Human Motion Generation via Cross-Space Constrained Sampling

Zhongyue Huang, Jingwei Xu and **Bingbing Ni**[*]
Shanghai Jiao Tong University, China[†]
{116033910063, xjwxjw, nibingbing}@sjtu.edu.cn

Abstract

We aim to automatically generate human motion sequence from a single input person image, with some specific action label. To this end, we propose a cross-space human motion video generation network which features two paths: a forward path that first samples/generates a sequence of low dimensional motion vectors based on Gaussian Process (GP), which is paired with the input person image to form a moving human figure sequence; and a backward path based on the predicted human images to re-extract the corresponding latent motion representations. As lack of supervision, the reconstructed latent motion representations are expected to be as close as possible to the GP sampled ones, thus yielding a cyclic objective function for cross-space (i.e., motion and appearance) mutual constrained generation. We further propose an alternative sampling/generation algorithm with respect to constraints from both spaces. Extensive experimental results show that the proposed framework successfully generates novel human motion sequences with reasonable visual quality.

1 Introduction

Video generation, especially human motion video generation, has been attracting increasing research attention. Early methods [Vondrick *et al.*, 2016; Villegas *et al.*, 2017a] directly apply/extend conventional 2D GAN (i.e., used to deal with 2D image generation) to generate 3D spatio-temporal video. However, these approaches usually yield low video quality (non-realistic looking) due to the high dimensional searching space. To this end, some recent methods [Yan *et al.*, 2017; Walker *et al.*, 2017] have attempted to constrain the generator with human skeleton information (e.g., skeleton figures or joint position maps), thus to output more realistic articulated human motions. However, these methods also have significant limitations. First, most of these algorithms require for each frame a corresponding specific skeleton pattern for image synthesis. In other words, a sequence of skeleton representation vectors (or joint positions) should be given in prior for video generation. However, in most cases, it is very hard to get this information, which greatly limits its application. Second, these methods always require pairs of image frames with same background and identical person for supervised training. However, to obtain such strong supervised training data is very expensive, which in turn forbids further scale up of algorithmic training.

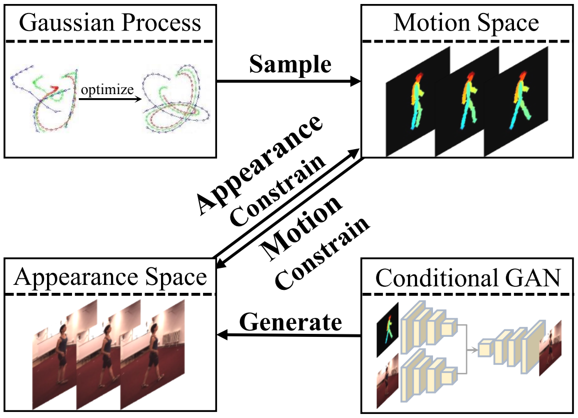

Figure 1: Motivation of cross-space human motion generation. To sample motions, we are conditioned on appearance constraints. To generate videos, we absorb constraints from skeleton motion space. By combining Gaussian Process and conditional GAN, we propose cross-space mutual constraint (i.e., motion and appearance).

To explicitly address these issues, this work proposes a new problem setting. Namely, given a **SINGLE** static image (with a human figure inside, denoted by ”input person”) and a video dataset of a specific human motion of some other persons (e.g., walking, dancing, denoted by ”target motion”), we aim to generate a novel video sequence of the input person acting some similar motion out. Note that the synthesized motion images should not follow exactly the same motion (i.e., the same joint position movements), and we DO allow randomness in terms of the motion. In other words, we shall first **sample** a proper sequence of (articulated) motion representations from the specific motion representation space of the target action type, and then according to these *generated* motion representations and the input human figure to further synthesize the full sequence of motion images.

[*] Corresponding Author: Bingbing Ni.

[†] Shanghai Institute for Advanced Communication and Data Science, Shanghai Key Laboratory of Digital Media Processing and Transmission, Shanghai Jiao Tong University, Shanghai 200240, China

It is therefore observed that for each time stamp, simultaneous sampling/generation of a particle in the motion representation space as well as an corresponding image in the appearance space is required, and moreover, the pair of samples in both space should be constrained by each other (i.e., make both sides compatible). Motivated by this observation, in this work we propose a cross-space human motion video generation network which features two paths: a forward path that first samples/generates a sequence of low dimensional motion vectors based on Gaussian Process (GP) (as GP is an effective latent space method modeling human motion), which is paired with the input person image to form a moving human figure sequence; and a backward path based on the predicted human images to re-extract the corresponding latent motion representations. As lack of supervision, the reconstructed latent motion representations are expected to be as close as possible to the GP sampled ones, thus yielding a cyclic objective function for cross-space (i.e., motion and appearance) mutual constrained generation. We further propose an alternative sampling/generation algorithm with respect to constraints from both spaces. As a form of self-supervision, the above framework **NO LONGER** needs pair of ground-truth image and input frame sharing same background and identical person for model training, makes the approach very flexible. Extensive experimental results show that the proposed framework successfully generates novel human motion sequences with reasonable visual quality.

2 Related Work

Gaussian Process. [Lawrence, 2004] proposes GPLVM that apply Gaussian Process [Seeger, 2004] to model articulated motion. [Wang *et al.*, 2008] further extends GPLVM to a dynamic version which models temporal dependent data. [Bui *et al.*, 2016] models the multiple properties of single object with hierarchical Gaussian Process. Many researchers utilize Gaussian Process for various applications like 3D human tracking [Sedai *et al.*, 2013] and human action classification [Li and Marlin, 2016]. However, traditional Gaussian Process can only sample low dimensional data like skeletons, and not high dimensional data like videos. To address this issue, our work generates video sequence using GP-GAN cross-space constrained sampling.

Video Generation. The early methods [Schödl *et al.*, 2000; Agarwala *et al.*, 2005] utilize video texture to generate video sequence. In the last years, there have been a number of works applying GAN on video generation. [Vondrick *et al.*, 2016] generates a spatio-temporal cuboid with two independent streams: a moving foreground pathway and a static background pathway. [Villegas *et al.*, 2017a] proposes a motion-content network that apply Convolutional LSTM to capture the temporal dynamics. Some works use pose information in human video generation. [Ma *et al.*, 2017] introduces a two-stage approach to generate a new image from a person image and a novel pose. [Yan *et al.*, 2017] further considers the continuity and smoothness of the video frames and generate a human motion video with a single iamge and human skeleton sequence. [Walker *et al.*, 2017], [Villegas *et al.*, 2017b] and [Denton and vighnesh Birodkar, 2017] first model possible movements of humans in the pose space, and then use the future poses to predict the future frames in the pixel space. These works require a lot of auxiliary constraint data such as the skeleton of each frame. However, our method overcomes this limitation with cross-space constraint and only needs the motion space.

3 Method

3.1 Problem Definition and Method Overview

As illustrated in Section 1, we first formulate the task mathematically as follows: given a static image (**"input person"**), denoted as $\mathbf{y}_0 \in R^{1\times C\times W\times H}$, and a specific human video sequence, denoted as $\mathbf{Y} \in R^{N\times C\times W\times H}$, we aim to first sample a motion sequence, denoted as $\hat{\mathbf{S}} \in R^{N\times D}$ (**"target motion"**), then generate a novel video, denoted as $\hat{\mathbf{Y}} \in R^{N\times C\times W\times H}$, of the input person $\mathbf{y}_0$ acting the target motion $\hat{\mathbf{S}}$ out to some extent of motion randomness. Here C, W and H represent the image channel, width and height. N denotes the length of the sequence. D is feature dimension.

There are two long standing problems on generating videos, i.e., motion consistency as well as smoothness. As stated above, the motion randomness is another important consideration. In the majority of the previous works [Villegas *et al.*, 2017b; Denton and vighnesh Birodkar, 2017], ground truth is treated as exclusive generation target, which significantly restricts the generation variety. So it is demanding to develop a novel generation scheme which could deal with these three aspects appropriately.

To this end, we propose a cross-space human motion video generation network. First, the proposed network generates a sequence of low dimensional motion representation vectors $\hat{\mathbf{S}}$ based on a Gaussian Process (**Motion Generation Module**). Second, the generated motion vectors $\hat{\mathbf{S}}$ are utilized to form a moving human figure sequence $\hat{\mathbf{Y}}$ (**Appearance Generation Module**) based on the static image $\mathbf{y}_0$. Third, taking above three aspects into consideration, we propose a cyclic framework to sample instances from both spaces in a collaborative way, where sampling in one space is conditioned on the other one. Detailed architecture of our model is shown in Figure 2.

3.2 Motion Generation Module

This module is based on Gaussian Process, which consists of a mapping from a latent space $\mathcal{T}$ to the motion space $\mathcal{S}$ and a dynamical model in the latent space. A latent variable mapping can be formulated as

$$\mathbf{S} = f\left(\mathbf{T}\right) + \epsilon, \tag{1}$$

where $\mathbf{S}$ is the motion sequence, $\mathbf{T} = \{\mathbf{t}_1, ..., \mathbf{t}_N\}$ denotes the latent representation, f is a non-linear mapping, and ϵ is a zero-mean white Gaussian noise process.

Inspired from Gaussian Process Dynamical Model (GPDM) [Wang *et al.*, 2008], we obtain the latent prior by marginalizing out the parameters

$$\begin{aligned} &p\left(\mathbf{T}|\bar{\alpha}\right) \\ &= \frac{p\left(\mathbf{t}_1\right)}{\sqrt{\left(2\pi\right)^{(N-1)d}\left|\mathbf{K}_T\right|^d}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{K}_T^{-1}\mathbf{T}_{2:N}\mathbf{T}_{2:N}^T\right)\right), \end{aligned} \tag{2}$$

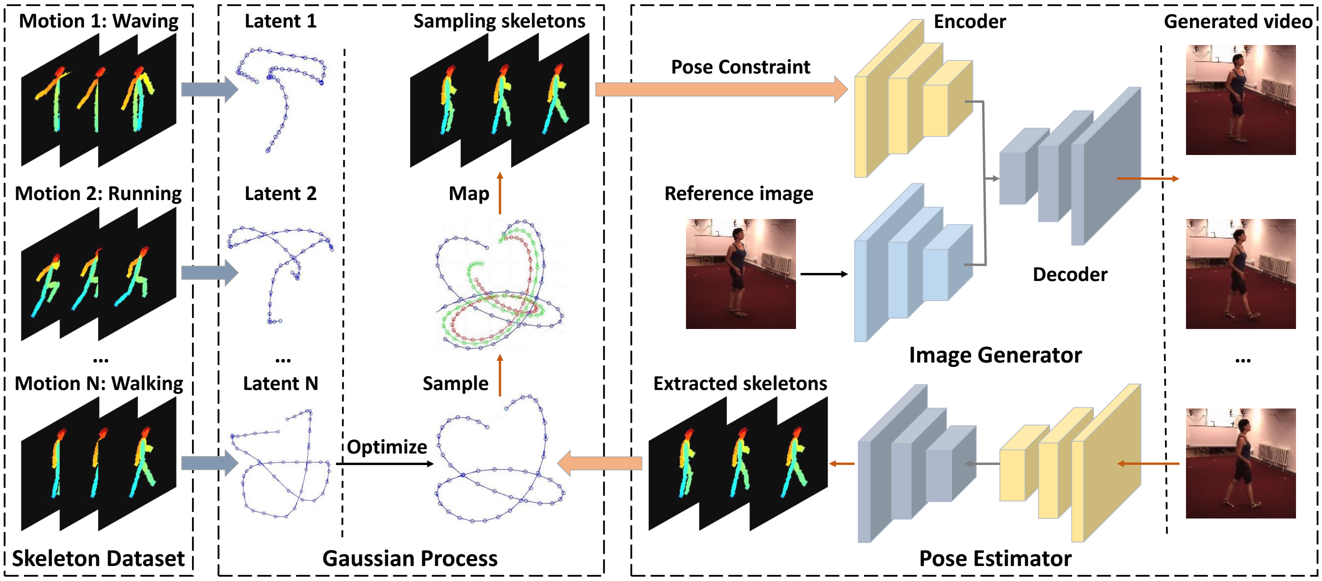

Figure 2: The cross-space human motion video generation framework. This network features two paths: a forward path that first samples/generates a sequence of low dimensional motion vectors based on Gaussian Process (GP), which is paired with the input person image to form a moving human figure sequence; and a backward path based on the predicted human images to re-extract the corresponding latent motion representations. The two paths construct cyclic cross-space mutual constraint, i.e., motion and appearance.

where $\mathbf{K}_T$ is a kernel matrix with hyperparameters $\bar{\alpha}$ where the elements are defined by a linear and RBF kernel

$$k_T\left(\mathbf{t},\mathbf{t}'\right) = \alpha_1 \exp\left(-\frac{\alpha_2}{2}\left\|\mathbf{t}-\mathbf{t}'\right\|^2\right) + \alpha_3 \mathbf{t}^T\mathbf{t}' + \alpha_4^{-1}\delta_{\mathbf{t},\mathbf{t}'}. \tag{3}$$

Similarly, GPDM marginalizes over f in closed form and obtains a Gaussian density

$$p\left(\mathbf{S}|\mathbf{T},\bar{\beta},\mathbf{W}\right) = \frac{|\mathbf{W}|^N}{\sqrt{(2\pi)^{ND}|\mathbf{K}_S|^D}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{K}_S^{-1}\mathbf{S}\mathbf{W}^2\mathbf{S}^T\right)\right), \tag{4}$$

where $\mathbf{K}_S$ is a kernel matrix with hyperparameters $\bar{\beta}$ and $\mathbf{W}$ is a scaling matrix. $\mathbf{K}_S$ is constructed by a RBF kernel

$$k_S\left(\mathbf{s},\mathbf{s}'\right) = \exp\left(-\frac{\beta_1}{2}\left\|\mathbf{s}-\mathbf{s}'\right\|^2\right) + \beta_2^{-1}\delta_{\mathbf{s},\mathbf{s}'}. \tag{5}$$

Thus, the generative model can be defined by

$$p\left(\mathbf{S},\mathbf{T},\bar{\alpha},\bar{\beta},\mathbf{W}\right) = p\left(\mathbf{S}|\mathbf{T},\bar{\beta},\mathbf{W}\right)p\left(\mathbf{T}|\bar{\alpha}\right)p\left(\bar{\alpha}\right)p\left(\bar{\beta}\right)p\left(\mathbf{W}\right), \tag{6}$$

where the priors are $p\left(\bar{\alpha}\right) \propto \prod \alpha_i^{-1}$, $p\left(\bar{\beta}\right) \propto \prod \beta_i^{-1}$ and

$$p\left(\mathbf{W}\right) = \prod_{m=1}^{D} \frac{2}{\kappa\sqrt{2\pi}} \exp\left(-\frac{w_m^2}{2\kappa^2}\right). \tag{7}$$

We follow the strategy of [Wang *et al.*, 2008] and use the two-stage map estimation algorithm to optimize the Gaussian Process model. We first estimate the hyperparameters $\Theta = \{\alpha,\beta,\mathbf{W}\}$. Then we update $\mathbf{T}$ with holding Θ fixed. In the fisrt step, we optimize

$$\mathcal{L}_\varepsilon\left(\Theta\right) = -\int_{\mathbf{T}} p\left(\mathbf{T}|\mathbf{S},\Theta\right)\ln p\left(\mathbf{S},\mathbf{T}|\Theta\right)d\mathbf{T}. \tag{8}$$

Using Hamiltonian Monte Carlo (HMC) to sample $\mathbf{T}$ from $p\left(\mathbf{T}|\mathbf{S},\bar{\alpha},\bar{\beta},\mathbf{W}\right)$, we obtain the approximation

$$\mathcal{L}_\varepsilon\left(\Theta\right) \approx -\frac{1}{R}\sum_{r=1}^{R}\ln p\left(\mathbf{S},\mathbf{T}^{(r)}|\Theta\right). \tag{9}$$

In the second step, we optimize

$$\mathcal{L}_\theta\left(\mathbf{T}\right) = \ln p\left(\mathbf{T},\Theta|\mathbf{S}\right). \tag{10}$$

The optimization procedure is illustrated in Algorithm 1.

3.3 Appearance Generation Module

Given a static image $\mathbf{y}_0$ and a sequence of skeletons $\hat{\mathbf{S}} = \{\hat{\mathbf{s}}_1,...,\hat{\mathbf{s}}_N\}$ produced by the motion generation module, this module aims to generate a video $\hat{\mathbf{Y}} = \{\mathbf{y}_1,...,\mathbf{y}_N\}$ of the input person acting the target motion out. The motion in the video should be coherent and the appearance ought to remain consistent over frames. To this end, we propose a conditional GAN, which is able to fully utilize the appearance as well as the skeleton information. On one hand, inspired from [Zhu *et al.*, 2017], pairs of skeleton and appearance images are stacked along the channel as input. On the other hand, the discriminator is required to distinguish ground truth from generated ones conditioned on the static image $\mathbf{y}_0$.

Adversarial Loss. After sampling a sequence of skeletons, we use the conditional GAN above to generate a completely new video sequence. Our generative model has two inputs of an image $\mathbf{y}_0$ and skeletons $\mathbf{S}$. We apply adversarial loss [Goodfellow *et al.*, 2014] as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{GAN}}\left(G,D\right) = &\ \mathbb{E}_{y\sim p_{data}(y)}\left[\log D\left(\mathbf{Y}\right)\right] \\ &+ \mathbb{E}_{s\sim p_{data}(s)}\left[\log\left(1-D\left(G\left(\mathbf{y}_0,\mathbf{S}\right)\right)\right)\right],\end{aligned} \tag{11}$$

where G tries to generate images $G\left(\mathbf{y}_0,\mathbf{S}\right)$, and D tries to distinguish between synthesized images $G\left(\mathbf{y}_0,\mathbf{S}\right)$ and real images $\mathbf{Y}$.

3.4 Cross-space Constraint

Only using adversarial loss is not feasible to generate satisfying results, because these two modules are trained independently, i.e., the motion correspondences can not be guaranteed. To this end, we propose cross-space constraint to bridge the gap between the skeleton and video space. The proposed constraint, consisting of consistency and smoothness in both two spaces, is shown as follows.

Pose Constraint (Consistency & Smoothness). There are many possible mapping functions that the generator may learn [Zhu *et al.*, 2017]. To pursue the most likely one, we utilize the pose information contained in the video space to ensure the pose consistency. To be specific, the poses of synthesized images should correspond to the input poses (**consistency**). We use a pose estimator F to extract the human skeletons from the generated images which should be near the input skeletons, and apply an L1 loss

$$\mathcal{L}_{\text{con}}\left(G\right) = \mathbb{E}_{s\sim p_{data}(s)}\left[\left\|F\left(G\left(\mathbf{y}_0, \mathbf{S}\right)\right) - \mathbf{S}\right\|_1\right]. \tag{12}$$

Another constraint is that the trajectory of human motion is usually smooth, e.g. the limb swing during walking, which is also contained in the video space. Note that the MAP estimation of Gaussian Process inside our motion generation module is closely related to the motion smoothness [Wang *et al.*, 2008]. So we use the log-likelihood to measure the **smoothness** of generated poses:

$$\mathcal{L}_{\text{smo}}\left(G\right) = -\text{log}(p(F(G(\mathbf{y}_0, \mathbf{S}))|\mathbf{T}, \Theta)). \tag{13}$$

Appearance Constraint (Consistency & Smoothness). In the video space, we introduce some constraint that can preserve smoothness of appearance between the input and output. We make the generator to approximate an identity mapping using a pixel-wise identity loss

$$\mathcal{L}_{\text{idp}}\left(G\right) = \mathbb{E}_{s\sim p_{data}(s)}\left[\left\|G\left(\mathbf{y}_0, \mathbf{S}\right) - \mathbf{Y}\right\|_1\right]. \tag{14}$$

To keep consistency of the person when generating the images, we use a classifier C to determine whether the synthesized images belong to the same category as the input image.

$$\mathcal{L}_{\text{idc}}\left(G\right) = \mathbb{E}_{s\sim p_{data}(s)}\left[-y\,\text{log}C\left(G\left(\mathbf{y}_0, \mathbf{S}\right)\right)\right]. \tag{15}$$

Full objective. Considering all above constraints, our full objective is

$$\begin{aligned}\mathcal{L}\left(G, D\right) = {} & \mathcal{L}_{\text{GAN}}\left(G, D\right) + \lambda\mathcal{L}_{\text{con}}\left(G\right) \\ & + \gamma\mathcal{L}_{\text{smo}}(G) + \alpha\mathcal{L}_{\text{idp}}\left(G\right) + \beta\mathcal{L}_{\text{idc}}\left(G\right),\end{aligned} \tag{16}$$

where λ, γ, α, β are the weights of different objectives. We aim to solve

$$G^* = \arg\min_G\max_D\mathcal{L}\left(G, D\right). \tag{17}$$

Optimization. During the optimization procedure, we iteratively perform sampling motion sequence and generating video sequence as follows:

- **Static Image to Skeleton Sequence.** Given a starting point $\mathbf{s}_0^*$, the corresponding latent representation $\mathbf{t}_0^*$ can be estimated by

  $$\hat{\mathbf{t}}_0^* = \arg\max_{\mathbf{t}} - \ln\left(p\left(\mathbf{t}|\mathbf{T}, \mathbf{S}, \Theta, \mathbf{s}_0^*\right)\right) \tag{18}$$

  Then we use a HMC sampler to generate a new latent sequence $\hat{\mathbf{T}}^*$, and apply a Gaussian prediction to get the final prediction $\hat{\mathbf{S}}^*$:

  $$\hat{\mathbf{s}}_t^* = \mathbf{S}^T\left(\mathbf{K}_S\right)^{-1}\left[\mathbf{K}_S\left(\hat{\mathbf{T}}^*, \hat{\mathbf{t}}_t^*\right)\right]. \tag{19}$$

  This procedure is optimized according to $\mathcal{L}_\varepsilon\left(\Theta\right)$ and $\mathcal{L}_\theta\left(\mathbf{T}\right)$.

- **Skeleton Sequence to Video Sequence.** Given a skeleton sequence, we use the proposed generator to produce new video sequence. And here the cross-space constraint $\mathcal{L}\left(G, D, F, C\right)$ is applied in both motion and video space.

We summarize the whole optimization procedure in Algorithm 1.

Algorithm 1 Optimization Algorithm

```
Input: M human images $\left\{\mathbf{y}^{(n)}\right\}_{n=1}^{M}$.
       Q video sequences $\left\{\mathbf{Y}^{(q)}\right\}_{q=1}^{Q}$.
       Q skeleton sequences $\left\{\mathbf{S}^{(q)}\right\}_{q=1}^{Q}$.
Output: parameters $\Theta^* = \{\alpha^*, \beta^*, \mathbf{W}^*\}$, matrix $\mathbf{T}^*$,
        generator $G^*$.
Initialize $\bar{\alpha} \Leftarrow (0.9, 1, 0.1, e), \bar{\beta} \Leftarrow (1, 1, e), \{w_k\} \Leftarrow 1$.
Initialize $G, D$ using xavier.
repeat
    // Motion Generation (Gaussian Process)
    Sample $\left\{\mathbf{T}^{(r)}\right\}_{r=1}^{R} \sim p\left(\mathbf{T}|\mathbf{S}, \bar{\alpha}, \bar{\beta}, \mathbf{W}\right)$.
    Construct $\left\{\mathbf{K}_S^{(r)}, \mathbf{K}_T^{(r)}\right\}_{r=1}^{R}$ from $\left\{\mathbf{T}^{(r)}\right\}_{r=1}^{R}$.
    for $j = 1 : J$ do
        for $k = 1 : D$ do
            $\mathbf{d} \Leftarrow \left[(\mathbf{S})_{1k}, ..., (\mathbf{S})_{Nk}\right]^T$.
            $w_k^2 \Leftarrow N\left(\mathbf{d}^T\left(\frac{1}{R}\sum_{r=1}^{R}\left(\mathbf{K}_S^{(r)}\right)^{-1}\right)\mathbf{d} + \frac{\mathbf{1}}{\kappa^{\mathbf{2}}}\right)^{-1}$.
        end for
        $\left\{\bar{\alpha}, \bar{\beta}\right\} \Leftarrow$ minimize $\mathcal{L}_\varepsilon\left(\Theta\right)$ (Equation 9)
            using SCG for K iterations.
        $\{\mathbf{T}\} \Leftarrow$ maximize $\mathcal{L}_\theta\left(\mathbf{T}\right)$ (Equation 10)
            using SCG for K iterations.
    end for
    // Appearance Generation (Conditional GAN)
    for $i = 1 : I$ do
        Sample $\hat{\mathbf{S}}$.
        for all batches $\mathbf{S}_b$ in $\hat{\mathbf{S}}$ do
            Minimize $\mathcal{L}\left(G, D\right)$ (Equation 16).
            Update G, D.
        end for
    end for
until Convergence
```

3.5 Implementation Details

We implement Gaussian Process based on [de G. Matthews *et al.*, 2017], and adapt the architecture for generative networks from [Zhu *et al.*, 2017]. All networks were trained using the

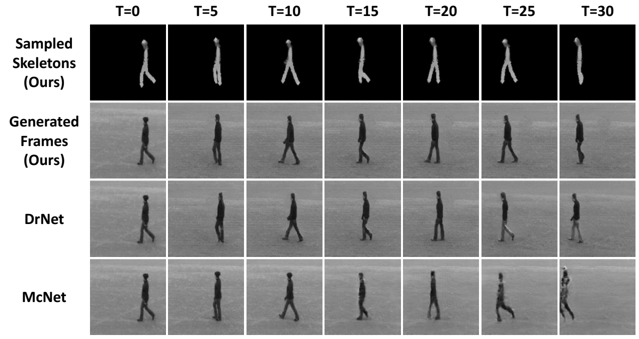

Figure 3: Qualitative comparison on KTH dataset. The first column (T=0) is real frames. We display the results every five frames. The first two rows are sampled skeletons and generated frames of our method. The last two rows are the results of DrNet and McNet.

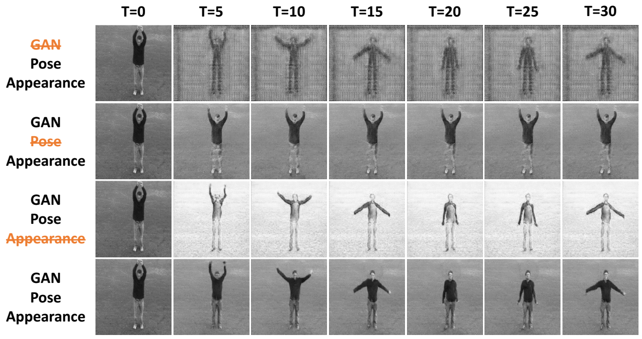

Figure 4: Examples generated with different loss terms on KTH dataset. The full losses are GAN loss, pose loss and apperance loss. Each row reduces one of them in the first three rows.

	handwaving	walking	running	all
Pose + Appearance	1.39	1.39	1.37	1.47
GAN + Appearance	1.43	1.41	1.44	1.37
GAN + Pose	2.03	1.47	1.71	2.04
Full losses	1.79	1.54	1.54	1.86
Real Data	2.08	1.60	1.69	2.05

Table 1: Inception Scores for different variants of our method on KTH dataset.

Adam solver with a leanrning rate as 0.0001 and a batch size of 10. We set $\lambda = \gamma = 10$ and $\alpha = \beta = 1$. As shown in [Zhu *et al.*, 2017], we apply a least-squares loss [Mao *et al.*, 2017] rather than the negative log likelihood for higher quality results. We also update the discriminator with the 50 previously generated images [Shrivastava *et al.*, 2017] instead of the latest one to reduce model oscillation. The skeletons are extracted by the state-of-the-art pose estimator (OpenPose) [Cao *et al.*, 2017]. And we fine tune the ResNet-18 [He *et al.*, 2016] pre-trained in the ImageNet as the classifier.

4 Experiment

In this section, we conduct both quantitative and qualitative experiments to evaluate the performance of proposed framework. To demonstrate the effectiveness of our model, we also conduct detailed comparison experiments with two strong baselines. Meanwhile we present in-depth analysis on the contribution of proposed cross-space constraint, which are the key component in our model. Details are given as follows.

4.1 Datasets

KTH Dataset. This dataset [Schuldt *et al.*, 2004] contains six types of human actions: walking, running, jogging, boxing, hand clapping and hand waving. There are 25 subjects with four different scenarios outdoors and indoors. And the videos are gray scale with a resolution of 120 × 160.

Human3.6M Dataset. This dataset [Ionescu *et al.*, 2014] offers poses and videos taken from 10 actors in 17 scenarios. All videos are recorded from four different views simultaneously in indoor environment. The main difficulties of generation lie in two aspects : (1) The Human3.6M dataset contains many subtle movements throughout all video sequences, for example random swing of limbs. (2) The appearance of 10 actors vary largely, which is very difficult for the model to learn.

4.2 Evaluation

Evaluation Setup. We compare the proposed model against two strong baselines: McNet [Villegas *et al.*, 2017a] and DrNet [Denton and vighnesh Birodkar, 2017], which achieve state-of-the-art performance in human motion generation task. To fairly compare with them, we follow the evaluation setup of DrNet, where the 10th frame is used as input image $\mathbf{y}_0$, to generate the following 30 frames. For KTH datasets, we use person 1-15 for training and 16-25 for testing. For Human3.6M dataset, we train on subject 1, 5, 6, 7, 8 and test on subject 9, 11. We resize the frames to resolution of 256×256. Note that not all the skeletons could be successfully detected by the OpenPose. To solve this problem, we manually annotate the skeletons of all the remaining failure cases.

Qualitative Evaluation. Figure 3 demonstrates qualitative results of our method and two baselines on KTH dataset. Here we remark two observations: (1) With the time step increasing, both the proposed model and DrNet well preserve the shape completeness of human figures, while it degrades rapidly on the results of MCNet. (2) Compared to DrNet, our model further successfully preserves the color of human body and pose transition during the long-term prediction (Note the comparison with DrNet from $T = 20$ to $T = 30$). These results clearly demonstrate that the articulated information (human skeleton) boosts the generation quality of human motion by a large margin. And more importantly, with the help of cross-space constraint, our model maintains the appearance consistency and motion continuity, **especially on the long-term generation task**.

Quantitative Evaluation. To quantitatively evaluate the quality of generated results, we conduct the Inception Score [Salimans *et al.*, 2016] comparison experiments with these two baselines. As commonly used evaluation metric on image generation task, it is considered more appropriate than the PSNR or SSIM on video generation domain. Following the evaluation setup of DrNet, the Inception Score is calculated

	handclapping	handwaving	walking	running	jogging	boxing	all
Prefers ours over McNet	76.2%	80.3%	85.0%	83.5%	84.6%	78.1%	81.5%
Prefers ours over DrNet	55.3%	57.4%	73.7%	72.1%	70.8%	60.6%	63.6%

Table 2: Human Preference Evaluation on KTH datasets.The last column is the results of all generated videos.

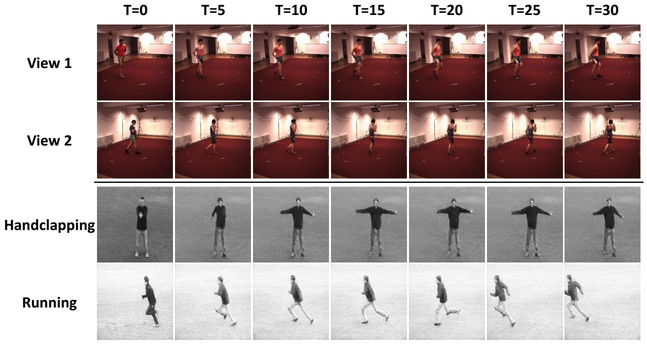

Figure 5: More examples of our method. The first two rows are experimented on Human3.6M dataset, while the others are tested on KTH dataset. We present the results of different views and different motions.

for 100 timesteps. As shown in Figure 6, the Inception Score of MCNet is unstable and decays rapidly with the increasing number of generated frames, which indicates degraded video quality during long-term generation task. The performance of DrNet is close to MCNet during the first few timesteps (from $T = 0$ to $T = 40$), and keeps relatively higher scores at latter time steps(from $T = 40$ to $T = 100$). The proposed model outperforms these two baselines during the whole generation procedure. It indicates that benefiting from the cross-space constraint, the sampled clear and reasonable motion guarantees the following high quality video generation.

Human Preference Evaluation. To comprehensively evaluate the performance of our model, we also conduct human preference experiment, according to the human psychophysical evaluation metric [Vondrick *et al.*, 2016]. We conduct 2000 comparisons in total with 50 different people (40 preference selection for each person), i.e., we collect 1000 comparisons against McNet and 1000 against DrNet. They are required to select one out of two videos they prefer under the criteria of appearance consistency and motion continuity. As shown in Table 2, our results are perceptually more realistic compared to the baselines. However, we observe nearly the same preference proportion for the hand clapping, hand waving and boxing motions. It mainly results from these actions usually involve tiny movement, which largely reduces the video generation difficulty. When generating more complicated motion like walking and running, our approach performs much better than these baselines, benefiting from the proposed cross-space constraint. Overall, out method outperforms the baselines by a large margin, i.e. 81.5% of people prefer ours over McNet and 63.6% over DrNet.

Ablation Study. For the key contribution of our work, **cross-space constraint**, we conduct detailed ablation study to evaluate the effectiveness of each component. As illustrated in Figure 4, we can observe that these losses have different effects on the generated videos. Specificly, removing the GAN loss leads to highly unnatural generation results which completely losses the original texture of the reference image. It indicates that the adversarial loss is important for video generation task. Removing the pose constraint, consisting of Equation 12 and 13, results in the poses nearly unchanged during the whole video sequence. It shows that pose loss mainly helps to learn the target motion transition generated by the motion generation module. The appearance constraint consists of Equation 14 and 15. Removing it makes the generator fail to preserve color consistency between the input and output. Even we input a dim image, the network generates very bright frames. Combining these three losses together, our model could largely benefit from both appearance and pose constrains as well as the cyclic training framework. The generation results are visually satisfying. Meanwhile we perform quantitative evaluations over these loss terms with results shown in Table 1. We compute Inception Score with three types of action in KTH dataset. Note that the reality and diversity are important criteria in Inception Score evaluation, which do not take the consistency into consideration. So one can notice that relatively higher scores in term of GAN+Pose, which mainly results from the lack of appearance constraint. Generally all three components have positive effect on our video generation task. The proposed cross-space constrained framework boosts the performance by a large margin.

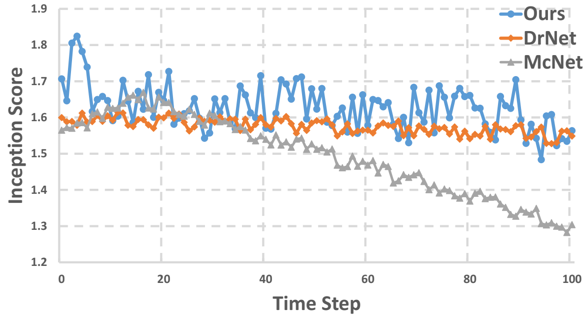

Figure 6: Quantitative comparison on KTH dataset. The Inception Score is calculated for 100 timesteps. X-axis indicates the timesteps of generated videos. Y-axis denotes Inception Score of frames in each timestep.

5 Conclusion

In this work, we propose a cross-space human motion video generation model. By combining Gaussian Process and conditional GAN, we introduce a cyclic objective for cross-space mutual constrained generation. Experiments show that our model generates high quality video sequence of the input person acting the target motion out.

Acknowledgments

This work was supported by National Science Foundation of China (U1611461, 61502301, 61521062). The work was partially supported by China's Thousand Youth Talents Plan, STCSM 17511105401 and 18DZ2270700.

References

[Agarwala *et al.*, 2005] Aseem Agarwala, Ke Colin Zheng, Chris Pal, Maneesh Agrawala, Michael F. Cohen, Brian Curless, David Salesin, and Richard Szeliski. Panoramic video textures. *international conference on computer graphics and interactive techniques*, 24(3):821–827, 2005.

[Bui *et al.*, 2016] Thang D. Bui, José Miguel Hernández-Lobato, Daniel Hernández-Lobato, Yingzhen Li, and Richard E. Turner. Deep gaussian processes for regression using approximate expectation propagation. *ICML*, pages 1472–1481, 2016.

[Cao *et al.*, 2017] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *2017 CVPR*, pages 1302–1310, 2017.

[de G. Matthews *et al.*, 2017] Alexander G. de G. Matthews, Mark van der Wilk, Tom Nickson, Keisuke Fujii, Alexis Boukouvalas, Pablo León-Villagrá, Zoubin Ghahramani, and James Hensman. Gpflow: a gaussian process library using tensorflow. *Journal of Machine Learning Research*, 18(40):1–6, 2017.

[Denton and vighnesh Birodkar, 2017] Emily L. Denton and vighnesh Birodkar. Unsupervised learning of disentangled representations from video. *NIPS*, pages 4417–4426, 2017.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in NIPS 27*, pages 2672–2680, 2014.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 CVPR*, pages 770–778, 2016.

[Ionescu *et al.*, 2014] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 36(7):1325–1339, 2014.

[Lawrence, 2004] Neil D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in NIPS 16*, pages 329–336, 2004.

[Li and Marlin, 2016] Steven Cheng-Xian Li and Benjamin M. Marlin. A scalable end-to-end gaussian process adapter for irregularly sampled time series classification. *NIPS*, pages 1804–1812, 2016.

[Ma *et al.*, 2017] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *NIPS*, pages 405–415, 2017.

[Mao *et al.*, 2017] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *2017 ICCV*, pages 2813–2821, 2017.

[Salimans *et al.*, 2016] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *NIPS*, pages 2234–2242, 2016.

[Schuldt *et al.*, 2004] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th ICPR 2004.*, volume 3, pages 32–36, 2004.

[Schödl *et al.*, 2000] Arno Schödl, Richard Szeliski, David H. Salesin, and Irfan Essa. Video textures. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000.

[Sedai *et al.*, 2013] Suman Sedai, Mohammed Bennamoun, and Du Q. Huynh. A gaussian process guided particle filter for tracking 3d human pose in video. *IEEE Transactions on Image Processing*, 22(11):4286–4300, 2013.

[Seeger, 2004] Matthias W. Seeger. Gaussian processes for machine learning. *International Journal of Neural Systems*, 14(2):69–106, 2004.

[Shrivastava *et al.*, 2017] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *2017 CVPR*, pages 2242–2251, 2017.

[Villegas *et al.*, 2017a] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. In *ICLR 2017*, 2017.

[Villegas *et al.*, 2017b] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. *ICML*, pages 3560–3569, 2017.

[Vondrick *et al.*, 2016] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *NIPS*, pages 613–621, 2016.

[Walker *et al.*, 2017] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *2017 ICCV*, pages 3352–3361, 2017.

[Wang *et al.*, 2008] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *pattern recognition and machine intelligence*, 30(2):283–298, 2008.

[Yan *et al.*, 2017] Yichao Yan, Jingwei Xu, Bingbing Ni, Wendong Zhang, and Xiaokang Yang. Skeleton-aided articulated motion generation. In *Proceedings of the 2017 ACMMM*, pages 199–207, 2017.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 ICCV*, pages 2242–2251, 2017.