Point-to-Point Video Generation

Tsun-Hsuan Wang*, Yen-Chi Cheng*, Chieh Hubert Lin, Hwann-Tzong Chen, Min Sun
National Tsing Hua University
{johnsonwang0810, charlescheng0117, hubert052702}@gmail.com
htchen@cs.nthu.edu.tw, sunmin@ee.nthu.edu.tw

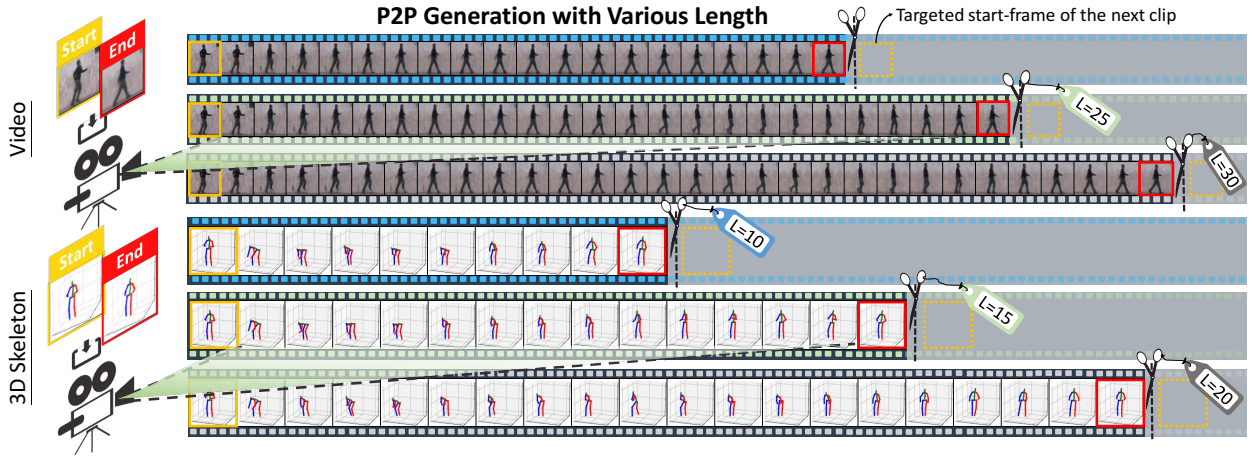

Figure 1: **Point-to-Point (P2P) Video Generation.** Given a pair of (orange) start- and (red) end-frames in the video and 3D skeleton domains, our method generates videos with smooth transitional frames of various lengths. The superb controllability of p2p generation naturally facilitates the modern video editing process.

Abstract

*While image synthesis achieves tremendous breakthroughs (*e.g.*, generating realistic faces), video generation is less explored and harder to control, which limits its applications in the real world. For instance, video editing requires temporal coherence across multiple clips and thus poses both start and end constraints within a video sequence. We introduce* ***point-to-point video generation*** *that controls the generation process with two control points: the targeted start- and end-frames. The task is challenging since the model not only generates a smooth transition of frames, but also plans ahead to ensure that the generated end-frame conforms to the targeted end-frame for videos of various lengths. We propose to maximize the modified variational lower bound of conditional data likelihood under a skip-frame training strategy. Our model can generate end-frame-consistent sequences without loss of quality and diversity. We evaluate our method through extensive experiments on Stochastic Moving MNIST, Weizmann Action, Human3.6M, and BAIR Robot Pushing under a series of scenarios. The qualitative results showcase the effectiveness and merits of point-to-point generation.*

*indicates equal contribution

1. Introduction

The significant advancements in deep generative models bring impressive results in a wide range of domains such as image synthesis, text generation, and video prediction. Despite the huge success, unconstrained generation is still a few steps away from practical applications since it lacks intuitive and handy mechanisms to incorporate human manipulation into the generation process. In view of this incapability, conditional and controllable generative models have received an increasing amount of attention. Most existing work achieves controllability by conditioning the generation on the attribute, text, user inputs, or scene graph [17, 39, 43, 45]. However, regardless of the considerable progress in still image generation, controllable video generation is yet to be well explored.

Typically, humans create a video through breaking down the entire story into separate scenes, taking shots for each scene individually, and finally merging every piece of footage to form the final edit. This requires a smooth transition across not only frames but also different video clips, posing constraints on both start- and end-frames within a video sequence so as to align with the preceding and subsequent context. We introduce *point-to-point video gener-*

ation (p2p generation) that controls the generation process with two control points—the targeted start- and end-frames. Enforcing consistency on the two control points allows us to regularize the context of the generated intermediate frames, and it also provides a straightforward strategy for merging multiple videos. Moreover, in comparison with the standard video generation setting [32], which requires a consecutive sequence for initial frames, p2p generation only needs a pair of individual frames. Such a setting is more accessible in real-world scenarios, *e.g.*, generating videos from images with similar content crawled on the Internet. Finally, p2p generation is preferable to attribute-based methods for more sophisticated video generation tasks that involve hard-to-describe attributes. Attribute-based methods heavily depend on the available attributes provided in the datasets, whereas p2p generation can avoid the burden of collecting and annotating meticulous attributes.

Point-to-point generation has two major challenges: *i*) The control point consistency (CPC) should be achieved without the sacrifice of generation quality and diversity. *ii*) The generation with various lengths should all satisfy the control point consistency. Following the recent progress in video generation and future frame prediction, we introduce a global descriptor, which carries information about the targeted end-frame, and a time counter, which provides temporal hints for dynamic length generation to form a conditional variational encoder (CVAE [31]). In addition, to balance among generation quality, diversity, and CPC, we propose to maximize the modified variational lower bound of conditional data likelihood. Besides, we inject an alignment loss to ensure that the latent space in the encoder and decoder aligns with each other. We further present the skip-frame training strategy to reinforce our model to be more time-counter-aware. Our model adjusts its generation procedure accordingly, and thus achieves better CPC. Extensive experiments are conducted on Stochastic Moving MNIST (or SM-MNIST) [32, 3], Weizmann Human Action [8], Human3.6M (3D skeleton data) [13], and BAIR Robut Pushing [5] to evaluate the effectiveness of the proposed method. A series of qualitative results further highlight the merits of p2p generation and the capability of our model.

2. Related Work

Our problem is most related to video generation [29, 33, 35] and the controllability of video generation [9, 10, 12, 20, 24, 41]. It also has a connection with video interpolation. We briefly review these topics in this section.

Video Generation. Many approaches use GANs [1, 33, 35] or adversarial loss during training for generating videos [1, 21, 23, 25, 30, 35, 36]. Vondrick *et al*. [35] use a generator with two pathways to predict the foreground and background, and a discriminator to distinguish a video as real or fake. On the other hand, it can be tackled by learning how to transform observed frames to synthesize the future frames [6, 15, 22, 36, 40]. Furthermore, strategies based on decomposing a video into a static part that can be shared along (*i.e*. content) and the varying part (*i.e*. motion) are also proposed to describe the video dynamics [4, 11, 33, 34, 38]. Denton *et al*. [4] encode motion and content into different subspaces and use an adversarial loss on the motion encoder to achieve disentanglement.

Several methods rely on VAE [18] to capture the uncertain nature in videos [2, 3, 7, 10, 19, 21, 37, 42]. Babaeizadeh *et al*. [2] extend [6] with variational inference framework such that their model can predict multiple frames of plausible futures on real-world data. Jayaraman *et al*. [14] predict the most certain frame first and break down the original problem such that the predictor can complete the semantic sub-goals coherently.

While the methods mentioned above achieve good results on video prediction, the generation process is often uncontrollable and hence leads to unconstrained outputs. In order to preserve the ability of generating diversified outputs while achieving control point consistency, we manage to build upon VAE for point-to-point video generation.

Video Interpolation (VI). The problem setting of p2p generation has connection to VI [16, 26, 27, 28, 46] but with essential differences. VI aims to increase the frame-rate of a video. Thus both the number of inserted frames and the time interval of interpolation are assumed to be small, whereas p2p generation involves a much longer-term synthesis of in-between frames, posing a different challenge. Besides, VI methods typically are deterministic (*i.e*., producing only one interpolated result). Instead, our work is akin to video generation where the synthesized frames are required to be both *temporally coherent* and *diverse* in context. Finally, automatic looping (*i.e*., generating a looping video given identical start-frame and end-frame) can be accomplished by p2p generation but not by VI (see Sec. 4.8 for detailed analysis).

Controllability on Video Generation. Several methods are proposed to guide the video generation process. Hu *et al*. [12] use an image and a motion stroke to synthesize the video. Hao *et al*. [9] condition on the start frame and a trajectory provided by user to steer the motion and appearance for the next frames. He *et al*. [10] propose an attribute-based approach for transient control by exploiting the attributes (*e.g*., identity, action) in the dataset. Text or language features can also be used as the instruction for controls [20, 24, 41]. Although the existing methods all provide freedom for controlling the generation, they come with some limitations. Conditioning on language would suffer from its ambiguous nature, which does not allow precise control [24]. Attribute control depends on the data labels and is not available in an unsupervised setting. User provided input is intuitive but requires annotations during training. Instead, our method *i*) only conditions on the target

frame which can be acquired without any cost, *ii*) can incorporate detailed descriptions of the control points (*e.g.*, the precise look and action of a person, or joints of a skeleton) to provide exact control, and *iii*) can be trained in a fully unsupervised fashion. The advantage over previous methods in having the controllability of start- and target-frames motivates our point-to-point generation.

3. Methodology

Given a pair of control points (the targeted start- and end-frames $\{x_1, x_T\}$) and the generation length T, we aim to generate a sequence $\hat{x}_{1:T}$ with the specified length such that their start- and end-frames $\{\hat{x}_1, \hat{x}_T\}$ are consistent with the control points. To maintain quality and diversity in p2p generation, we present a conditional video generation model (Sec. 3.2) that maximizes the modified variational lower bound (Sec. 3.3). To further improve CPC under various lengths, we propose a novel skip-frame training strategy (Sec. 3.4) and a latent alignment loss (Sec. 3.5).

3.1. VAE and Video Generation

Variational Autoencoder (VAE) leverages a simple prior $p_\theta(\mathbf{z})$ (*e.g.*, Gaussian) and a complex likelihood $p_\theta(\mathbf{x}|\mathbf{z})$ (*e.g.*, a neural network) on latent variable $\mathbf{z}$ to maximize the data likelihood $p_\theta(\mathbf{x})$, where $\mathbf{x} = [x_1, x_2, \dots, x_T]$. A variational neural network $q_\phi(\mathbf{z}|\mathbf{x})$ is introduced to approximate the intractable latent posterior $p_\theta(\mathbf{z}|\mathbf{x})$, allowing joint optimization over θ and ϕ,

$$\begin{aligned}\log p_\theta(\mathbf{x}) &= \log \int_{\mathbf{z}} p_\theta(\mathbf{x}|\mathbf{z}) p(\mathbf{z})\, d\mathbf{z} \\ &\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}) - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))\,.\end{aligned} \tag{1}$$

The intuition behind the inequality is to reconstruct data $\mathbf{x}$ with latent variable $\mathbf{z}$ sampled from the posterior $q_\phi(\mathbf{z}|\mathbf{x})$, simultaneously minimizing the KL-divergence between the prior $p(\mathbf{z})$ and posterior $q_\phi(\mathbf{z}|\mathbf{x})$.

Video generation commonly adopts VAE framework accompanied by a recurrent model (*e.g.*, LSTM), where the VAE handles generation process and the recurrent model captures the dynamic dependencies in sequential generation. However, in VAE, the simple choice for prior $p(\mathbf{z})$ such as a fixed Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is confined to drawing samples randomly at each timestep regardless of temporal dependencies across frames. Accordingly, existing works resort to parameterizing the prior with a learnable function $p_\psi(z_t|x_{1:t-1})$ conditioned on previous frames $x_{1:t-1}$. The variational lower bound throughout the entire sequence is

$$\begin{aligned}\mathcal{L}_{\theta,\phi,\psi}(x_{1:T}) = \sum_{t=1}^{T} \big[\, &\mathbb{E}_{q_\phi(z_{1:t}|x_{1:t})} \log p_\theta(x_t|x_{1:t-1}, z_{1:t}) \\ &- D_{\mathrm{KL}}(q_\phi(z_{1:t}|x_{1:t})||p_\psi(z_t|x_{1:t-1}))\,\big]\,.\end{aligned} \tag{2}$$

In comparison with a standard VAE, the former term describes the reconstruction sampled from the posterior $q_\phi(z_{1:t}|x_{1:t})$ conditioned on data up to the *current frame*. The latter term ensures that the prior $p_\psi(z_t|x_{1:t-1})$ conditioned on data up to the *previous frame* does not deviate from the posterior. Meanwhile, it also serves as a regularization on the learning of posterior. In this work, we inherit and modify the network architecture of [3], and adapt $\mathcal{L}_{\theta,\phi,\psi}(x_{1:T})$ for p2p generation.

3.2. Global Descriptor and Time Counter

For a deep network to achieve p2p generation under various lengths, *i*) the model should be aware of the information of control points and *ii*) the model should be able to perceive time lapse and generate the targeted end-frame at the designated timestep. While the targeted start-frame is already fed as an initial frame, we adopt a straightforward strategy to incorporate the control points into the model at every timestep by feeding features encoded from the targeted end-frame h_T to our model. Besides, to enforce our model to be aware of when to generate the targeted end-frame given the generation length T, we introduce a time counter $\tau_t \in [0, 1]$, where $\tau_t = 0.0$ indicates the beginning of the sequence and $\tau_t = 1.0$ indicates reaching the targeted end-frame. As shown in Fig. 2(a), q_ϕ and p_ψ are modeled by a shared-weight encoder and two different LSTMs, and p_θ is modeled by the third LSTM along with a decoder to map latent vectors to image space. The inference process during training at timestep t is shown as

$$\begin{aligned} h_T &= Enc(x_T), \quad \tau_t = t/T, \\ \mu_\phi^t, \sigma_\phi^t &= \mathrm{LSTM}_\phi(h_t, h_T, \tau_t),\ h_t = Enc(x_t), \\ z_t &\sim \mathcal{N}(\mu_\phi^t, \sigma_\phi^t), \\ g_t &= \mathrm{LSTM}_\theta(h_{t-1}, z_t, \tau_t),\ h_{t-1} = Enc(x_{t-1}), \\ \hat{x}_t &= Dec(g_t)\,. \end{aligned} \tag{3}$$

During test time, as we have no access to current x_t, the latent variable z_t is sampled from the prior distribution p_ψ,

$$\begin{aligned} \mu_\psi^t, \sigma_\psi^t &= \mathrm{LSTM}_\psi(h_{t-1}, h_T, \tau_t), \\ z_t &\sim \mathcal{N}(\mu_\psi^t, \sigma_\psi^t)\,. \end{aligned} \tag{4}$$

Recall that the KL divergence in (2) enforces the alignment between q_ϕ and p_ψ, allowing p_ψ to serve as a proxy of q_ϕ at test time. Besides, by introducing the global descriptor h_T and time counter τ_t, (2) is extended to a variational lower bound of conditional data likelihood $\mathcal{L}_{\theta,\phi,\psi}(x_{1:T}|\mathbf{c})$, where $\mathbf{c}$ is the conditioning on the targeted end-frame and time counter. In addition, we further propose a latent space alignment loss within h_t and g_t to mitigate the mismatch between the encoding and the decoding process, as shown in (6).

3.3. Control Point Consistency on Prior

Although introducing the time counter and the global descriptor of control points provides the model with capability of achieving CPC, we are not able to further reinforce the generated end-frame to conform to the targeted

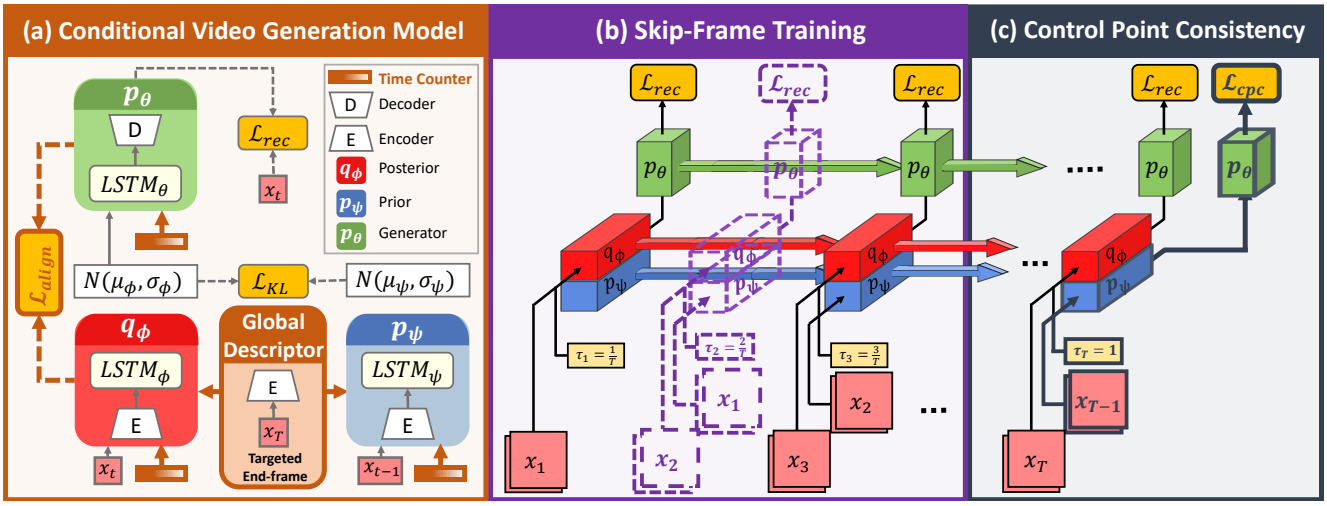

Figure 2: **An overview of the novel components of p2p generation.** (a) Our model is a VAE consisting of posterior q_ϕ, prior p_ψ, and generator p_θ (all with an LSTM for temporal coherency). We use KL-divergence to encourage p_ψ to be similar to q_ϕ. To control the generation, we encode the targeted end-frame x_T into a global descriptor. Both q_ϕ and p_ψ are computed by considering not only the input frame (x_t or x_{t-1}), but also the "global descriptor" and "time counter". We further use the "alignment loss" to align the encoder and decoder latent space to reinforce the control point consistency. (b) Our skip-frame training has a probability to skip the input frame in each timestamp where the input will be ignored completely and the hidden state will not be propagated at all **(see the dashed line)**. (c) The control point consistency is achieved by posing CPC loss on p_ψ without harming the reconstruction objective of q_ϕ **(highlighted in bold)**.

end-frame. While the conditioning happens to be a part of the reconstruction objective, naively increasing the weight α_{cpc} at timestep T in the reconstruction term of (2), *i.e.*, $\sum_{t=1}^{T-1} \mathbb{E}_{q_\phi} \log p_\theta(x_t) + \alpha_{\text{cpc}} \mathbb{E}_{q_\phi} \log p_\theta(x_T)$, results in unstable training behavior and degradation of generation quality and diversity. To tackle this problem, we propose to single out the CPC from the reconstruction loss on the posterior and pose it on the prior. The modified lower bound of conditional data likelihood with a learnable prior p_ψ is

$$\begin{aligned}
&\mathcal{L}^{\text{p2p}}_{\theta,\phi,\psi}(x_{1:T}|\mathbf{c}) = \\
&\sum_{t=1}^{T} \big[\mathbb{E}_{q_\phi(z_{1:t}|x_{1:t},\mathbf{c})} \log p_\theta(x_t|x_{1:t-1}, z_{1:t}, \mathbf{c}) \\
&\quad - D_{\text{KL}}(q_\phi(z_{1:t}|x_{1:t}, \mathbf{c})||p_\psi(z_t|x_{1:t-1}, \mathbf{c}))\big] \\
&\quad + \mathbb{E}_{p_\psi(z_T|x_{1:T-1},\mathbf{c})} \log p_\theta(x_T|x_{1:T-1}, z_{1:T}, \mathbf{c}) \,.
\end{aligned} \tag{5}$$

While the first two terms are the same as the bound of conditional VAE (CVAE), the third term of the above formulation benefits a more flexible tuning on the behavior of the additionally-introduced condition without degrading the maximum likelihood estimate in the first term.

3.4. Skip-Frame Training

A well-functioning p2p generation model should be aware of the time counter in order to achieve CPC under various lengths. However, most video datasets have a fixed frame rate. As a result, the model may exploit the fixed frequency across frames and ignore the time counter. We introduce skip-frame training to further enhance the model to be more aware of the time counter. Basically, we randomly drop frames while computing the reconstruction loss and KL divergence (the first two terms in (5)). The LSTMs are hence forced to take time counter into consideration so as to handle the random skipping in the recurrence. Such adaption in the maximum likelihood estimate of posterior q_ϕ further incorporates CPC into the learning of posterior.

3.5. Final Objective

To summarize, our final objective that maximizes the modified variational lower bound of conditional data likelihood under a skip-frame training strategy is

$$\begin{aligned}
&\mathcal{L}^{\text{full}}_{\theta,\phi,\psi}(x_{1:T}|\mathbf{c}) = \\
&\sum_{t=1}^{T} M_t \Big[\mathbb{E}_{q_\phi(z_{1:t}|x_{1:t},\mathbf{c})} \log p_\theta(x_t|x_{1:t-1}, z_{1:t}, \mathbf{c}) \\
&\qquad - \beta D_{\text{KL}}(q_\phi(z_{1:t}|x_{1:t}, \mathbf{c})||p_\psi(z_t|x_{1:t-1}, \mathbf{c})) \\
&\qquad - \alpha_{\text{align}}||h_t - g_t||_2\Big] \\
&+ \alpha_{\text{cpc}} \mathbb{E}_{p_\psi(z_T|x_{1:t-1},\mathbf{c})} \log p_\theta(x_T|x_{1:T-1}, z_{1:T}, \mathbf{c}) \,,
\end{aligned} \tag{6}$$

where $M_t \sim \text{Bernoulli}(1 - p_{\text{skip}})$, $M_T = 1$. β, α_{cpc}, and α_{align} are hyperparameters to balance among KL term, CPC, and latent space alignment. The constant $p_{\text{skip}} \in [0, 1]$ determines the rate of skip-frame training.

4. Experiment

To evaluate the effectiveness of our method, we conduct qualitative and quantitative analysis on four datasets: SM-

Method	SSIM (± indicates 95% confidence interval)				PSNR (± indicates 95% confidence interval)			
	S-Best ↑	S-Div (1E-3) ↑	S-CPC ↑	R-Best ↑	S-Best ↑	S-Div ↑	S-CPC ↑	R-Best ↑
SVG [3]	$0.780^{\pm 0.006}$	$2.349^{\pm 0.076}$	$0.621^{\pm 0.004}$	$0.850^{\pm 0.005}$	$15.774^{\pm 0.161}$	$0.816^{\pm 0.019}$	$12.105^{\pm 0.047}$	$18.001^{\pm 0.201}$
+ C	$0.768^{\pm 0.002}$	$2.482^{\pm 0.048}$	$0.729^{\pm 0.003}$	$0.840^{\pm 0.004}$	$15.373^{\pm 0.049}$	$0.914^{\pm 0.014}$	$14.024^{\pm 0.054}$	$17.751^{\pm 0.094}$
+ C + A	$0.755^{\pm 0.003}$	$2.377^{\pm 0.085}$	$0.735^{\pm 0.005}$	$0.816^{\pm 0.005}$	$15.117^{\pm 0.103}$	$0.804^{\pm 0.014}$	$14.141^{\pm 0.069}$	$16.884^{\pm 0.147}$
Ours	$0.755^{\pm 0.004}$	$2.525^{\pm 0.052}$	$\mathbf{0.769}^{\pm 0.005}$	$0.832^{\pm 0.005}$	$15.265^{\pm 0.079}$	$0.815^{\pm 0.009}$	$\mathbf{15.185}^{\pm 0.096}$	$17.581^{\pm 0.172}$

Table 1. Evaluation on SM-MNIST (+*C*: CPC loss on p_ψ only. +*C*+*A*: CPC loss and Alignment loss. *Ours*: Our full model).

Method	SSIM (± indicates 95% confidence interval)				PSNR (± indicates 95% confidence interval)			
	S-Best ↑	S-Div (1E-3) ↑	S-CPC ↑	R-Best ↑	S-Best ↑	S-Div ↑	S-CPC ↑	R-Best ↑
SVG [3]	$0.819^{\pm 0.008}$	$1.992^{\pm 0.351}$	$0.734^{\pm 0.008}$	$0.819^{\pm 0.009}$	$25.234^{\pm 0.355}$	$1.904^{\pm 0.357}$	$22.236^{\pm 0.242}$	$25.039^{\pm 0.400}$
+ C	$0.814^{\pm 0.005}$	$2.574^{\pm 0.402}$	$0.730^{\pm 0.004}$	$0.808^{\pm 0.006}$	$24.898^{\pm 0.110}$	$2.186^{\pm 0.346}$	$22.028^{\pm 0.084}$	$24.624^{\pm 0.211}$
+ C + A	$0.823^{\pm 0.005}$	$1.225^{\pm 0.178}$	$0.767^{\pm 0.009}$	$0.822^{\pm 0.005}$	$25.092^{\pm 0.186}$	$1.266^{\pm 0.170}$	$22.855^{\pm 0.197}$	$24.848^{\pm 0.145}$
Ours	$0.824^{\pm 0.004}$	$1.106^{\pm 0.078}$	$\mathbf{0.783}^{\pm 0.003}$	$0.842^{\pm 0.006}$	$24.993^{\pm 0.103}$	$1.039^{\pm 0.057}$	$\mathbf{23.334}^{\pm 0.105}$	$25.660^{\pm 0.154}$

Table 2. Evaluation on Weizmann (+*C*: CPC loss on p_ψ only. +*C*+*A*: CPC loss and Alignment loss. *Ours*: Our full model).

Method	S-Best ↓	S-Div ↑	S-CPC ↓	R-Best ↓
SVG [3]	$6.49^{\pm 0.31}$	$0.68^{\pm 0.05}$	$10.83^{\pm 0.90}$	$5.75^{\pm 0.17}$
+ C	$8.25^{\pm 0.65}$	$0.64^{\pm 0.06}$	$12.08^{\pm 0.65}$	$8.97^{\pm 0.53}$
+ C + A	$4.96^{\pm 0.18}$	$0.80^{\pm 0.03}$	$6.66^{\pm 0.82}$	$4.74^{\pm 0.17}$
Ours	$4.46^{\pm 0.35}$	$0.88^{\pm 0.06}$	$\mathbf{0.72}^{\pm 0.06}$	$1.23^{\pm 0.04}$

Table 3. Evaluation on Human3.6M (with MSE).

Method	S-Best ↑	S-Div (1E-3) ↑	S-CPC ↑	R-Best ↑
SVG [3]	$0.845^{\pm .006}$	$0.716^{\pm .166}$	$0.775^{\pm .008}$	$0.926^{\pm .003}$
SV2P [2]	$0.841^{\pm .010}$	$0.186^{\pm .021}$	$0.770^{\pm .009}$	$0.847^{\pm .004}$
Ours	$0.847^{\pm .004}$	$0.664^{\pm .049}$	$\mathbf{0.824}^{\pm .015}$	$0.907^{\pm .006}$

Table 4. Evaluation on BAIR Robot Pushing (with SSIM).

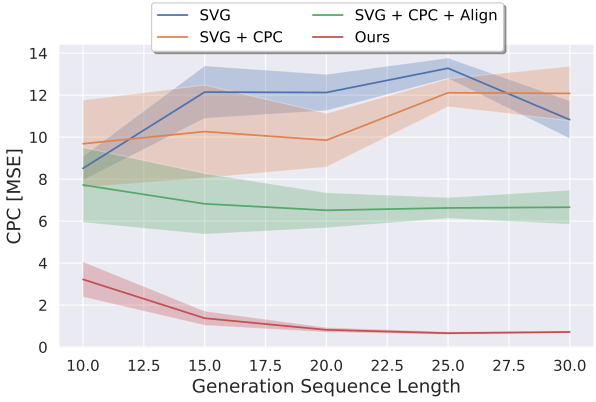

Figure 3: Control Point Consistency (CPC) with various generation lengths shows that our final model (in red) is more stable and can steadily approach the targeted end-frame. (Figures are best viewed in color.)

MNIST [3], Weizmann Action [8], Human3.6M [13], and BAIR Robut Pushing [5] to measure the CPC, quality, and diversity. The following section is organized as follows: we start by describing the datasets in Sec. 4.1 and the evaluation metrics in Sec. 4.2; the quantitative results are shown in 4.3-4.6; the qualitative results are presented in Sec. 4.7; finally the comparisons with VI are discussed in 4.8.

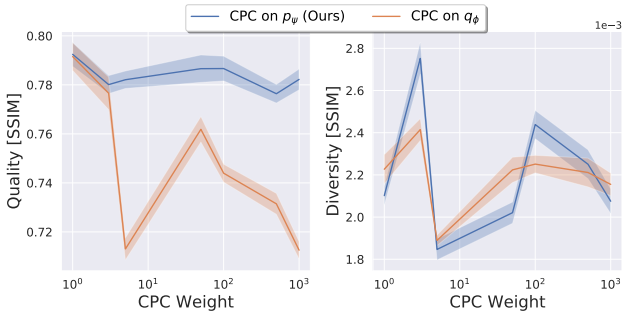

Figure 4: We show the generation quality and diversity with different CPC weights. The results show that posing CPC on prior is more stable than on posterior; the latter is sensitive to large CPC weights and tends to harm the quality.

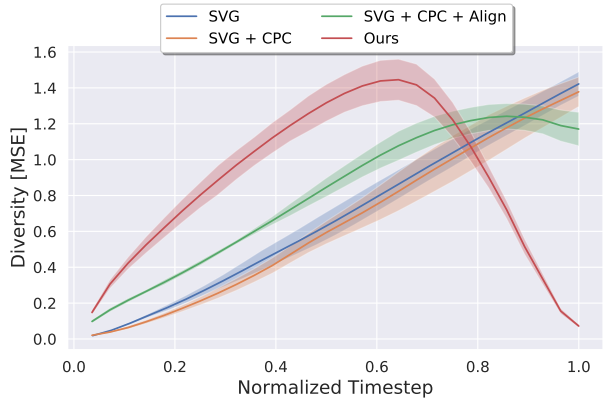

Figure 5: The generation diversity through normalized time-steps shows that *Ours* (in red) presents a desired behavior—diversity increases until the middle of generation, and then converges (decreases) at targeted end-frames.

4.1. Datasets

We evaluate our method on four common testbeds: **Stochastic Moving MNIST** is introduced by [3] (a modified version from [32]). The training sequence is gener-

Figure 6: Generation with various lengths on BAIR Pushing.

ated by sampling one or two digits from the training set of MNIST, and then the trajectory is formed by sampling starting locations within the frame and an initial velocity vector, $(\Delta x, \Delta y) \in [-4, 4] \times [-4, 4]$. The velocity vector will be re-sampled each time the digits reach the border. **Weizmann Action** contains 90 videos of 9 people performing 10 actions. We center-crop each frame and follow the setting in [10] to form the training and test sets. **Human3.6M** is a large-scale dataset with 3.6 million 3D human poses captured by 11 professional actors, providing more than 800 sequences in total. We use normalized 3D skeletons of 17 joints for experiments. Following [42], we use subjects 1, 5, 6, 7, and 8 for training and subjects 9 and 11 for testing. **BAIR Robut Pushing [5]** features a robotic arm moving randomly to push diverse objects. With the large degree of stochasticity and cluttered background, it is widely used for evaluating video prediction/generation.

4.2. Evaluation Metrics

We measure the structural similarity (SSIM) and peak signal-to-noise ratio (PSNR) for SM-MNIST, Weizmann, and BAIR as [2, 3, 6]. For Human3.6M, we calculate mean squared error (MSE) as [42]. To assess the learning of p_ψ and q_ϕ, we adopt the concept of [42] by introducing *Sampling* and *Reconstruction* metric (referred to as "S-" and "R-"), where the evaluation is performed on generation from prior and posterior respectively. For each test sequence, we generate 100 samples and compute the following metrics in 95% confidence interval:

- **Control Point Consistency (S-CPC)**: We compute the mean SSIM/PSNR/MSE between the generated end-frame and the targeted end-frame since CPC should be achieved for all samples.
- **Quality (S-Best)**: We compute the best SSIM/PSNR/MSE among all samples as [2, 3, 42]. It is a better way to assess the quality for stochastic methods because the best sample's score allows us to check if the true outcome is included in the generations.
- **Diversity (S-Div)**: Adapting the concept from [44], we compute the variance of SSIM/PSNR across all samples with the ground-truth sequences as reference. For MSE, we calculate the variance of difference between generated and ground-truth sequences instead because MSE only measures the distance between joints while ignoring their relative positions, which will result in biased estimation for diversity.

4.3. Quantitative Results

We show quantitative analysis on generation quality, diversity, and CPC over SM-MNIST, Weizmann, Human3.6M, and BAIR—in Tables 1, 2, & 3, as well as the comparison with more baselines in Table 4. From R-Best we know that the posteriors learn well in all setting. In Tables 1, 2, & 3, the model with CPC+Alignment losses (+C+A) outperforms the model with only CPC loss (+C) in S-CPC. This shows the effectiveness of alignment loss. Recall from Sec. 3.2 that there are two LSTMs that separate the encoder and decoder, the alignment loss aligns the two latent spaces to alleviate the mismatch between the encoding and the decoding process. Moreover, the model (*Ours*) with skip-frame training further improves over +C+A in S-CPC, where the gain mainly results from a better usage of time counter. Finally, S-CPC gain in Weizmann is less than SM-MNIST and Human3.6M since unlike the latter two, its data are captured in cluttered background with visible noise that is more challenging for CPC. On the other hand, when compared with more baselines [2, 3], our method successfully models the robot's movements while maintaining CPC without hurting diversity and quality as shown in Table 4.

On the generation quality, all four tables show comparable results in S-Best, which means that our method is able to maintain the quality while achieving CPC. Besides, the S-Best in Table 3 demonstrates an interesting finding that *Ours* not only achieves extremely superior performance in S-CPC but also in S-Best. The main reason is that Human3.6M contains 3D skeletons with highly diverse actions, giving rise to considerably flexible generation. A long-term generation may easily deviate from the others, causing high S-Best error, but our method gradually converges to the targeted end-frames, confining the S-Best error (see Sec. 4.5).

Regarding the generation diversity, our method attains either comparable or better performance in Tables 1 and 3. This suggests that our method generates diverse samples while reaching the same targeted end-frame. However, our method suffers from a larger performance drop on S-Div in Table 2. This is expected since Weizmann data often involve video sequences with unvarying actions, *e.g.* walking in a fixed speed, and therefore, posing constraints at the end-frame significantly reduces the possible generation trajectories and thus leads to low diversity. Overall, our method has a significant improvement on CPC while reaching comparable generation quality and diversity with the baseline.

Figure 7: Given a pair of (orange) start- and (red) end-frames, we show the results of various-length generation on Weizmann and Human3.6M (The number beneath each frame indicates the timestamp). Our model can achieve high-intermediate-diversity and targeted end-frame consistency while being aware of various-length generation at the same time.

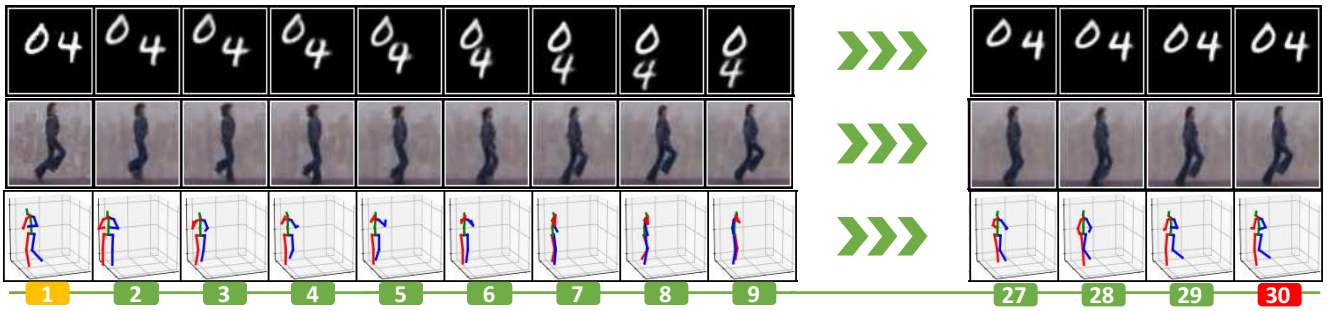

Figure 8: We set the (orange) start- and (red) end-frame with the same frame to achieve loop generation. Our model can generate videos that form infinite loops while preserving diversity. See more results in the supplementary materials.

4.4. CPC in Generation with Various Lengths

We show the CPC performance of all models under generation of different lengths on Human3.6M dataset in Fig. 3. The models achieve CPC under various lengths even though they have only seen the sequences with length around 30, showing that our models generalize well to various lengths. It is worth noting that with skip-frame training (red line), our model achieves CPC even further compared with other variations since it is able to leverage the information provided from the time counter. However, our method performs a bit worse at length 10 comparing to longer lengths because the model has less time budget for planning its trajectory and the training data do not contain any sequences with length less than 20.

4.5. Diversity Through Time

We evaluate the diversity of our method by investigating its behaviour through time in Fig. 5. The downward trend can be observed around the end of the green line, which means it tries to reach the targeted end-frame as the time-counter approaches the end. However, with the skip-frame training (red line), the diversity becomes higher around the middle segment and converges near the start- and end-frame. Our full model knows its precise status such as how far it is to the end-frame or how much time budget remains, and thus can plan ahead to achieve CPC. Since our model perceives well about its time budget, it can "go wild", *i.e.*, explore all possible trajectories while still being capable of getting back to the targeted end-frame on time.

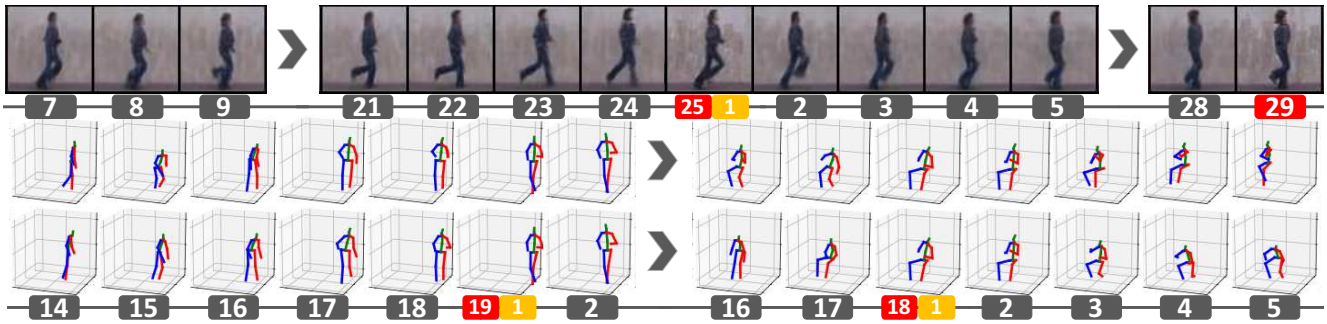

Figure 9: Given multiple pairs of (orange) start- and (red) end-frames, we can merge multiple generated clips into a longer video, which is similar to the modern video editing process. The number beneath each frame indicates the timestamp.

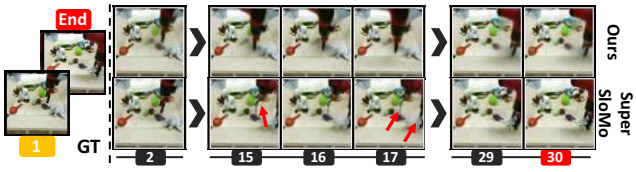

Figure 10: Longer time-interval video generation on BAIR.

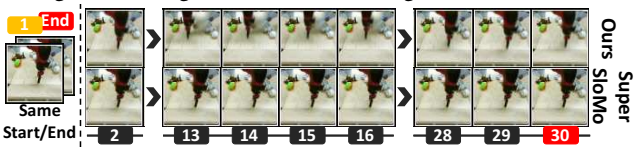

Figure 11: Automatic looping generation on BAIR Pushing.

4.6. CPC Weight on Prior vs. Posterior

We assess the effect of posing different CPC weights on prior p_ψ versus posterior q_ϕ by comparing the quality and diversity in SSIM (Fig. 4). With different weights, the behavior of diversity for p_ψ and q_ϕ is comparable. However, CPC on p_ψ (blue line) does not result in degradation throughout all CPC weights in comparison with posing CPC on q_ϕ. This shows that our method is more robust to different CPC weights.

4.7. Qualitative Results

Generation with various lengths. In Fig. 6, we show roughly how p2p generation works by comparing with [3] on BAIR dataset. Fig. 7 shows various examples across other datasets. Our model maintains high CPC for all lengths while producing diverse results.

Multiple control-points generation. In Fig. 9, we show the generated videos with multiple control points. The first row highlights transition across different attributes or actions (*i.e.*, "run" to "skip" in Weizmann dataset). The second and third rows show two generated videos with the same set of multiple control points (*i.e.*, stand; sit and lean to the left side). Note that these are two unique videos with diverse frames in transitional timestamps. By placing each control point as a breakpoint in a generation, we can achieve fine-grained controllability directly from frame exemplars.

Loop generation. Figs. 8 & 11 show that our method can be used to generate infinite looping videos by forcing the targeted start- and end-frames to be the same.

4.8. Comparison with Video Interpolation

To elaborate the essential difference between VI and p2p generation, we conduct a task of inserting 28 frames between start- and end-frame where the temporal distance between the targeted start- and end-frames is large (Fig. 10). Note that Super SloMo [16] produces artifacts such as distortion or two robot arms (indicated by red arrows in the 15th and 17th frames). VI methods typically are deterministic approaches while p2p generation is able to synthesize diverse frames (see Fig. 7). Finally, automatic looping can be accomplished by p2p generation but not by VI. Given the same start- and end-frames, we confirm that Super SloMo [16] will interpolate all the same frames as if the video is freezing (Fig. 11).

5. Conclusion

The proposed point-to-point (p2p) generation controls the generation process with two control points—the targeted start- and end-frames—to provide better controllability in video generation. To achieve control point consistency (CPC) while maintaining generation quality and diversity, we propose to maximize the modified variational lower bound for conditional video generation model, followed by a novel skip-frame training strategy and a latent space alignment loss to further reinforce CPC. We show the effectiveness of our model via extensive quantitative analysis. The qualitative results further highlight the merits of p2p generation. However, our current model cannot handle high-resolution videos. Modeling all the details such as small objects or noisy background in high-res videos is still an open problem for the existing video generation/prediction methods. We will explore this direction in the future. Overall, our work opens up a new dimension in video generation that is promising for further exploration.

*We thank MOST-107-2634-F-007-007, MOST-106-2221-E-007-080-MY3, MOST Joint Research Center for AI Technology, and All Vista Healthcare for their supports.

References

[1] Sandra Aigner and Marco Körner. Futuregan: Anticipating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing autoencoder gans. *arXiv preprint arXiv:1810.01325*, 2018.

[2] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. In *Proceedings of the International Conference on Learning Representations*, 2018.

[3] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *Proceedings of the International Conference on Machine Learning*, 2018.

[4] Emily L. Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, pages 4414–4423, 2017.

[5] Frederik Ebert, Chelsea Finn, Alex Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. In *CoRL*, 2017.

[6] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems*, pages 64–72, 2016.

[7] Katerina Fragkiadaki, Jonathan Huang, Alex Alemi, Sudheendra Vijayanarasimhan, Susanna Ricco, and Rahul Sukthankar. Motion prediction under multimodality with conditional stochastic networks. *arXiv preprint arXiv:1705.02082*, 2017.

[8] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, December 2007.

[9] Zekun Hao, Xun Huang, and Serge Belongie. Controllable video generation with sparse trajectories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7854–7863, 2018.

[10] Jiawei He, Andreas Lehrmann, Joseph Marino, Greg Mori, and Leonid Sigal. Probabilistic video generation using holistic attribute control. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–467, 2018.

[11] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems*, pages 515–524, 2018.

[12] Qiyang Hu, Adrian Waelchli, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. Video synthesis from a single image and motion stroke. *arXiv preprint arXiv:1812.01874*, 2018.

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[14] Dinesh Jayaraman, Frederik Ebert, Alexei A. Efros, and Sergey Levine. Time-agnostic prediction: Predicting predictable video frames. In *Proceedings of the International Conference on Learning Representations*, 2019.

[15] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *Advances in Neural Information Processing Systems*, pages 667–675, 2016.

[16] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: high quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[17] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1219–1228, 2018.

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations*, 2014.

[19] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018.

[20] Yitong Li, Martin Renqiang Min, Dinghan Shen, David E. Carlson, and Lawrence Carin. Video generation from text. *arXiv preprint arXiv:1710.00421*, 2017.

[21] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P Xing. Dual motion gan for future-flow embedded video prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1744–1752, 2017.

[22] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4463–4471, 2017.

[23] William Lotter, Gabriel Kreiman, and David Cox. Unsupervised learning of visual structure using predictive generative networks. In *Workshop Track of International Conference on Learning Representations*, 2016.

[24] Tanya Marwah, Gaurav Mittal, and Vineeth N Balasubramanian. Attentive semantic video generation using captions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1426–1434, 2017.

[25] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *Proceedings of the International Conference on Learning Representations*, 2016.

[26] Simone Meyer, Abdelaziz Djelouah, Brian McWilliams, Alexander Sorkine-Hornung, Markus H. Gross, and Christopher Schroers. Phasenet for video frame interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[27] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[28] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[29] MarcAurelio Ranzato, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*, 2014.

[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] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015.

[32] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using lstms. In *Proceedings of the International Conference on Machine Learning*, pages 843–852, 2015.

[33] 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.

[34] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. In *Proceedings of the International Conference on Learning Representations*, 2017.

[35] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances in Neural Information Processing Systems*, pages 613–621, 2016.

[36] Carl Vondrick and Antonio Torralba. Generating the future with adversarial transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1020–1028, 2017.

[37] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3332–3341, 2017.

[38] Olivia Wiles, A Koepke, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. In *Proceedings of the British Machine Vision Conference*, 2018.

[39] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2018.

[40] Tianfan Xue, Jiajun Wu, Katherine Bouman, and Bill Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2016.

[41] Shohei Yamamoto, Antonio Tejero-de Pablos, Yoshitaka Ushiku, and Tatsuya Harada. Conditional video generation using action-appearance captions. *arXiv preprint arXiv:1812.01261*, 2018.

[42] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 265–281, 2018.

[43] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016.

[44] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. In *Proceedings of the International Conference on Learning Representations*, 2019.

[45] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.

[46] Xiaoou Tang Yiming Liu Ziwei Liu, Raymond Yeh and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.