$ in all experiments, i.e. query, key and value tokens originate from the same sequences. 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 $D^{\mathrm{J}}_{\mathrm{s}}$ 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

All networks in the RNN variant of our method are implemented as 1-layer LSTM with hidden size 1024, while Transformer networks are implemented as a single self-attention block with one head. 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.

We set λ in equation 11 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

Fig. 4. 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.

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.

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) [38] are used to measure the distance of the generated samples to the ground truth data distribution. While the FID gives a score of static face realism, the 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 [35] 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 4.

4.3 Models comparison

Quantitative comparison. As it is the first unconditional head motion prediction method, we must rely on a broader literature to assess the performances of SUHMo. Human pose prediction, which consists in predicting future positions of body joints given a short observed sequence or an action label, is arguably the closest task to

Table 1. Model comparison on the head motion generation task from a single reference frame on CONFER and VoxCeleb2. The FVD and *t*-FID are sequential metrics computed on fixed sequence lengths reported as subscript. Here all metrics are computed over the 40 last predicted time steps.

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

the one we address here. We therefore compare to two state-of-the-art architectures for human pose prediction, **HiT-DVAE** [4] and **ACTOR** [32], which both implement Transformer-based Variational Autoencoder (VAE) architectures [22], and that we train on our talking head datasets. One notable difference arises from the fact that human pose prediction datasets are usually composed of several modes corresponding to a predefined set of actions, and synthesis models typically account for this by conditioning the generation on an action label. A minimal amount of changes is therefore necessary to adapt the previous models to our setting: we replace in particular the action conditioning in ACTOR by the observed initial frame. We also seek to compare with audio-conditioned talking head generation models. Although it is not possible to evaluate them directly in the absence of a speech signal, we take inspiration from common practices in talking head generation to build two additional baselines. The **Δ-based** model reproduces the SUHMo-RNN method, but similarly to Zhou et al. [54] and Das et al. [11] produces displacements from a fixed set of reference points, in this case the initial landmark positions. **MLE**, for maximum likelihood estimation, 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 5). The same is true for models trained with a L_2

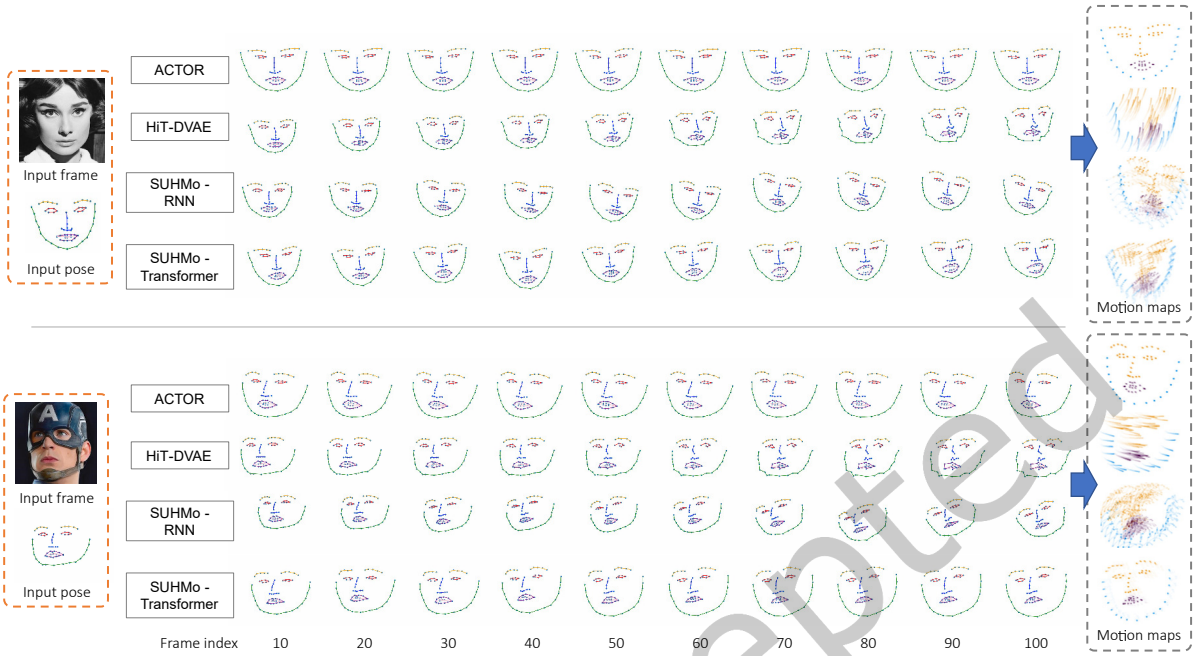

Fig. 5. Qualitative evaluation of results from different models on in-the-wild images, and for sequence generation of one hundred frames. Models are trained on CONFER dataset.

reconstruction loss (the likelihood-based method), 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 5, 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 after 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.

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 6. 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

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

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 the 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 improve the overall quality of the generated motion sequences in the unconditional generation setting.

5 LIMITATIONS

Although several landmarks-to-image methods exist, most are still either unfit for large pose changes or require an additional fine-tuning step on unseen identities, which limits their usage on the produced motion sequences. However our method is not specific to 2D landmarks, which if needed may be replaced at almost no cost by any other representation from which to reenact output videos. A second limitation stems from the autoregressive nature of our model. Although this enables smooth and realistic dynamics, as the output sequence length grows error inevitably accumulates, distorting the face and limiting in practice the maximum length. Possible solutions include block-wise autoregression [30] or discretizing the action space [39], although the former does not totally avoid error accumulation while finding the optimal codebook dimension to model the complex facial configurations is an open problem in the latter case.

6 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

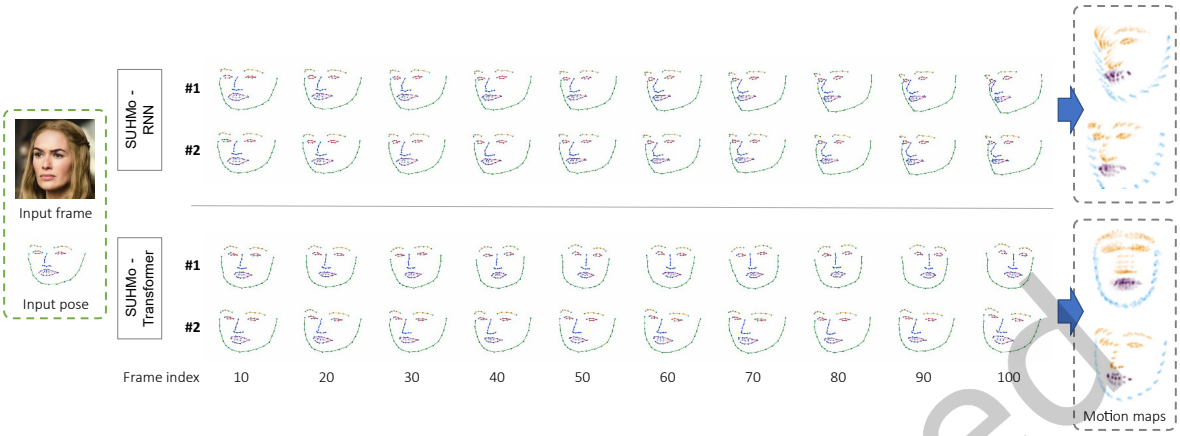

Fig. 6. Illustrative examples of diverse results given the same reference pose, for both variants of SUHMo method. Models are trained on VoxCeleb2 dataset.

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).

REFERENCES

[1] Louis Airale, Dominique Vaufreydaz, and Xavier Alameda-Pineda. 2022. Socialinteractiongan: Multi-person interaction sequence generation. *IEEE Transactions on Affective Computing* (2022).

[2] Sadegh Aliakbarian, Fatemeh Saleh, Lars Petersson, Stephen Gould, and Mathieu Salzmann. 2021. Contextually plausible and diverse 3d human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11333–11342.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 214–223.

[4] Xiaoyu Bie, Wen Guo, Simon Leglaive, Lauren Girin, Francesc Moreno-Noguer, and Xavier Alameda-Pineda. 2022. HiT-DVAE: Human Motion Generation via Hierarchical Transformer Dynamical VAE. *arXiv preprint arXiv:2204.01565* (2022).

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

[6] Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. 2016. Mode regularized generative adversarial networks. *arXiv preprint arXiv:1612.02136* (2016).

[7] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. 2020. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision*. Springer, 35–51.

[8] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7832–7841.

[9] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. 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. 2018. VoxCeleb2: Deep Speaker Recognition. In *INTERSPEECH*.
[11] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. 2020. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *European conference on computer vision*. Springer, 408–424.
[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.
[13] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. 2022. FaceFormer: Speech-Driven 3D Facial Animation with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18770–18780.
[14] Christos Georgakis, Yannis Panagakis, Stefanos Zafeiriou, and Maja Pantic. 2017. The conflict escalation resolution (confer) database. *Image and Vision Computing* 65 (2017), 37–48.
[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2672–2680.
[16] David Greenwood, Iain Matthews, and Stephen Laycock. 2018. Joint learning of facial expression and head pose from speech. Interspeech.
[17] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2255–2264.
[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017), 6626–6637.
[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.
[20] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. 2022. EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model. *arXiv preprint arXiv:2205.15278* (2022).
[21] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–12.
[22] Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. *2nd International Conference on Learning Representations (ICLR)* (2014).
[23] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems* 32 (2019).
[24] Jogendra Nath Kundu, Maharshi Gor, and R Venkatesh Babu. 2019. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 8553–8560.
[25] Jae Hyun Lim and Jong Chul Ye. 2017. Geometric gan. *arXiv preprint arXiv:1705.02894* (2017).
[26] Xiao Lin and Mohamed R Amer. 2018. Human motion modeling using dvgans. *arXiv preprint arXiv:1804.10652* (2018).
[27] Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. 2018. Pacgan: The power of two samples in generative adversarial networks. *Advances in neural information processing systems* 31 (2018).
[28] Julieta Martinez, Michael J Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2891–2900.
[29] Moustafa Meshry, Saksham Suri, Larry S Davis, and Abhinav Shrivastava. 2021. Learned spatial representations for few-shot talking-head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13829–13838.
[30] Max Morrison, Rithesh Kumar, Kundan Kumar, Prem Seetharaman, Aaron Courville, and Yoshua Bengio. 2022. Chunked Autoregressive GAN for Conditional Waveform Synthesis. In *International Conference on Learning Representations*. https://openreview.net/forum?id=v3aeIsY_vVX
[31] Naima Otberdout, Mohamed Daoudi, Anis Kacem, Lahoucine Ballihi, and Stefano Berretti. 2020. Dynamic facial expression generation on hilbert hypersphere with conditional wasserstein generative adversarial nets. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 2 (2020), 848–863.
[32] Mathis Petrovich, Michael J Black, and Gül Varol. 2021. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10985–10995.
[33] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zięba, Stavros Petridis, and Maja Pantic. 2023. Diffused Heads: Diffusion Models Beat GANs on Talking-Face Generation. *arXiv preprint arXiv:2301.03396* (2023).
[34] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–13.
[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.
[36] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–11.
[37] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. 2018. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1526–1535.

[38] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717* (2018).
[39] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).
[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[41] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. 2017. Learning to generate long-term future via hierarchical prediction. In *international conference on machine learning*. PMLR, 3560–3569.
[42] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Generating videos with scene dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*. 613–621.
[43] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2020. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision* 128, 5 (2020), 1398–1413.
[44] Suzhe Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. 2021. Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion. In *IJCAI*.
[45] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. 2022. One-shot talking face generation from single-speaker audio-visual correlation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 2531–2539.
[46] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8798–8807.
[47] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. 2020. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137* (2020).
[48] Lingyun Yu, Jun Yu, Mengyan Li, and Qiang Ling. 2020. Multimodal inputs driven talking face generation with spatial–temporal dependency. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 1 (2020), 203–216.
[49] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. 2020. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference on Computer Vision*. Springer, 524–540.
[50] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 9459–9468.
[51] Ruiqi Zhao, Tianyi Wu, and Guodong Guo. 2021. Sparse to dense motion transfer for face image animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1991–2000.
[52] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 9299–9306.
[53] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4176–4186.
[54] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. 2020. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–15.