

30 Nov 2021

arXiv:2107.07224v2 [cs.CV]

StyleVideoGAN: A Temporal Generative Model using a Pretrained StyleGAN

Gereon Fox

gfox@mpi-inf.mpg.de

Ayush Tewari

atewari@mpi-inf.mpg.de

Mohamed Elgharib

elgharib@mpi-inf.mpg.de

Christian Theobalt

theobalt@mpi-inf.mpg.de

Max Planck Institute for Informatics

Saarland Informatics Campus

Saarbrücken, Germany

Abstract

Generative adversarial models (GANs) continue to produce advances in terms of the visual quality of *still* images, as well as the learning of temporal correlations. However, few works manage to combine these two interesting capabilities for the synthesis of video content: Most methods require an extensive training dataset to learn temporal correlations , while being rather limited in the resolution and visual quality of their output. We present a novel approach to the video synthesis problem that helps to greatly improve visual quality and drastically reduce the amount of training data and resources necessary for generating videos. Our formulation separates the *spatial* domain, in which individual frames are synthesized, from the *temporal* domain, in which motion is generated. For the spatial domain we use a pre-trained StyleGAN network, the latent space of which allows control over the appearance of the objects it was trained for. The expressive power of this model allows us to embed our training videos in the StyleGAN latent space. Our temporal architecture is then trained not on sequences of RGB frames, but on sequences of StyleGAN latent codes. The advantageous properties of the StyleGAN space simplify the discovery of temporal correlations. We demonstrate that it suffices to train our temporal architecture on only 10 minutes of footage of 1 subject for about 6 hours. After training, our model can not only generate new portrait videos for the training subject, but also for *any* random subject which can be embedded in the StyleGAN space.

1 Introduction

Recent advances of generative adversarial networks (GANs), notably StyleGAN [13, 14, 15], show impressive capabilities in learning manifolds of photorealistic images at high resolution (up to 1024^2). This is especially true for images of human faces. However, these improvements are only starting to carry over to the domain of *videos*: While existing methods for videos show promising results in modeling content and motion [19, 24, 27, 28, 33], they usually are subject to at least a subset of the following limitations: small spatial resolution (

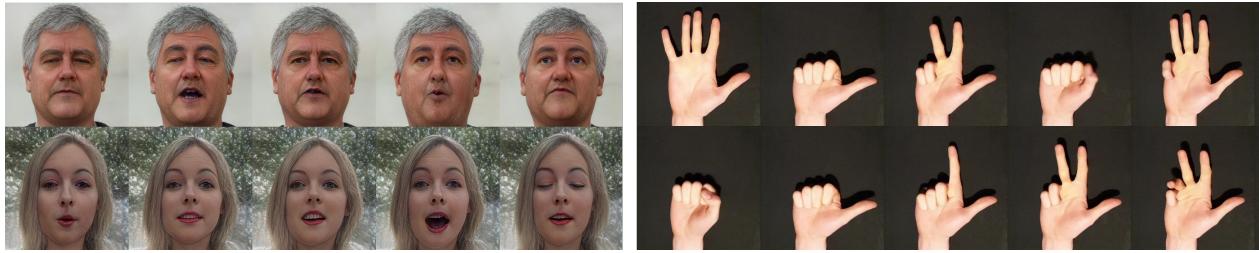


Figure 1: Face and hand videos generated by our method

$\leq 128^2$); spatial artifacts ; constrained motion ; necessity of large amounts of training data ; large computational cost for training (memory, time; see Table 2 and Section 4) .

To address these problems, we present a novel approach to unconditional video generation: Our goal is to learn a generator for nearly photorealistic, high-resolution videos (up to 1024×1024), by training on a video dataset that contains no more than 10 minutes of video footage. In addition, we limit the training footage to depict only a single subject. However, we want the trained model can generate motion not only for the training subject, but for *many* different *random* subjects. As the proving ground for our method we chose the generation of portrait videos, because (1) portraits are an attractive target for animation, and (2) high-quality training data and StyleGAN models for this domain are readily available. To demonstrate that our method is also applicable to other domains very different from portrait, we also show how it can be applied to the generation of complex hand motion.

Our key idea for addressing this task is to embed our training video set into the \mathcal{W}^+ latent space of a pretrained StyleGAN model [22] . Embedding videos in \mathcal{W}^+ turns them from sequences of RGB frames into sequences of \mathcal{W}^+ vectors. While an RGB frame has $1024 \cdot 1024 \cdot 3 = 3145728$ dimensions, a \mathcal{W}^+ code only has $18 \cdot 512 = 9216$ dimensions. This means that the embedding allows our generative model to be trained in a much lower-dimensional space, which simplifies the discovery of temporal correlations. Also, this transformation completely eliminates the necessity to actually render any video frames at training time, which greatly reduces the amount of memory and the time required to train our model. Our model is supervised completely in \mathcal{W}^+ space, unlike any other existing approach.

There is one more major advantage of the embedding approach, that allows us to make our model generate motion for a great multitude of subjects, even though it has seen only one subject at training time: The linear separability properties of the \mathcal{W}^+ space [20, 25, 26] allow us to analyze the shape of the \mathcal{W}^+ embedding of the training footage. Using such an analysis, we present an “**offset trick**” which allows us to transfer the motion of a generated video to a different subject.

The ideas described so far already allow us to generate very high resolution videos with a minimal demand of training data and computational resources, by training a recurrent Wasserstein GAN [3] on temporal volumes of 25 time steps, more than what most previous methods can afford. However, in order to have our model generate videos of longer duration, our generator needs to continue its output sequence beyond 25 time steps at test time. This can be achieved by making the generator a recurrent neural network (RNN). Previous work [27] has pointed out, though, that just using an RNN is not sufficient. We validate this observation by demonstrating that a vanilla RNN may tend to “loop”, i.e., repeat the same motion over and over. We address this problem with a novel “**gradient angle penalty**” term which successfully prevents looping. In summary, we make the following contributions:

- We present a novel approach for unconditional video generation that is supervised in the latent space of a pretrained image generator, without having to render video frames at training time, leading to large savings in computational resources.
- We are the first to demonstrate how the properties of StyleGAN’s \mathcal{W}^+ space can greatly reduce the amount of training data necessary to train a video generative model.
- We present a novel “gradient angle penalty” loss that helps in the generation of videos that are longer than the temporal window seen at training time.
- We demonstrate that our approach is applicable even to domains as complicated as hand motion, where there are more challenging articulations and self-occlusions.

2 Related Work

2.1 Generative Models for Videos

Several methods have been proposed for learning a generative model of videos [2, 5, 6, 7, 11, 12, 17, 19, 23, 24, 27, 28, 31, 32, 33, 34, 36]. While such approaches show interesting results, they are limited to low spatial resolutions such as 128x128 [19, 28, 32, 34], 256x256 [2, 6, 11, 24] or 512x512 [12]. An exception is the work of Tian *et al.* [27], which can generate videos at 1024x1024. Furthermore, most approaches struggle with generating realistic videos of long durations. We now discuss these existing approaches in more detail:

Saito *et al.* [23] presented an approach for video synthesis using Wasserstein GAN losses and a novel parameter clipping method. The network architecture, like ours, uses a shared image generator for each frame. However, the image generator is not pretrained, and the loss functions are defined on the final video space, and not the latent space of the image generator. Saito *et al.* demonstrate results on videos upto 128x128 resolution. Tulyakov *et al.* [28] decompose the generation of videos into a content part and a motion part. Their “MoCoGAN” is trained in an unsupervised manner using motion and content discriminators. Yushchenko *et al.* [36] formulated video generation by means of Markov Decision Processes and extended into the framework of Tulyakov *et al.* [28]. Acharya *et al.* [2] learned to progressively grow the generative model starting from low-resolution and short duration, which enabled the synthesis of videos at 256x256 resolution for the first time. Clark *et al.* [6] divided their discriminator into a spatial component and a temporal component, which also allows the generation of videos at resolutions up to 256x256 and duration up to 48 frames. More recently, Saito *et al.* [24] proposed a memory-efficient approach for training that scales linearly with resolution. Instead of directly training on the full temporal window, it uses a stack of sub-generators that are trained on different temporal resolutions. Earlier sub-generators process high frame-rate input with low resolution information while the later sub-generators process low frame-rate input with high resolution information.

Weissenborn *et al.* [33] proposed an autoregressive video generation model that generalizes the Transformer architecture [30] using a three-dimensional self-attention mechanism. To reduce computational complexity, images are produced as sequences of smaller, sub-scaled image slices, akin to [18]. Munoz *et al.* [19] modelled frames as points in a latent space, without any 3D convolutions. Their generator consists of a sequence generator and an image generator. Adversarial losses contain a 2D discriminator and a 3D discriminator.

Very recently, and most related to our work, Tian *et al.* [27] formulated video generation as the problem of finding a suitable trajectory through the latent space of a pretrained and

fixed image generator [4, 15], for example StyleGAN. Despite this commonality and their ability to produce output at resolution 1024×1024 as a result, there is a number of important differences between their approach and ours:

- Their discriminator supervises the generator in the image domain, which is a much higher-dimensional and more redundant domain than \mathcal{W}^+ .
- Their design inherently relies on the image generator being available for forward and backward passes at training time, which increases the required amounts of GPU memory and computation time immensely, compared to our approach.
- Their method requires a diverse training set. When trained on a single video (like our method is), their results show very limited motion, as we demonstrate in Section 4

3 Method

We train a Wasserstein GAN [3], consisting of generator G and critic C .

The input to G is a pair (i, s) , with both $i \sim \mathcal{N}(0, 1)^{32}$ and $s \sim \mathcal{N}(0, 1)^{32 \times (t-1)}$ being Gaussian samples. The number t of time steps is fixed to 25 at training time, but can be larger at test time, because the generator is an RNN. The output of G is a sequence of t latent codes $w_k \in \mathcal{W}^+$, with $0 \leq k < t$. To train the generator, we first use pSp [22], an encoder-based inversion method for StyleGAN, to embed the training video in \mathcal{W}^+ space. This embedding provides the source of “real” samples for the critic to distinguish from the generator’s output. No frames are rendered during training, StyleGAN is absent. This leads to considerable savings in training time and memory consumption, in particular compared to the method by Tian *et al.* (see Section 4). Only at test time do we forward the output of our generator into StyleGAN.

Note that although we focus on demonstrating our pipeline on portrait videos, no part of it other than the preprocessing step is inherently face-specific.

Data Preprocessing Our training set consists of 1 single video of < 10 minutes, sometimes much less. Before we can embed its frames in \mathcal{W}^+ via pSp, we preprocess them in the same way that the training data for the respective StyleGAN model has been preprocessed. In the case of faces this means that we compute face crops similarly as for FFHQ [14]: Since the preprocessing there was not designed with temporal smoothness in mind, we had to slightly alter the choice of landmarks used for face alignment (we use only the eye corners, never the mouth) and applied temporal lowpass filters to the rotation and scale of the face bounding boxes. This gave decent results, best seen in our supplemental video, or in Fig. 3.

Applying pSp to each frame of the training video reliably led to sequences of \mathcal{W}^+ codes that, when rendered with StyleGAN, showed a decently smooth video again (Fig. 3). The identity of the training subject was not always preserved perfectly, but this is not of interest, since our goal is anyways to generate motion for a great multitude of random actors.

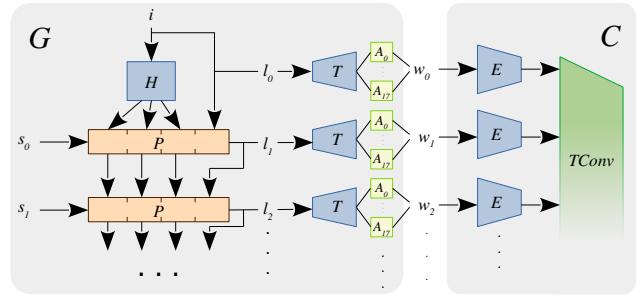


Figure 2: Our Wasserstein GAN.

This embedding provides the source of “real” samples for the critic to distinguish from the generator’s output. No frames are rendered during training, StyleGAN is absent. This leads to considerable savings in training time and memory consumption, in particular compared to the method by Tian *et al.* (see Section 4). Only at test time do we forward the output of our generator into StyleGAN.

Generator Our generator is a stack P of 4 GRU cells that processs the “per-time-step randomness” s . To initialize the GRU memory, we have the MLP H “hallucinate” some memory contents for the first three cells, whereas the last is initialized with i :

$$(h_{0,0}, h_{0,1}, h_{0,2}) := H(i) \quad h_{0,3} := i$$

After this initialization, P can produce a sequence of low-dimensional latent codes $l_k \in \mathbb{R}^{32}$ according to the following recurrence:

$$((h_{k+1,0}, \dots, h_{k+1,3}), l_{k+1}) := P(s_k, (h_{k,0}, \dots, h_{k,3}))$$

for $0 \leq k < t - 1$. Another MLP, $T : \mathbb{R}^{32} \rightarrow \mathbb{R}^{512}$ “translates” these latent codes from the space in which motion is generated to a higher dimensional one (this is reminiscent of StyleGAN’s mapping network, translating from \mathcal{Z} to \mathcal{W}). A set of learned affine transformations then maps to $\mathcal{W}^+ \subseteq \mathbb{R}^{18 \times 512}$, which gives the final output $w_0, \dots, w_{t-1} \in \mathcal{W}^+$ of our generator.

We do not claim particular novelty of this generator design. Related works [19, 27, 28] have presented similar designs. The novelty of our work lies in the ideas outlined in Section 1, i.e., supervision in \mathcal{W}^+ instead of image space, the “offset trick”, the loss functions, and the resulting massive reduction in the amount of data and resources required for training.

Only at test time, not at training time, do we forward w_0, \dots, w_{t-1} to StyleGAN for rendering of actual video frames. Note that that t might be considerably larger than the 25 time steps used at training time. In Section 4 we report numbers for $t = 250$.

Critic In contrast to G , our critic C [3] can rely on a fixed t and thus does not need to be recurrent. Instead, we first have another 6-layer MLP $E : \mathcal{W}^+ \rightarrow \mathbb{R}^{32}$ “extract” a learned set of relevant features from each \mathcal{W}^+ step and then let a temporally convolutional network consume the resulting sequence. We derived the architecture of this network from DCGAN [21], turning spatial convolutions into temporal ones and eliminating batch normalization layers, to enforce the Lipschitz constraint required for Wasserstein GANs by a gradient penalty [8].

Loss Terms & Training By training we minimize $\mathcal{L} = \mathcal{L}_{\text{WGAN}} + \lambda_{\text{GP}} \mathcal{L}_{\text{GP}} + \lambda_{\text{GAP}} \mathcal{L}_{\text{GAP}}$, where $\mathcal{L}_{\text{WGAN}} + \lambda_{\text{GP}} \mathcal{L}_{\text{GP}}$ is the usual WGAN-GP loss [8] (with $\lambda_{\text{GP}} = 50$) and \mathcal{L}_{GAP} is a novel *gradient angle penalty* (with $\lambda_{\text{GAP}} = 100$): When training our model only with the WGAN-GP loss we have observed (Section 4) that synthesizing videos for $t >> 25$ can lead to motion that seems to be “looping”, i.e. the same motion pattern is repeated over and over. Based on observations reported in previous work [27] we suspect that P learns to simply ignore the “per-time-step randomness” s and rely exclusively on $(h_{0,0}, \dots, h_{0,3})$, without modifying it much in the course of the sequence. There seems to be a tendency to make i determine the entire course of the sequence, which makes looping very likely. To counteract this, we present a new loss formulation that makes sure that the gradient of the producer output with respect to s is at least a certain fraction of the gradient with respect to i . We set:

$$\mathcal{L}_{\text{GAP}} := \left(\max \left(0, \frac{\pi}{4} - \varphi \right) \right)^2; \quad \text{with} \quad \varphi := \arctan \left(\frac{\left\| \left[\frac{\partial d}{\partial s_0}, \dots, \frac{\partial d}{\partial s_{t-2}} \right] \right\|}{\left\| \frac{\partial d}{\partial i} \right\|} \right),$$



Figure 3: Projecting face videos into \mathcal{W}^+ via pSp reasonably maintains the identity of the actor and leads to temporally smooth results.

where $d := \text{norm}(l_{t-1} - l_0)$ is the normalized Euclidean distance between the last time step and the first time step generated by P . The function `norm` here normalizes the components of the difference vector according to running statistics that are tracked during training, such that we can expect the distribution of each component to have mean 0 and variance 1. The angle φ will be close to 0 if the output of P depends mostly on i , which we want to prevent. Unless stated otherwise, we trained our models with Adam [16] for 350 epochs. We exponentially average the weights of the generator throughout training using a momentum of 0.995.

The offset trick Although our network is trained on only 1 single actor, it should be able to generate motion for a large set of randomly generated actors. We achieve this by making use of the advantageous properties of StyleGAN’s \mathcal{W}^+ space, that have been used for face editing before [9, 25, 26]: Our main assumption is that given a point in \mathcal{W}^+ , the directions into which one would need to shift this point in order to change the identity of the actor that it depicts are mostly orthogonal to those directions that would change the pose/expression/articulation of the actor. If this assumption is justified (which we demonstrate in Section 4 and in our supplemental video), it should be possible to first generate a motion trajectory for our training subject and then shift this trajectory along a direction that is orthogonal to those latter directions, to transfer it to a different actor that also exists in \mathcal{W}^+ . To find the directions responsible for pose/expression/articulation we consider the \mathcal{W}^+ embedding of our training set. A simple PCA tells us those 32 directions in which the points corresponding to our training video frames extend the furthest. Since our training frames span the relevant range of motion states but always show the same actor, we can assume that shifting points in these directions changes the state, but not the identity of the actor. Given this PCA basis and having sampled a motion trajectory $w_0, \dots, w_{t-1} \in \mathcal{W}^+$ for our training actor, we now randomly sample a point from StyleGAN’s \mathcal{Z} space, render it using StyleGAN and then embed it in \mathcal{W}^+ using pSp, obtaining w_{new} . This new point shows a random, new actor, that already is in a particular (likely nonneutral) state. For example, in the case of faces, w_{new} might correspond to a person with the mouth closed. We must not naively use this point as the starting point for our “transferred” motion trajectory, because the motion that we generated for the training actor might start with a mouth-closing motion. Applying this motion to a mouth that is already closed would likely lead to strong artifacts (see Fig. 4). Instead, we project w_{new} onto the PCA basis, resulting in w'_{new} . The point w'_{new} shows our training actor in the same state as the new actor. The difference $\Delta := w_{\text{new}} - w'_{\text{new}}$ is the exact offset by which we want to shift our motion trajectory, i.e. the new trajectory is $w_0 + \Delta, w_1 + \Delta, \dots, w_{t-1} + \Delta$. For a graphical explanation of this process, please refer to our supplemental video.

As illustrated in Fig. 4 and as demonstrated in our supplemental video, thanks to the disentangled representation of images in \mathcal{W}^+ , this simple offset operation is sufficient to transfer motion sampled for our training actor to new random actors. Also in our supplemental video we demonstrate that not embedding the new actor with pSp or naively offsetting the sequence without using the PCA basis leads to much stronger artifacts.

4 Results

Training data & Metrics For our ablation study and comparison to related work, we used sequences of faces talking into a commodity RGB camera, that were all less than 10 minutes long. In the quantitative evaluation of trained models, we use Fréchet Inception Distance



Figure 4: Top (left & right): A motion trajectory for the training actor. Bottom left: Naively shifting the motion trajectory from the training actor to some random new actor in \mathcal{W}^+ leads to strong visual artifacts. Bottom right: Projecting the code w_{new} of the new actor onto the PCA basis of the training actor first (Section 3) establishes an anchor point w'_{new} in the training actor’s point cloud. The offset by which we shift the training motion maps this anchor point to the \mathcal{W}^+ code of the new actor, reliably avoiding artifacts.

Model	Reference	FID (\downarrow)		FVD (\downarrow)	
		Short	Long	Short	Long
Ours	Original \mathcal{W}^+	54.1 \pm 0.1	54.2 \pm 1.2	627.6 \pm 25.5	629.1 \pm 24.7
		1.1 \pm 0.1	3.9 \pm 1.5	42.9 \pm 12.9	84.0 \pm 17.7
Ours \ \mathcal{L}_{GAP}	Original \mathcal{W}^+	53.9 \pm 0.5	58.8 \pm 4.8	603.7 \pm 39.4	727.1 \pm 159.1
		1.1 \pm 0.0	7.0 \pm 4.2	33.2 \pm 3.7	178.4 \pm 94.4
Tian [27]	\mathcal{W}	4.07	97.9	706.3	2130.3
Tulyakov [28]	Original	87.9	87.6	2849.3	2845.1
Saito [24]	Original	108.8	169.0	1211.4	2339.2
Munoz [19]	Original	75.5	-	755.4	-

Table 1: We compare FID and FVD scores between the training sets (or in the case of our method: the \mathcal{W}^+ embedding of the training video) and the sets of generated samples. For a description of “Short” and “Long” see Section 4 . For our method we report averages and standard deviations for 5 independently trained models.

(FID) [10] for spatial quality, and Fréchet Video Distance (FVD) [29] for the quality of motion. The reference sets for all methods are their training datasets, preprocessed as required by the particular method. For our method we report scores in relation to both the actual original footage, as well as its \mathcal{W}^+ embedding (which is what our method is trained on).

Each method was trained on the training video, depicting only one actor, as our task demands. We then generated two sets of videos from each model: The “Short” set consists of 2048 videos that are as long as the temporal window the particular method considered at training time (see column “ t ” in Table 2). The “Long” set consists of 128 video segments, that all have at least 128 frames. The technique by Munoz *et al.* is not able to produce samples longer than its training window, which is why Table 1 contains no numbers for this set. FID scores are computed on 8,000 frames randomly sampled from the reference and generated sets. FVD scores are computed on 2048 videos from each of the two sets, with the duration of the videos again equal to the default temporal window length of each method.

We also evaluate the temporal consistency of facial identity using a variant of the Average Content Distance (ACD) [28]: For each generated frame, we obtain the identity features from a popular facial recognition library [1] and compute the average L2-distances between all pairs of frames of the video. This score is then averaged over all generated videos.

Method	Sample	Resolution	t
Ours (random actor)		1024^2	25
Tian <i>et al.</i> [27]		1024^2	16
Munoz <i>et al.</i> [19]		128^2	16
Tulyakov <i>et al.</i> [28]		64^2	16
Saito <i>et al.</i> [24]		192^2	16

Table 2: Qualitative comparison to previous methods. Tian *et al.* does synthesize motion for the training identity by default, whereas other previous methods can synthesize motion *only* for the training identity. While all methods appear to generate decent spatial quality, the spatial and/or temporal resolutions of previous methods is severely limited by their computational resource demands. The temporal quality of the generated videos can only be judged from our supplemental video.

Video Generation Fig. 5 shows sequences generated by three different models, each for the respective training identity. As shown in Fig. 1 however, the “offset trick” allows us to generate motion for randomly sampled identities as well, even though our training datasets always contain only 1 actor. All videos are synthesized at a resolution of 1024×1024 and even though our method was trained only on a temporal window of 25 frames, we can easily generate videos that are much longer, e.g. 1500 frames. The quality of motion can only be judged in our supplemental video, not on paper.

Comparison to Previous Methods We compare to several previous techniques by training them all on the same dataset and evaluating the metrics described above.

The very recent method by Tian *et al.* [27] also learns a trajectory in the StyleGAN latent space, which makes it the most related to ours. We compare to their cross-domain setting, i.e. the StyleGAN generator they use was pretrained on FFHQ [14], but the motion is learned from our training video. The model that the authors kindly trained for us does by default not generate videos of the training identity, but instead samples random identities from StyleGAN’s \mathcal{W} space (*not* \mathcal{W}^+ !). This is a problem for the computation of FID and FVD scores, which always compare the generated distribution (random identities) to the training distribution (our specific training identity). We thus need to force the model to generate samples depicting



Figure 5: Videos generated by our method, for the training identities.

our training identity, as otherwise it would have been impossible to compute fair scores. We achieve this by sampling random frames from our training video and projecting [14] them to \mathcal{W} . The resulting \mathcal{W} points are then “injected” as the initial code for the model to condition its generated sequences on. As mentioned in previous works [22, 25], \mathcal{W} cannot represent real images as faithfully as \mathcal{W}^+ , which is why we did not use our training video as the reference set, but its \mathcal{W} embedding. Even though the method is able to produce videos at resolution 1024×1024 , we show (in our supplemental material) that the model the authors trained on our dataset is not able to generate a lot of facial motion, i.e. while the camera is panning, the facial expression is very static. This is reflected in Table 1. Training this method is rather expensive and requires about 5 days on 8 Quadro RTX 8000 GPUs for a resolution of 1024×1024 , i.e. 40 GPU days in total. In contrast, our method is trained on a single Quadro RTX 8000 GPU in around 6 hours, a speedup of factor 160.

The methods by Saito *et al.* [24] and Tulyakov *et al.* [28] generate realistic motion, but are very limited in terms of spatial resolution (192^2 and 64^2 respectively). As we show in our supplemental video, the results for Munoz *et al.* [19] include strong structural artifacts. All three methods are not capable of generalizing their output to random identities after being trained on only one subject, i.e. their training set would have to be much larger to make them generate the diversity of output identities Tian *et al.* and our method achieve. We attempted to compare to further methods [33, 34], but could not get access to their code.

Tab. 1 reports Fréchet distances for all methods. We have also computed ACD scores for our method (random actors) and for Tian *et al.*: While our method averages at 5.68 (over 5 models) Tian *et al.* achieve a score of **0.54**. We attribute this large difference to the very limited facial motion generated by Tian *et al.*, that of course makes it much easier to preserve the identity. For visual impressions of this observation please see our supplemental video.

Evaluation of the Gradient Angle Penalty Table 1 shows that while removing \mathcal{L}_{GAP} from our training objective slightly improves the scores for “short” samples, it considerably increases the FVD scores for “long” samples. However, since the primary purpose of \mathcal{L}_{GAP} is to prevent looping, which is maybe not effectively captured by FVD, we have recorded a very short training sequence (one single sentence, spoken three times, 20 seconds in total) that provoked strong looping artifacts in 19 out of 20 independently trained models if \mathcal{L}_{GAP} was absent, but led to looping only in 6 out of 20 models that were trained with the loss in place. This suggests that \mathcal{L}_{GAP} is indeed making looping artifacts much less likely.

Proof of concept: Hands To demonstrate that our method should in principle be applicable to content categories other than talking faces, we have conducted a proof-of-concept experiment for hands: We recorded the right hand of a subject for 1 hour, performing various types of motions (like showing numerals or performing a set of gestures), resulting in a dataset of around 100k frames. The only constraint was for the hand to always turn the palm to the camera and to never leave the recording space. This dataset we successfully used for training a StyleGAN model and the corresponding pSp inverter, both for resolution 256×256 . With these models available, we were able to train our temporal model with a temporal window of 75 time steps, on several test sequences (each about 8000 frames). Results are shown in Fig. 1 and in our supplemental video.

Proof of concept: Cars As a third domain, we applied our method to the category of cars. We used the official StyleGAN2 checkpoint for the LSUN-Car dataset [35] and trained pSp



Figure 6: A synopsis of an experiment on cars: The videos we recorded (left, license plate censored) were embedded into \mathcal{W}^+ (center) using a pSp model trained on the recorded video (because we did not have a more general pSp of sufficient quality available). We then trained our temporal architecture on the embedding, which allowed us to generate new motion for the car (right). Results on more cars are shown in our supplemental material.

from scratch. LSUN-Car is a much more challenging dataset than FFHQ, because the data is not as "clean" and because there is no alignment. We thus did not manage to train pSp to the same level of precision as the official FFHQ checkpoints. As a result, the \mathcal{W}^+ embeddings contained clearly visible artifacts and the identity of the car drifted depending on the orientation. In order to nevertheless demonstrate that the core concept of our method works, we trained pSp a second time, but this time on the frames of our recordings. This way we easily achieved decent embeddings, showcased in Fig. 6 and in our supplemental videos. The disadvantage of this approach is that we cannot demonstrate the offset trick in this case, as our pSp model has only ever seen the training car and cannot embed cars randomly sampled from StyleGAN's latent space. Of course we would have preferred to choose an object category (other than faces) for which there is a *general* high-quality, temporally stable embedding method. However to the best of our knowledge, nobody has demonstrated such a method yet. We show qualitative results in Fig. 6 and in our supplemental material.

5 Conclusion

We have presented a temporal GAN for the unconditional generation of high-quality videos. Based on embedding the footage of only 1 actor into the latent space of StyleGAN, we are able to train our model with a minimal amount of resources and can nevertheless generate diverse motion of arbitrary length for a great number of random actors at high spatial resolution. Although these abilities also have their limitations (see supplemental), we hope that our work can pave the way for future innovations in video generation.

Acknowledgements We thank the the authors of [27] for training their model on the training data we sent them. We also thank Pramod Rao for his invaluable support in conducting the experiments for our evaluation section. This work was supported by the ERC Consolidator Grant 4DReply (770784).

References

- [1] https://github.com/ageitgey/face_recognition.
- [2] Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Towards high resolution video generation with progressive growing of sliced Wasserstein GANs, 2018.
- [3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN, December 2017.
- [4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019.
- [5] L. Chai, Y. Liu, W. Liu, G. Han, and S. He. CrowdGAN: Identity-free interactive crowd video generation and beyond. *PAMI*, 2020.
- [6] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets, 2019.
- [7] Emily L Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NeurIPS*, 2017.
- [8] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NeurIPS*, volume 30. Curran Associates, Inc., 2017.
- [9] Erik Häkkinen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: Discovering Interpretable GAN Controls. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *NeurIPS*, volume 33, pages 9841–9850. Curran Associates, Inc., 2020.
- [10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, 2017.
- [11] Kibae Hong, Youngjung Uh, and Hyeran Byun. ArrowGAN : Learning to generate videos by learning arrow of time, 2021.
- [12] Emmanuel Kahembwe and Subramanian Ramamoorthy. Lower dimensional kernels for video discriminators. *Neural Networks*, 132:506–520, 2020.
- [13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018.
- [14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.
- [15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020.

- [16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015.
- [17] Willi Menapace, Stephane Lathuiliere, Sergey Tulyakov, Aliaksandr Siarohin, and Elisa Ricci. Playable video generation. In *CVPR*, 2021.
- [18] Jacob Menick and Nal Kalchbrenner. Generating high fidelity images with subscale pixel network and multidimensional upscaling. In *ICLR*, 2019.
- [19] Andres Munoz, Mohammadreza Zolfaghari, Max Argus, and Thomas Brox. Temporal shift GAN for large scale video generation. In *WACV*, 2021.
- [20] Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. Face identity disentanglement via latent space mapping. *ACM Transactions on Graphics (Proceedings of SIGGRAPH-Asia)*, 39(6), 2020.
- [21] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.
- [22] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a StyleGAN encoder for image-to-image translation. In *CVPR*, 2021.
- [23] M. Saito, E. Matsumoto, and S. Saito. Temporal generative adversarial nets with singular value clipping. In *ICCV*, 2017.
- [24] Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal GAN. *IJCV*, 128(10-11), 2020.
- [25] Yujun Shen, Jinjin Gu, Xiaou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *CVPR*, 2020.
- [26] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. StyleRig: Rigging StyleGAN for 3D control over portrait images. In *CVPR*. IEEE, 2020.
- [27] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=6puCSjH3hwA>.
- [28] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. In *CVPR*, 2018.
- [29] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2019.
- [30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NeurIPS*, volume 30. Curran Associates, Inc., 2017.

-
- [31] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics, October 2016. URL <http://arxiv.org/abs/1609.02612v3>; <http://arxiv.org/pdf/1609.02612v3.pdf>.
 - [32] Yaohui Wang, Francois Bremond, and Antitza Dantcheva. InMoDeGAN: Interpretable motion decomposition generative adversarial network for video generation, 2021.
 - [33] Dirk Weissenborn, Jakob Uszkoreit, and Oscar Täckström. Scaling autoregressive video models. In *ICLR*, 2020.
 - [34] Shuquan Ye, Chu Han, Jiaying Lin, Han Guoqiang, and Shengfeng He. Coherence and identity learning for arbitrary-length face video generation. In *ICPR*, 2020.
 - [35] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, abs/1506.03365, 2015.
 - [36] V. Yushchenko, N. Araslanov, and S. Roth. Markov Decision Process for video generation. In *ICCVW*, 2019. doi: 10.1109/ICCVW.2019.00190.

arXiv:2107.07224v2 [ccs.CV] 30 Nov 2021

StyleVideoGAN: A Temporal Generative Model using a Pretrained StyleGAN

SUPPLEMENTAL MATERIAL –

Cereon Fox

gfox@mpi-inf.mpg.de

Ayush Tewari

atewari@mpi-inf.mpg.de

Mohamed Elgharib

elgharib@mpi-inf.mpg.de

Christian Theobalt

theobalt@mpi-inf.mpg.de

Max Planck Institute for Informatics

Saarland Informatics Campus

Saarbrücken, Germany

1 Quantitative results for additional identities

In Tables 1 and 2 we give more quantitative results, on the additional identities depicted in Fig. 1. Our method outperforms the state of the art for these subjects as well.

Model	Reference	FID (\downarrow)		FVD (\downarrow)	
		Short	Long	Short	Long
Ours	Original \mathcal{W}^+	62.9 ± 0.2	65.1 ± 0.8	846.7 ± 19.1	944.1 ± 51.7
Ours \ \mathcal{L}_{GAP}	Original \mathcal{W}^+	0.6 ± 0.0	2.0 ± 0.2	39.2 ± 19.2	51.8 ± 18.1
Tulyakov [7]	Original	62.9 ± 0.1	64.8 ± 3.0	848.8 ± 8.0	926.2 ± 58.5
Saito [5]	Original	0.7 ± 0.0	4.3 ± 2.8	57.0 ± 93.1	110.7 ± 74.9
Munoz [3]	Original	76.5	77.8	1318.7	1338.7

Table 1: FID and FVD scores for subject # 2. All metrics were computed as described in the main paper.

2 Experimental Details

2.1 Training details

We trained all methods with their default hyperparameters, except for Saito et al, where we had to adjust the batch size to 2 and set `clstm channels = 512`. The authors of

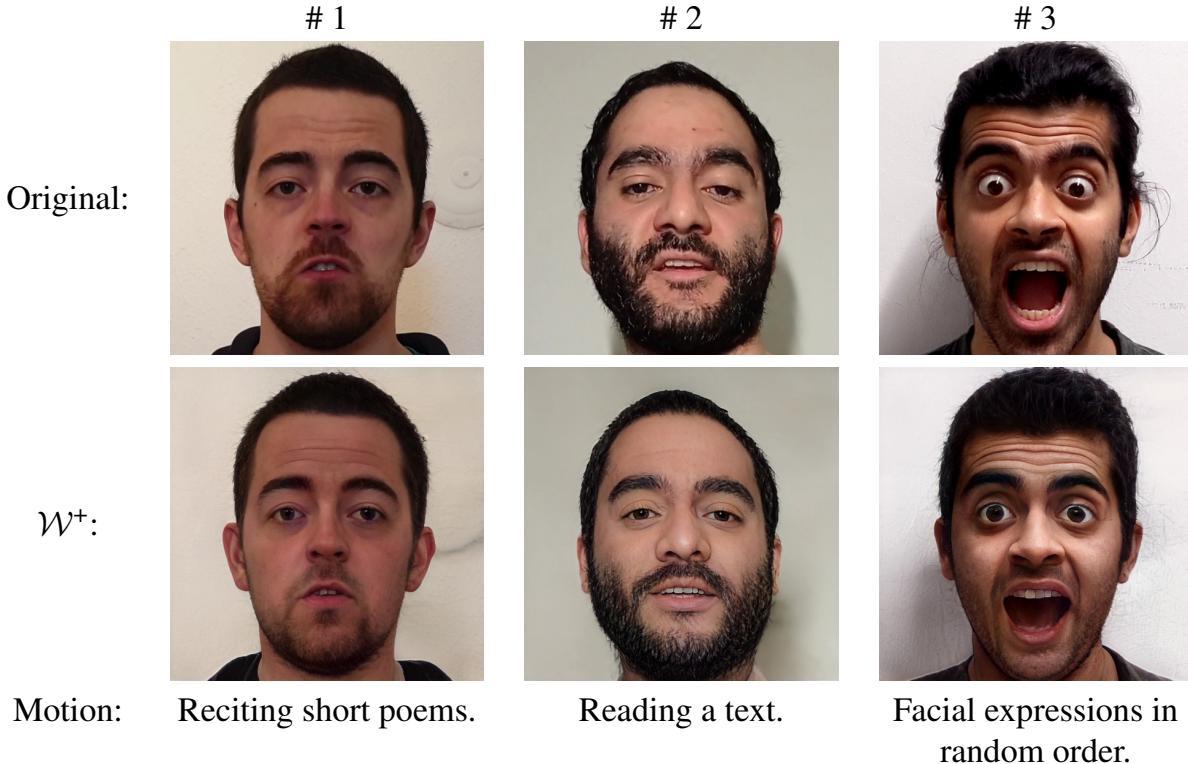


Figure 1: Screenshots from the training sequences we used for quantitative evaluation. The numbers in this supplemental document have been computed for subjects # 2 and # 3. The ones reported in the main paper are for # 1.

Tian *et al.* [6] kindly trained their technique on the training data we sent them. We trained all methods with at least the computational resources that our method uses, but usually gave them much more training time. Table 3 gives a comprehensive overview.

Each training video contains 1 single actor/object. We computed quantitative evaluations for 3 actors, reported in the main paper and this supplemental. For the proof of concept on hands, all training data was recorded from one actor. In the case of cars, we show qualitative results of 3 different cars. We use a batch size of 128, learning rate of 0.005, and exponential averaging of the weights with a momentum of 0.997 in all experiments. All temporal generative networks are trained for 350 epochs.

2.2 Evaluation details

FID scores have been computed by sampling 8000 frames from both the training set (as preprocessed for the particular method) and the set of generated videos. We used the FID implementation in <https://github.com/mseitzer/pytorch-fid>.

FVD scores have been computed by sampling 2048 video slices from both the training set (as preprocessed for the particular method) and the set of generated videos. For each method, and regardless of the length of the samples we generated (“short” versus “long”), the videos we sampled were always 25 frames long for our method and 16 frames long for the previous methods. We used the FVD implementation in

Model	Reference	FID (\downarrow)		FVD (\downarrow)	
		Short	Long	Short	Long
Ours	Original \mathcal{W}^+	52.7 ± 0.2	53.7 ± 0.5	589.2 ± 9.7	625.3 ± 3.1
Ours \ \mathcal{L}_{GAP}	Original \mathcal{W}^+	3.7 ± 0.2	4.8 ± 0.4	61.4 ± 4.2	98.6 ± 11.3
Tulyakov [7]	Original	52.3 ± 0.5	55.2 ± 2.1	590.9 ± 15.2	679.0 ± 28.9
Saito [5]	Original	3.4 ± 0.1	8.2 ± 2.0	54.0 ± 3.6	133.5 ± 28.5
Munoz [3]	Original	123.0	141.1	1163.3	1500.3
	Original	82.1	270.3	823.9	2090.8
	Original	83.3	-	1037.0	-

Table 2: FID and FVD scores for subject # 3. This subject was not talking, but instead performing some simple face motions in a random order (like smiling or acting surprised). The training video is only 1 minute and 20 seconds in length.

	GPU Used	GPU Memory	Training time (max)
Ours	Quadro RTX 8000	48 GB (8GB used)	approx. 6 hours
Tulyakov <i>et al.</i> [7]	GeForce RTX 2080	12 GB	approx. 6 hours
Munoz <i>et al.</i> [3]	Quadro RTX 8000	48 GB	approx. 2 days
Saito <i>et al.</i> [5]	GeForce RTX 2080	12 GB	approx. 15 hours

Table 3: Training details for the various methods. The authors of Tian *et al.* [6] trained their method for us.

ACD scores have been computed always on 128 “long” samples drawn from the trained models. For our method these long samples were 400 frames long. For Tian *et al.* they were 128 frames long. Having a good ACD becomes harder as sequences grow longer.

3 Limitations

Even though we are able to further to the state of the art in video generation, in particular with respect to the amounts of computational resources and training data necessary to generate a large amount of diverse videos, our method has several limitations: For one, the quality of our generated videos strictly depends on the quality of the underlying StyleGAN model and its corresponding pSp inverter. For example, in the case of faces we have observed that nontrivial video backgrounds tend to not be represented in a temporally stable way. The importance of temporally stable embedding is also underlined by an experiment we made with a BigGAN model [1] instead of a StyleGAN model: We used a SOTA optimization-based method [2] to embed several short videos (to the best of our knowledge there are no encoder-based inversion methods, see [8]). The embeddings contain strong temporal noise, making training our method pointless (see supplemental video). For the sources of the videos, see Table 8

Another limitation is the fact that while our approach does not contain any inherently face-specific components and even though we are showing a proof-of-concept for animating hands and cars, it is still unclear whether all the advantageous properties of StyleGAN’s \mathcal{W}^+ space can be made use of in any arbitrary domain, e.g. if our offset trick will work there.

4 Detailed architecture

For the sake of completeness, Tables 4 to 7 give detailed specifications for the architecture components that we outlined in Figure 2 of the main paper.

Input	Module	Outputs (Dimensionality)
i	4 layers (Fully Connected + LeakyReLU)	$m (3 \times 32)$
m	BatchNorm	$(h_{0,0}, h_{0,1}, h_{0,2}) (3 \times 32)$

Table 4: The “hallucinator” H is responsible for producing some initial contents for the GRU memory. For each one of the four stacked GRU cells it produces a vector of length 32.

Input	Module	Outputs (Dimensionality)
$s_0, (h_{0,0}, \dots, h_{0,3})$	GRU (4 stacked cells)	$(h_{1,0}, \dots, h_{1,3}), l_1 (3 \times 32, 32)$
$s_1, (h_{1,0}, \dots, h_{1,3})$	GRU (4 stacked cells)	$(h_{2,0}, \dots, h_{2,3}), l_2 (3 \times 32, 32)$
$s_2, (h_{2,0}, \dots, h_{2,3})$	GRU (4 stacked cells)	$(h_{3,0}, \dots, h_{3,3}), l_3 (3 \times 32, 32)$
...

Table 5: The feature generator P consists of four stacked GRU cells. Its hidden state is initialized with the output of H (Table 4) and it translates a sequence of random vectors s_k into intermediate latent codes l_{k+1} for $0 \leq k < t - 1$ in a recurrent fashion.

Input	Module	Outputs (Dimensionality)
l_k	BatchNorm + Affine transform + Pixel-Norm	$l'_k (512)$
l'_k	4 layers (FullyConnected + LeakyReLU)	$v'_k (512)$
v'_k	BatchNorm + Affine transform	$v_k (512)$
v_k	18 parallel layers (FullyConnected + LeakyReLU + BatchNorm)	$w_k (18 \times 512)$

Table 6: The latent mapper T is an MLP that transforms the outputs l_k of the feature generator P into StyleGAN latent codes $w_k \in \mathcal{W}^+$: After a 4-layer MLP that widens the dimensionality from 32 to 512, we employ 18 independent fully connected layers (with LeakyReLU activation), in a way similar to how StyleGAN broadcasts its \mathcal{W} vectors to its 18 Style layers.

Input	Module	Outputs (Dimensionality)
w_k	FullyConnected + LeakyReLU (2 layers)	$e'_k (512)$
e'_k	FullyConnected + LeakyReLU (4 layers)	$e_k (32)$
e_0, \dots, e_{t-1}	1D-version of the DCGAN critic [4], with 32 input channels	Critic Output (1)

Table 7: The latent critic takes a sequence w_0, \dots, w_{t-1} of StyleGAN latent codes as input, and produces a scalar output.

BigGAN category	YouTube key
indigo bird (014)	DZOiCmxSU2k
green mamba (064)	hG4Wvp0U18A
bison (347)	L4eOhuLDfeU
gazelle (353)	jMIIiB9DnRXg

Table 8: The sequences we embedded into the BigGAN latent space were excerpts of YouTube videos.

References

- [1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019.
- [2] Minyoung Huh, Jun-Yan Zhu, Richard Zhang, Sylvain Paris, and Aaron Hertzmann. Transforming and projecting images to class-conditional generative networks. In *ECCV*, 2020.
- [3] Andres Munoz, Mohammadreza Zolfaghari, Max Argus, and Thomas Brox. Temporal shift GAN for large scale video generation. In *WACV*, 2021.
- [4] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.
- [5] Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal GAN. *IJCV*, 128(10-11), 2020.
- [6] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=6puCSjH3hwA>.
- [7] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. In *CVPR*, 2018.
- [8] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. GAN inversion: A survey. *CoRR*, abs/2101.05278, 2021. URL <https://arxiv.org/abs/2101.05278>.