Self-Supervised Video GANs: Learning for Appearance Consistency and Motion Coherency

Sangeek Hyun, Jihwan Kim, Jae-Pil Heo*
Sungkyunkwan University

Abstract

A video can be represented by the composition of appearance and motion. Appearance (or content) expresses the information invariant throughout time, and motion describes the time-variant movement. Here, we propose self-supervised approaches for video Generative Adversarial Networks (GANs) to achieve the appearance consistency and motion coherency in videos. Specifically, the dual discriminators for image and video individually learn to solve their own pretext tasks; appearance contrastive learning and temporal structure puzzle. The proposed tasks enable the discriminators to learn representations of appearance and temporal context, and force the generator to synthesize videos with consistent appearance and natural flow of motions. Extensive experiments in facial expression and human action public benchmarks show that our method outperforms the state-of-the-art video GANs. Moreover, consistent improvements regardless of the architecture of video GANs confirm that our framework is generic.

1. Introduction

Generative Adversarial Networks (GANs) [16] are one of the major research topics in the spotlight, due to their impressive capability to model the data distribution in an unsupervised way. The recent advances in the aspects of objectives [5, 4, 26] and architectures [20, 21, 22] alleviate the chronic problems of GANs such as the mode collapse and training instability. Thanks to these sustained research efforts, the latest techniques enable us to synthesize visually plausible and diverse images.

With these advances in the image domain, the problem of generating videos has emerged in recent years. Pioneering attempts [37, 30] have started with mapping a latent vector to a video with spatio-temporal convolutions. On the other hand, the following methods [35, 38] have proposed the video generation frameworks mainly targeting to decompose spatio-temporal latent space into motion and content subspaces. Their efforts to disentangle the latent space to appearance and motion ones have reduced the complexity of generating videos.

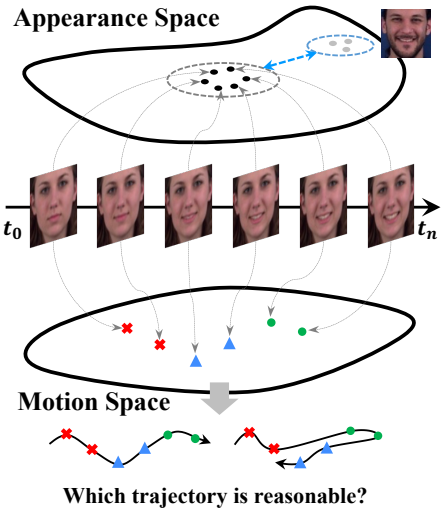


Figure 1. **Illustration of conditions for realistic videos.** Video can be represented as the composition of appearance and its motions. For the natural and realistic video, appearances and motions in the same video have to be consistent throughout time. In other words, appearance representation should be similar among frames of the same video compared to the frames from different videos, and the flow of the motion would be fit into the temporal context.

In spite of the aforementioned successes, extending GANs to the video domain is still challenging. One of the main causes is high dimensionality of videos. Compared to images, videos have one more dimension, time. The time dimension exponentially expands the video space from the image space with respect to the number of frames. In this huge space, perceptually satisfying videos could account for an extremely small portion because they have to fulfill not only spatial realness but also temporal coherency. Therefore, it is obvious that mapping low-dimensional latent vectors to visually plausible videos is a lot more complex than the case of images.

From this perspective, we suggest to reduce the video

*Corresponding author

space. That is, we have to contract the possible space by exploiting the prior knowledge about "realistic" videos. This provides essential constraints towards the realness of the synthesized videos. For instance, let us assume real videos consist of only grayscale frames instead of color ones. In this case, typical RGB representations are redundant since videos are fully represented in grayscale. Hence, we can reduce the space by forcing the videos to have a single channel with the prior knowledge in the assumption. By this, the task of GANs can be turned into a simpler one because it shrinks many possible mappings.

Then what are the major components to achieve for generating "realistic" videos? We hypothesize two prominent constraints for realistic videos; consistency of appearance and coherency of motion. Fig. 1 describes the two types of conditions. First, appearance should be consistent over time, especially for short videos. For example, the identity of an actor should be retained over time. Second, the motion should be naturally progressed with coherency along time. A physically impossible human movement can be a counterexample for it. These constraints should be satisfied when generating realistic videos.

To meet these necessary conditions, we present Self-supervised Video GANs (SVGAN), which imposes explicit constraints on appearance and motion with two pretext self-supervision tasks; appearance contrastive learning and temporal structure puzzle. Appearance contrastive learning makes the discriminator to learn the representations of appearance which is invariant throughout time in videos. On the other hand, temporal structure puzzle forces the discriminator to figure out whether the video is coherent or not in temporal ordering. Furthermore, we distill these explicit constraints to the generator so that it synthesizes videos satisfying those conditions in a collaborative way. Eventually, these constraints significantly reduce the huge video space to its small portion of spots where realistic videos can exist so that the video generation problem becomes less complex. Different from previous approaches [37, 30, 35, 38] mainly focusing on the architecture of the generator, we bring the focus into the objective of the discriminator in video GANs. The proposed self-supervision tasks directly involve the objective of the discriminator as a new direction towards realistic videos in the literature.

Our main contributions are summarized as follows:

- We mainly focus on the objectives of the discriminator and its effects on video GANs with self-supervision. To our best knowledge, it is the first attempt to shed light on the discriminator objectives in video GANs.
- We propose Self-supervised Video GANs (SVGAN), which explicitly imposes constraints on GANs with two pretext self-supervision tasks; appearance contrastive learning and temporal structure puzzle. They constrain GANs to synthesize videos with (1) invariant appearance through time and (2) naturally progressed flow of motion, respectively.
- Our extensive experiments on challenging benchmarks of facial expressions and human actions validate that our method significantly enhances video generation performance of the state-of-the-art techniques regardless of the generator architectures which previous approaches have mostly focused on.

2. Related Work

2.1. Video GANs

Generative Adversarial Networks (GANs) [16] is one of the rapidly growing research topics in computer vision community. Especially, in the image domain, there are lots of recent advances [7, 21, 28, 44, 20, 22] which generate large-scale and high-fidelity images. Different from that impressive progress, extending GANs to the temporal domain is rather tough. The major challenge towards video GANs is the fact that realistic videos need to fulfill both consistent and plausible appearances and natural motion.

To overcome the aforementioned challenge, there are several studies to model video distribution based on GANs. The most pioneering work is VGAN [37], which generates foreground and background of video separately with a two-stream spatio-temporal generator. TGAN [30] introduces a temporal generator that produces a set of frame latent vectors from a single video latent vector to yield video frames. Those methods basically focus on mapping an entire video into a single point in the latent space. On the other hand, MoCoGAN [35] decomposes spatio-temporal latent space into motion and appearance subspaces to reflect the dynamic characteristics of the temporal domain. Moreover, they newly utilize dual discriminators, video and image discriminators, for stability while training GANs. Similarly, G^3AN [38] proposes a three-stream generator to disentangle motion and appearance with a self-attention module for a spatio-temporal consistency of videos. Also, there are researches which mainly concentrate on generating high-fidelity videos [31, 1, 10]. In addition, there are maximum likelihood based video generation models [41, 42], but we do not directly compare those models with ours since our work focus on enhancing the performance of GANs.

The majority of previous methods focus on searching for the generator architecture optimized to the video generation task. Unlike those methods, here we take an orthogonal approach, contributing to the objectives of discriminators in video GANs.

2.2. Self-supervised Learning

Self-supervised learning is one of the promising ways to learn feature representations without any human super-

vision. Generally, it produces pseudo labels to solve a predefined pretext task, which is commonly achieved by transforming input data. In the image domain, one type of recent advances employs geometrical transformations (or verifications) such as relative patch prediction [12], solving jigsaw puzzle [29] and rotation prediction [15].

Contrastive learning is another stream in self-supervised learning, which maximizes the agreement between representations of two different views (e.g. transformation) of the given image. For instance, Exemplar-CNN [13] and instance discrimination [40] treat each image instance as an individual class and perform a classification task in parametric and non-parametric ways, respectively. Recently, MoCo [17] introduces a momentum encoder for consistency of feature representation during training and a dynamic dictionary for a large number of negative samples to enhance the chance to pick the hard-negative samples. On the other hand, SimCLR [8] suggests a simple framework for contrastive learning with an extensive analysis of a model architecture, augmentation methods and loss functions.

In the video domain, several recent methods [14, 25, 39] focus on learning spatio-temporal representations for action recognition and video retrieval tasks. ShuffleLearn [27] and clip order prediction [43] learn to predict the correct temporal order of shuffled images and video clips, respectively. Similarly, video jigsaw [2] and solving cubic puzzle [23] learn to match 3D permutation in an image and video level.

Recently, those impressive representations of self-supervised learning are brought into image GANs [9, 34, 19]. They introduce an auxiliary task (e.g. rotation prediction) to the discriminator for maintaining visual representations of the real distribution during unstable training process of GANs. In contrast to aforementioned methods, we bring self-supervised learning into the generation of videos. To our best knowledge, this is the first self-supervised approach in the field of video GANs. Furthermore, we assign two pretext tasks which are tailored to the video generation task for enhancing consistency of appearance and coherency of motion.

3. Self-supervised Video GANs

Video GANs are the framework to generate the video, a sequence of frames. Similar to general GANs for images, it learns to map the random noise into the video space by minimizing the gap between distributions of real data and fake samples. However, different from the architectures of image GANs, recent video GANs [35, 38] deploy dual discriminators; image discriminator D_I and video discriminator D_V. Each discriminator solves the binary classification problem (real and fake) on its domain, and provides feedback to the generator G that learns to deceive discriminators. As a result, the typical objective of Video GANs is defined as:

$$\begin{aligned} &\min_{G} \max_{D_I, D_V} V(G, D_I, D_V) = \\ &\mathbb{E}_{x\sim P_I(x)} \log(D_I(x)) + \mathbb{E}_{x\sim P_{\hat{I}}(x)} \log(1 - D_I(x)) + \\ &\mathbb{E}_{v\sim P_V(v)} \log(D_V(v)) + \mathbb{E}_{v\sim P_{\hat{V}}(v)} \log(1 - D_V(v)), \end{aligned} \tag{1}$$

where P_I and $P_{\hat{I}}$ are real and fake distributions of the images, and P_V and $P_{\hat{V}}$ are those of the videos.

In addition to adversarial learning, we assign (1) appearance contrastive learning to the image discriminator D_I, and (2) temporal structure puzzle to the video discriminator D_V as shown in Fig. 2. These two pretext tasks put direct constraints to GANs so that they generate videos with consistent representations of appearance over time, and natural motion. Now, we deliver the two supervision tasks in detail through the following two subsections. Then we cover collaborative learning, which gives the constraints to the generator G, followed by the subsection on the full objectives of SVGAN. Note that we consistently regard the discriminator as the encoder for the pretext tasks.

3.1. Appearance Contrastive Learning

Let us first solidify our definition of the word "appearance". In this paper, appearance is information which is invariant with respect to the temporal axis. For example, the identity of a particular person in a facial expression video can be seen as a subset of appearance in our definition. In this case, appearance consistency indicates that an identical appearance has to be maintained within the entire frames in the same video.

Here, we strictly re-define the meaning of appearance consistency as follows: appearance representations extracted from any pairs of frames in the same video have to be **relatively closer** than representations of frames from the other videos. This definition allows us to bring about the concept of contrastive learning for video generation.

Recent approaches of contrastive learning [8, 17] in the image domain have a goal to maximize the agreement between two different views of a single image. With transformations $t, t^{'} \in T$, they regard the randomly augmented images x_t and $x_{t'}$ for a given image x as positive samples, and augmented images from the other images $\tilde{x}$ as negative samples. Then, the objective of contrastive learning maximizes the cosine similarity between positive samples higher than similarity to the negative samples. In this case, if we regard a set of transformations T as the sampling along the temporal axis, the objective of contrastive learning can correspond to our definition of appearance consistency.

From this perspective, we present appearance contrastive loss L_A. Let us elaborate appearance contrastive learning formally. We denote a video as $V = \{x_1, x_2, ..., x_n\}$, where x_i is the i^{th} frame. In order to learn the time-invariant representation called appearance, we sample two frames of dis-

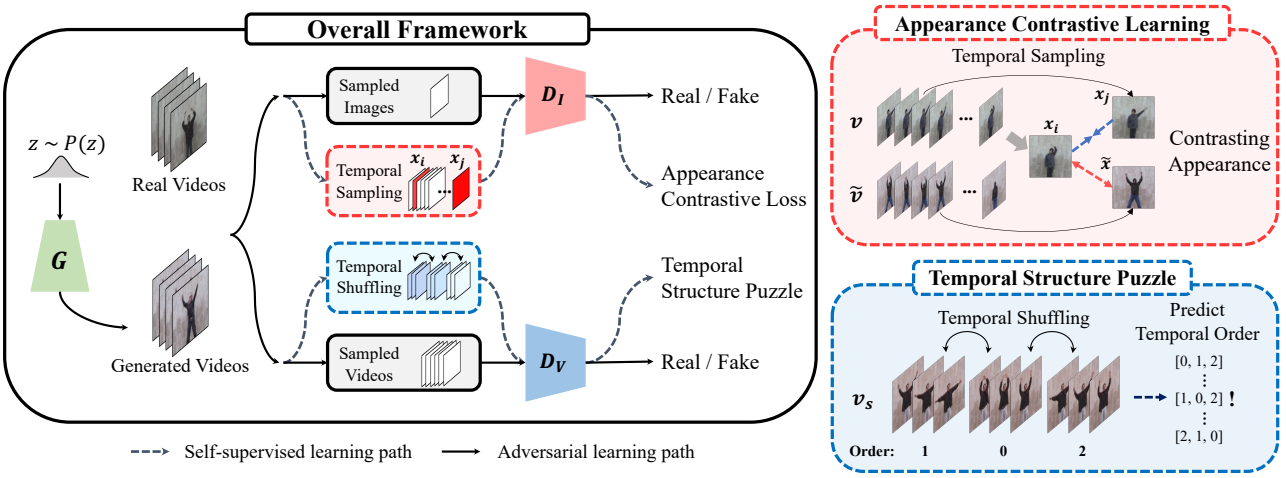


Figure 2. **Overall framework of the proposed method.** The framework has two mainstreams; adversarial learning and self-supervised learning paths. Adversarial learning follows a typical real vs. fake training scheme of video GANs with dual discriminators, D_I and D_V for images and videos, respectively. The pretext tasks for self-supervised learning are appearance contrastive learning and temporal structure puzzle. Appearance contrastive learning makes D_I to encode time-invariant representations of appearance throughout the video, which consequently guides the generator G to maximize the similarity of representations among frames of the same video. Temporal structure puzzle encourages D_V to learn coherent representations of temporal structure by correcting the shuffled order, which makes G synthesize natural videos with coherent motions.

tinct time steps $\{x_i, x_j\}_{i \neq j}$ from the same video, and a set of negative samples $\tilde{X} = \{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_k\}$ from the other videos. Afterwards, by forwarding the images into the encoder (or the discriminator) $f(x)$, we obtain the representation vectors $h_i, h_j = f(x_i), f(x_j)$. Then objective function of appearance contrastive learning is defined as:

$$L_A = -\log \frac{\exp(\text{sim}(h_i, h_j)/\tau)}{\sum_{\tilde{x} \in \tilde{X}} \exp(\text{sim}(h_i, f(\tilde{x}))/\tau)}, \quad (2)$$

where $\text{sim}(\cdot, \cdot)$ is cosine similarity, τ denotes a temperature parameter, and h_i and h_j are representations of x_i and x_j, respectively.

As in typical self-supervised GANs [9], we assign the above pretext task to the discriminator. Specifically, it is assigned to the image discriminator D_I since appearance contrastive learning is an image-level self-supervision task. However, the generator implicitly learns consistent appearance by only adversarial learning with the discriminator. Thus, we also assign the proposed task to the generator so that it also minimizes the loss of the discriminator by synthesizing appearance consistent videos. It is collaborative learning between the discriminator and the generator different from adversarial learning. We further elaborate this collaborative learning in Sec. 3.3.

3.2. Temporal Structure Puzzle

The motion in the video can be represented as differences among consecutive frames. The important point for natural motion is temporal coherency which indicates the global context of the movements of objects. Hence, the lack of this component can lead to visually or semantically unsatisfying videos. For instance, videos that consist of frames that are randomly shuffled seem unnatural to humans.

In this point of view, we present temporal structure puzzle that is a modified version of the clip order prediction task [43]. Here, we denote a partial consecutive sequence of frames from a video as a clip. Temporal structure puzzle is a task of correcting the order of temporally shuffled clips sampled from a single video. For more details, we first uniformly divide the video into multiple disjoint clips and concatenate them with a random order to form a new video. The encoder then classifies the shuffled permutation from the feature of a temporally re-organized video extracted by the encoder. This process makes the model to learn representation about motion and temporal structure of the video since the encoder has to know the flow of motion in the entire sequence to predict correct order.

The main distinction against the previous method [43] is the learned representation of the encoder from the task. In [43], they forward divided clips into the encoder individually. Then, encoded clips are concatenated and classified using a few additional layers. This may lead the encoder not directly to learn about the re-ordering task, since the encoder has only a single clip as input, not the entire video. Instead, we feed not a clip but a temporally re-organized video to the encoder. This re-organized video enables the encoder to directly learn the task with a larger temporal receptive field. Especially for generating short videos, this large receptive field is more proper because one clip has an

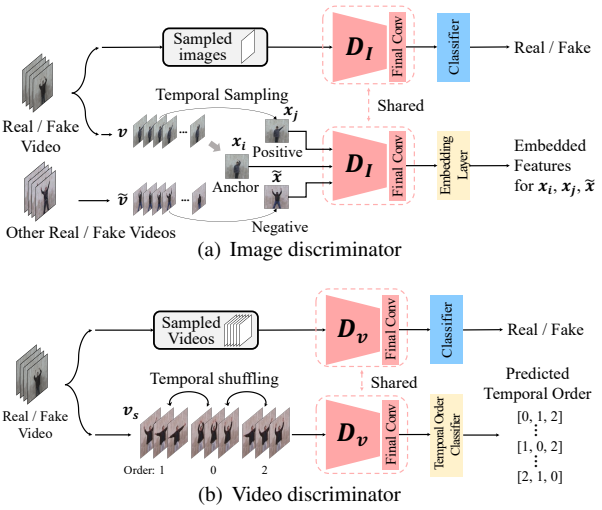


Figure 3. **Detailed illustrations of discriminators.** Image discriminator classifies image as real or fake and sovles appearance contrastive learning. Video discriminator classifies video as real or fake and matches temporal order of shuffled video.

extremely short temporal semantics.

Now, we formulate the temporal structure puzzle task. The task is to classify the correct permutation over a set of possible shuffling permutations $S = \{s_1, s_2, ..., s_P\}$. Let us denote a video v as a sequence of clips $\{c_1, c_2, ..., c_N\}$, where c_i is the i^{th} clip, and we assume the number of frames of each clip is the same for the sake of simplicity. As mentioned above, the video v is represented as a concatenation of the clips from a real video. We further shuffle the order of these clips of the video v based on a permutation s_i. A permutation can be represented as an order of clips $\langle c_{k_1}, c_{k_2}, ..., c_{k_N} \rangle$, where $k_1, k_2, ..., k_N \in \{1, 2, ..., N\}$ and $k_1 \neq k_2 \neq ... \neq k_N$. For a given video v_{s_i}, which is shuffled on the permutation s_i, the encoder outputs a probability distribution $p^{v_{s_i}}$ over S. We can denote the probability of the jth permutation s_j as $p_j^{v_{s_i}}$. Then temporal structure loss L_T for v_{s_i} is defined as:

$$L_T = -\sum_{j=1}^{P} y_j \log(p_j^{v_{s_i}}), \tag{3}$$

where y_j is the ground-truth probability of the video belonging to the j^{th} permutation.

Detailed process of discriminators is illustrated in Fig. 3. Similar to the appearance contrastive learning which is applied to the image discriminator, the temporal structure puzzle with the video-level self-supervision task is assigned to the video discriminator D_V. Moreover, we give the feedback of the discriminator to the generator as collaborative learning, which is more specifically elaborated in the next subsection.

3.3. Collaborative Learning for Generator

The discriminator which has representations of appearance and motion may be sufficient to train the generator capable of producing realistic videos. However, when feedback is given to the generator through backpropagation from the self-supervision objectives of the discriminators, the generator is further encouraged to synthesize videos satisfying the two constraints; consistent appearance and coherent motion. This collaborative learning accelerates to meet our constraints with the synergy between the discriminators and the generator, which is a different way from adversarial learning.

To this end, we follow the training scheme of SS-GAN [9]. Basically, the generator and the discriminators learn to solve the true vs. fake prediction task in an adversarial way. In contrast, for self-supervision tasks, they are trained in a collaborative way. In other words, the generator is guided to produce videos that the discriminators can easily solve the pretext tasks. The discriminators are basically good at solving the tasks with real videos since they are trained from the given tasks, and thus the generator is encouraged to imitate the real data. In this condition, advantages of the proposed self-supervision tasks are intuitively explainable. For instance, appearance contrastive loss encourages the generator to synthesize frames that maintain consistency of contents like the way we defined in Sec. 3.1. On the other hand, temporal structure loss forces the generator to yield natural motions aligned with the global context in videos.

3.4. Full Objectives

As mentioned in previous sections, we apply collaborative self-supervision loss and adversarial loss to both the generator G and discriminators D_I and D_V. Specifically, two discriminators, D_V and D_I, have videos and sampled images as inputs, respectively. When regarding each D_I and D_V as encoder (or classifier) in objectives L_A and L_T, total loss functions are defined as follows:

$$L_D = V_I(G, D_I) + V_V(G, D_V) + \lambda_D(L_A^R + L_T^R), \tag{4}$$

$$L_G = -V_I(G, D_I) - V_V(G, D_V) + \lambda_G(L_A^G + L_T^G), \tag{5}$$

$$\begin{aligned} V_I(G, D_I) = \mathbb{E}_{x \sim P_I(x)} \log\left(D_I(x)\right) \\ + \mathbb{E}_{x \sim P_{\hat{I}}(x)} \log\left(1 - D_I(x)\right), \end{aligned} \tag{6}$$

$$\begin{aligned} V_V(G, D_V) = \mathbb{E}_{v \sim P_V(v)} \log\left(D_V(v)\right) \\ + \mathbb{E}_{v \sim P_{\hat{V}}(v)} \log\left(1 - D_V(v)\right), \end{aligned} \tag{7}$$

where $V(G, D)$ is typical GANs loss proposed in Goodfellow et al. [16], and λ_D and λ_G are hyperparameters that balance two self-supervision loss functions. $L_{A,T}^R$ and $L_{A,T}^G$ are proposed self-supervised loss with real and generated input data. Also, P_V and P_I represent distributions of real

videos and images, while $P_{\hat{V}}$ and $P_{\hat{I}}$ are fake distributions of those.

4. Experiments

Our framework is generic for video GANs with the dual discriminator architecture. Therefore, we evaluate our framework on top of two state-of-the-art video GANs, MoCoGAN [35] and G^{3}AN [38]. Note that, MoCoGAN and G^{3}AN use different configurations for training, which can lead to performance variation. Specifically, MoCoGAN uses a single frame rate, while G^{3}AN deploys diverse frame rates for data augmentation. We directly follow the settings of them to validate the merits of our framework clearly. Besides, we perform all the experiments on unconditional video generation without exploiting any high-level human supervision. Our framework is denoted as **SVGAN** in the tables.

4.1. Datasets

Facial expression datasets We conduct experiments on two facial expression datasets, MUG Facial Expression dataset [3] and UvA-NEMO Smile dataset [11]. MUG contains 1254 videos with 86 subjects, and we use 54 of them, which are available online. UvA-NEMO consists of 1240 smiling videos of 400 identities. For face datasets, we crop the facial regions and resize videos into a spatial resolution of 64×64.

Action datasets As for human actions, we evaluate our model on the Weizmann Action [6] and UCF101 [33] public benchmarks. Weizmann contains 81 videos of 9 people, and UCF101 has 13220 videos with 101 action categories. For both datasets, videos are spatially scaled to 64×64 as did in [35, 38, 30].

4.2. Implementation Details

For the generator architecture, we use default settings as provided by authors except for a few additional layers of self-supervision task head. Therefore, there is only a marginal computational overhead compared to the baseline model. We deploy the ADAM [24] optimizer with a learning rate of 0.0002, 0.5 for β_1, and 0.999 for β_2. We use the batch size as 32 for all experiments. Besides, the length of the generated video is 16 as did [35, 38]. For weighting parameters of self-supervision tasks, we set $\lambda_G = 0.1$ for all experiments. We set $\lambda_D = 1.0$ for MoCoGAN for all benchmarks. Regarding to G^{3}AN, we use $\lambda_D = 0.5$ for all of the datasets except for UCF101, in which we deploy $\lambda_D = 0.1$.

Appearance contrastive learning For appearance contrastive loss, we additionally use data augmentation utilized in [8] when training the discriminator for feature learning. Specifically, two temporally distinct frames transformed with data augmentation are sampled for each video. Note that, we keep the sampled frames distant enough from each other. Also, we do not use any additional methods of preserving a large set of negative samples such as a memory bank [40] or a momentum encoder [17], to maintain the same computational cost with the baselines [35, 38]. We use a temperature parameter $\tau = 0.07$.

	Weizmann	MUG	UvA	UCF101	
	FVD ↓	FVD ↓	FVD ↓	FVD ↓	IS ↑
MoCoGAN	194.34	102.20	46.20	869.41	1.46
SVGAN	**189.65**	**90.79**	**40.56**	**822.48**	**1.48**
G^{3}AN	117.69	89.73	56.97	687.67	1.93
SVGAN	**105.51**	**67.62**	**39.62**	**643.55**	**1.96**

Table 1. Quantitative results on action and facial expression datasets with respect to FVD and IS. Each SVGAN denotes a self-supervised approach applied to each above baseline.

Temporal structure puzzle For temporal structure puzzle, we divide 16 consequent frames into 4 segments of 4 frames, and concatenate them again with the random order. Besides, we use all possible permutations of 4! for training. We deploy random crop to prevent a shortcut to predict the order as did in [43] for UCF101 [33] because they are more likely to have simple cues from the background.

4.3. Evaluation Metric

Recently, Frechet Video Distance (FVD) [36] is introduced to evaluate the visual quality of generated videos, which is a temporal counterpart of Frèchet Inception Distance (FID) [18]. Specifically, FVD computes Frèchet Distance between real-world video distribution P_R and fake video distribution P_G under the condition that P_R and P_G are multivariate Gaussian. FVD is defined as follows:

$$|\mu_R - \mu_G|^2 + Tr(\Sigma_R + \Sigma_G - 2(\Sigma_R\Sigma_G)^{\frac{1}{2}}), \quad (8)$$

where μ_R and Σ_R are the mean and covariance matrix of P_R and μ_G and Σ_G are those of P_G.

We also deploy Inception Score (IS) [32] as another metric. IS mainly evaluates the diversity of the generated videos. It can be obtained as follows:

$$\exp(\mathbb{E}_x \mathrm{KL}(p(y|x)||p(y)), \quad (9)$$

where $p(y|x)$ and $p(y)$ are conditional and marginal class distributions, respectively, and KL means Kullback-Leibler divergence.

	UvA	UCF101	
	FVD ↓	FVD ↓	IS ↑
Baseline	46.20	869.41	1.46
+ Appearance contrastive L_A	41.56	853.89	1.49
+ Temporal structure L_T	43.40	856.76	**1.52**
Full objectives (L_A + L_T)	**40.56**	**822.48**	1.48

Table 2. Ablation study on the proposed self-supervision loss functions with MoCoGAN in FVD and IS. Note that, UvA and UCF101 datasets have higher diversities compared to the other ones.

4.4. Quantitative Results

Comparison with the state-of-the-art We firstly evaluate our framework on top of the state-of-the-art methods to validate the performance gain. As shown in Table 1, our method consistently achieves better FVD scores as well as higher IS ones on all of the benchmarks by remarkable margins. These results confirm that SVGAN generates more realistic samples in terms of both visual quality and diversity. In other words, the synthesized videos retain more consistent appearance with natural movements. Note that, these benefits of SVGAN become more clear with a larger model capacity of G^{3}AN [38]. More importantly, we find these performance improvements consistent with respect to the architecture of video GANs and their training strategies. Thus, it shows that SVGAN can be widely deployed with video GANs to enhance their quality of generation.

Moreover, one interesting point is that the performance is improved even in the extremely small-scale dataset such as Weizmann, which has only 81 videos of 9 identities. It validates that our framework can work with datasets in various scales, while self-supervised approaches are usually utilized in large-scale datasets for feature learning.

Ablation study on pretext tasks To figure out the effects of each self-supervision task, we conduct ablation studies of loss functions on UvA and UCF101 datasets which have higher diversities compared to the other datasets. To this end, we measure FVD scores when removing the individual objectives from our framework. Note that, we use MoCoGAN as the baseline for all ablation studies and set $\lambda_G = 0.2$ when applying a single self-supervision task.

In Table 2, we observe that FVD is gradually improved when we assign more self-supervision tasks. Furthermore, the lowest FVD score achieved with two pretext tasks emphasizes that they are complementary. Since appearance contrastive learning forces the consistency of entire frames by maximizing their similarity and temporal structure puzzle prompts the dynamics of motions, there are synergistic effects in video generation performance when both tasks are jointly deployed.

Pretext task (L_A)	UvA	UCF101	
	FVD ↓	FVD ↓	IS ↑
-	46.20	869.41	1.46
Rotation	46.05	859.30	1.38
Ours	**40.56**	**822.48**	**1.48**

Table 3. Comparison with appearance contrastive learning against rotation prediction task based on MoCoGAN in terms of FVD and IS.

Comparison with rotation prediction task For validating the effectiveness of proposed self-supervision task on video generation, we compare our method to commonly used pretext task, rotation prediction [15] as did in Self-supervised GANs[9]. Since rotation prediction is an image-level task, we replace proposed appearance contrastive learning L_A with it.

As reported in Table 3, our method outperforms the rotation pretext task with respect to FVD in all tested datasets. Interestingly, the rotation prediction task shows a marginal improvement on the UvA dataset of facial expression. This is a similar observation with [9] that rotation pretext task may not enhance the generation quality in the human face dataset, since it is hard to learn semantic information by predicting rotation for face images. Differently, we found that the proposed appearance contrastive loss effectively works on the UvA dataset. Furthermore, our method shows a large enhancement in terms of FVD in UCF101 dataset while the performance of the rotation task becomes worse in IS than that of baseline in spite of a small improvement of FVD.

Analysis on temporal sampling in L_A In appearance contrastive learning, we sample two frames at different times as transformations. Here, we analyze the effect of this temporal sampling for contrasitve learning. First, we sample a single frame from a video instead of sampling different ones over time. Then we apply random spatial transformations for the identical image to produce two different views as in [8]. Afterward, we compute the contrastive loss described in Sec. 3.1 to train the video GANs.

Table 4 shows the results of analysis on temporal sampling. In the table, we observe the clear drop in performance of our framework without temporal sampling. This result confirms that temporal sampling with contrastive learning brings more consistency of appearance for realistic videos.

Analysis on input types in L_T We also conduct experiments to validate the feasibility of proposed temporal structure puzzle. To this end, we extract features of each clip of 4 frames independently by the discriminator as did in [43]. Then we concatenate these features in a pairwise way and give them to the classifier. We expect this clip-level input

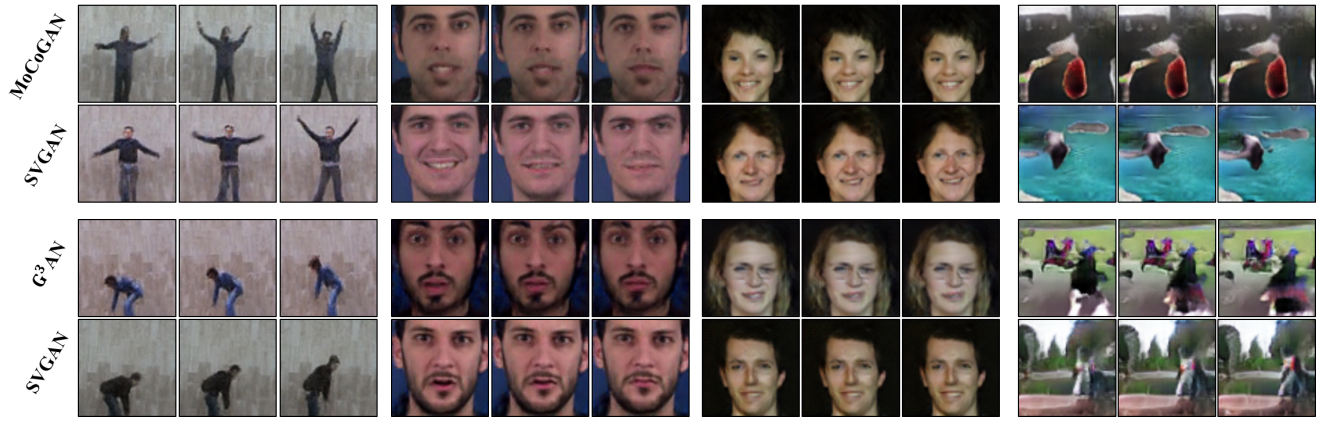


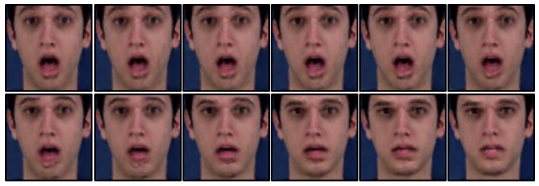
Figure 4. Qualitative comparison with state-of-the-arts on Weizmann (left-most), MUG (left-center), UvA-NEMO (right-center) and UCF101 (right-most). Each "SVGAN" denotes a self-supervised approach applied to each above baseline model.

	Temporal sampling	UvA	UCF101
MoCoGAN	-	46.20	869.41
Ours (L_A)	-	42.54	921.73
Ours (L_A)	✓	**41.56**	**853.89**

Table 4. Analysis on temporal sampling in appearance contrastive learning. Temporal sampling indicates sampling frames with distinct time step for contrastive learning. Reported scores are FVDs.

	Input types	UvA	UCF101
MoCoGAN	-	46.20	869.41
Ours (L_T)	Clip-level	55.44	1004.68
Ours (L_T)	Video-level	**43.40**	**856.76**

Table 5. Analysis on input types in temporal structure puzzle. "Video-level" and "Clip-level" denote the input types of ours and the original one [43], respectively. Reported scores are FVDs.



(a) MUG

(b) Weizmann

Figure 5. Generated examples of SVGAN based on G^{3}AN in MUG and Weizmann datasets. Each sequence is progressed from left-top to right-bottom.

degrades the performance since this way provides a smaller temporal receptive field compared to our video-level input.

Table 5 shows that video-level input provides better FVD scores. Note that, when we forward each clip independently, the performance gets even worse than that of the baseline. Therefore, we reach conclusion that forwarding clips separately is not proper to generate short videos since it makes the discriminator only extract semantics of an extremely short clip.

4.5. Qualitative Results

We further analyze unconditional video generation performance in a qualitative way compared to the previous state-of-the-art methods [35, 38]. As reported in Fig. 4, the proposed SVGAN generates more realistic videos compared to the baseline models in the aspect of the visual quality. Besides, we present videos sampled from SVGAN based on G^{3}AN with 12 frames in Fig. 5.

5. Conclusion

In this paper, we have proposed Self-supervised Video GANs with two pretext tasks; appearance contrastive learning and temporal structure puzzle. Each pretext task explicitly encourages the discriminator to learn representations of appearance which is time-invariant information, and temporal structure which is flow of the motions. Furthermore, we distill the knowledge about appearance and motion to the generator for synthesizing appearance consistent and temporally coherent videos. Extensive experiments have validated the improvements of generation quality over the state-of-the-art video GANs regardless of the architecture of generators.

Acknowledgement. This work was supported in part by Samsung Research Funding & Incubation Center for Future Technology (SRFC-IT1901-01), and the IITP funded by the Korea Government MSIT under Grants 2020-0-00973, 2019-0-00421, and 2020-0-01821.

References

[1] Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Towards high resolution video generation with progressive growing of sliced wasserstein gans. *CoRR*, abs/1810.02419, 2018.

[2] Unaiza Ahsan, Rishi Madhok, and Irfan Essa. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 179–189. IEEE, 2019.

[3] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. The mug facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*, pages 1–4. IEEE, 2010.

[4] Martín Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[5] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 2017.

[6] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1395–1402. IEEE, 2005.

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of Machine Learning and Systems 2020*, pages 10719–10729. 2020.

[9] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12154–12163, 2019.

[10] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv*, pages arXiv–1907, 2019.

[11] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *European Conference on Computer Vision*, pages 525–538. Springer, 2012.

[12] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[13] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2015.

[14] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3636–3645, 2017.

[15] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

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

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[18] 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 *Advances in neural information processing systems*, pages 6626–6637, 2017.

[19] Rui Huang, Wenju Xu, Teng-Yok Lee, Anoop Cherian, Ye Wang, and Tim Marks. Fx-gan: Self-supervised gan learning via feature exchange. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 3194–3202, 2020.

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[23] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8545–8552, 2019.

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[25] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017.

[26] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2813–2821. IEEE Computer Society, 2017.

[27] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016.

[28] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[29] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[30] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE international conference on computer vision*, pages 2830–2839, 2017.

[31] Masaki Saito and Shunta Saito. Tganv2: Efficient training of large models for video generation with multiple subsampling layers. *CoRR*, abs/1811.09245, 2018.

[32] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2226–2234, 2016.

[33] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CRCV-TR-12-01*, November, 2012.

[34] Ngoc-Trung Tran, Viet-Hung Tran, Bao-Ngoc Nguyen, Linxiao Yang, et al. Self-supervised gan: Analysis and improvement with multi-class minimax game. In *Advances in Neural Information Processing Systems*, pages 13253–13264, 2019.

[35] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018.

[36] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation. In *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019*. OpenReview.net, 2019.

[37] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances in neural information processing systems*, pages 613–621, 2016.

[38] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3an: Disentangling appearance and motion for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5264–5273, 2020.

[39] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018.

[40] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[41] Jianwen Xie, Ruiqi Gao, Zilong Zheng, Song-Chun Zhu, and Ying Nian Wu. Learning dynamic generator model by alternating back-propagation through time. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5498–5507, 2019.

[42] Jianwen Xie, Ruiqi Gao, Zilong Zheng, Song-Chun Zhu, and Ying Nian Wu. Motion-based generator model: Unsupervised disentanglement of appearance, trackable and intrackable motions in dynamic patterns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12442–12451, 2020.

[43] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019.

[44] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363. PMLR, 2019.