International Journal of Computer Vision (2023) 131:2699–2722
https://doi.org/10.1007/s11263-023-01832-8

MANUSCRIPT

Conditional Temporal Variational AutoEncoder for Action Video Prediction

Xiaogang Xu[1] · Yi Wang[2] · Liwei Wang[3] · Bei Yu[3] · Jiaya Jia[3]

Received: 21 March 2022 / Accepted: 24 May 2023 / Published online: 18 June 2023

Abstract
To synthesize a realistic action sequence based on a single human image, it is crucial to model both motion patterns and diversity in the action video. This paper proposes an Action Conditional Temporal Variational AutoEncoder (ACT-VAE) to improve motion prediction accuracy and capture movement diversity. ACT-VAE predicts pose sequences for an action clip from a single input image. It is implemented as a deep generative model that maintains temporal coherence according to the action category with a novel temporal modeling on latent space. Further, ACT-VAE is a general action sequence prediction framework. When connected with a plug-and-play Pose-to-Image network, ACT-VAE can synthesize image sequences. Extensive experiments bear out our approach can predict accurate pose and synthesize realistic image sequences, surpassing state-of-the-art approaches. Compared to existing methods, ACT-VAE improves model accuracy and preserves diversity.

Keywords Variational AutoEncoder · Action modeling · Temporal coherence · Adversarial learning

1 Introduction

Human action video prediction aims to generate future human action from single or multiple input human images (Kim et al., 2019; Villegas et al., 2017; Wichers et al., 2018; Zhao et al., 2018). This topic is actively studied recently, for its importance to understand and improve human motion modeling and benefit in a variety of video applications, e.g, motion re-target (Aberman et al., 2019; Villegas et al., 2018). In this work, we focus on synthesizing image sequences from a single image and controlling their action types via the input of action labels (Kim et al., 2019), as shown in Fig. 1.

✉ Xiaogang Xu
xgxu@zhejianglab.com

Yi Wang
wangyi@pjlab.org.cn

Liwei Wang
lwwang@cse.cuhk.edu.hk

Bei Yu
byu@cse.cuhk.edu.hk

Jiaya Jia
leojia@cse.cuhk.edu.hk

[1] Zhejiang Lab, Hangzhou City, Zhejiang Province, China

[2] Shanghai AI Laboratory, Shanghai City, China

[3] The Chinese University of Hong Kong, Hong Kong, China

Due to the diversity in human motion, action video prediction is highly ill-posed with multiple possible solutions. Conventional deterministic models utilizing regression are useful, but over-smooth image sequences may be produced (Finn et al., 2016; Jia et al., 2016; Kalchbrenner et al., 2017; Villegas et al., 2017), giving mean estimation of future action. Recent deep generative approaches alleviate this problem by using Generative Adversarial Networks (GAN) (Cai et al., 2018; Wichers et al., 2018), Variational AutoEncoder (VAE) (Kim et al., 2019; Li et al., 2018), and Variational Recurrent Neural Network (VRNN) (Castrejon et al., 2019; Denton & Fergus, 2018) to model motion diversity explicitly. Methods of Cai et al. (2018); Wichers et al. (2018); Kim et al. (2019); Li et al. (2018) use latent variables with identical independent distributions to capture motion patterns and diversity in every frame. Without temporal coherence among latent variables, action video prediction accuracy is bounded (The prediction accuracy refers to the difference between the synthesized pose sequences' distribution and the real pose sequences' distribution within each category), which has been proved by existing works that consider the issue of coherence (Castrejon et al., 2019; Mao et al., 2019, 2020). Meanwhile, works of Castrejon et al. (2019); Denton and Fergus (2018) introduced unitive temporal coherence for all actions while ignored the distinction among different action categories. We consider temporal coherence from two aspects. For the first

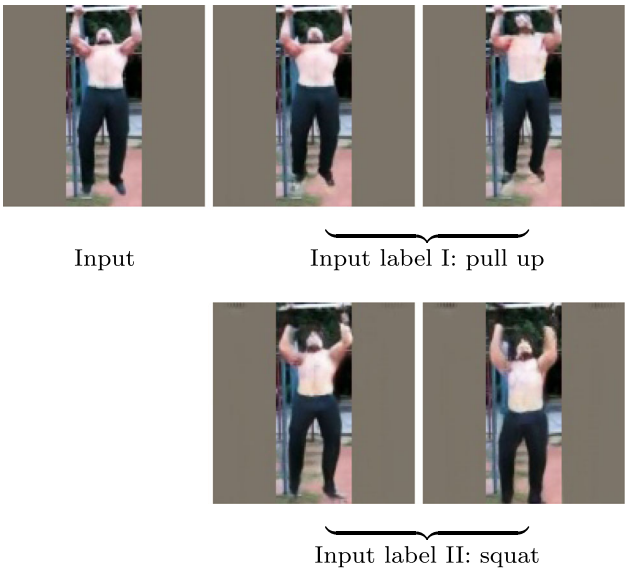


Fig. 1 Given an input image, our method can synthesize future sequences and control their action types with the input action label

aspect, frames at different time steps should be consistent to ensure the causality of the action sequences, which has been considered by existing methods. For the second aspect, the synthesized image sequence should be in accord with the given input action label, which has not been achieved by previous works.

In this paper, we treat the human pose as the high-level structure for action video prediction, and the predicted pose sequences are utilized as guidance for the synthesis of image sequences. This setting can avoid the interference of action-irrelevant appearance (Villegas et al., 2017; Zhao et al., 2018) and thus usually outperform strategies of directly hallucinating images.

To achieve modeling for the human pose, we propose Action Conditional Temporal Variational AutoEncoder (ACT-VAE) to describe the motion patterns and diversity, individually maintaining temporal coherence for each action category (so-called "individual temporal coherence"). It is built upon a distinctive Recurrent Neural Network (RNN) (Mikolov et al., 2010; Greff et al., 2016) to maintain such coherence. Similar to Kim et al. (2019); Yang et al. (2018), we employ human key points as the representation of pose, and ACT-VAE predicts key points of the future pose sequence, based on the pose of the input image as well as an action label, as shown in Fig. 2. We introduce action labels to the input and intermediate states of RNN for explicitly controlling what action to generate. Besides of the individual temporal coherence, compared with existing approaches (Kim et al., 2019; Castrejon et al., 2019; Wichers et al., 2018; Yuan & Kitani, 2020), we incorporate novel temporal modeling on latent variables into ACT-VAE that improves motion prediction accuracy. It updates the latent variable at each time step via the previous action features and the latent variables. Moreover, extensive experiments validate its notable precision improvement for forecasting and comparable diversity in prediction with state-of-the-art methods (Kim et al., 2019; Castrejon et al., 2019; Yuan & Kitani, 2020). In summary, our novelty in ACT-VAE contains two aspects. First, our proposed new network structure (new network conditions and intermediate states) leads to novel temporal modeling, which is different from current approaches, including VAE (Li et al., 2018; Kim et al., 2019; Lee et al., 2018; Yan et al., 2018; Babaeizadeh et al., 2017; Kumar et al., 2019; Razavi et al., 2019; Aliakbarian et al., 2020), RNN (Villegas et al., 2017; Wichers et al., 2018; Zhao et al., 2018), and their combinations (Castrejon et al., 2019; Denton & Fergus, 2018; Minderer et al., 2019). Second, we propose the corresponding training strategy. Different from previous methods, our training strategy optimizes the distribution gap between the synthesized and real pose sequences, simultaneously allowing the identical sampling pipeline for the generation during the training and inference procedure.

Furthermore, ACT-VAE can synthesize image sequences by connecting it with a plug-and-play network that maps pose to images. To this, we design a Pose-to-Image (P2I) network to convert the predicted pose sequence from ACT-VAE into the image sequence with realistic appearances. To improve the synthesis, we explicitly disentangle the foreground part from the image sequence via an attention mechanism, and enhance synthesized results further by introducing action conditional batch normalization (ACBN) to the P2I network. The foreground attention mechanism can push the model to pay more attention to the areas that need to be synthesized; ACBN points out that different action features in the generator should be separated for better synthesis effects. These two principles can be applied in various temporal human action modeling networks to improve their performance.

Extensive experiments on Penn-action dataset (Zhang et al., 2013) and Human3.6M dataset (Ionescu et al., 2013) and NTU RGB+D dataset (Shahroudy et al., 2016) show the effectiveness of our method. Our overall contribution is threefold.

- We explicitly model individual temporal coherence for human action video prediction of diverse action types.
- We build ACT-VAE with novel temporal modeling on latent variables, improving the accuracy of action video prediction to a new level and simultaneously keep comparable diversity with existing methods.
- We show that ACT-VAE is very general and is applicable to synthesize plausible videos by connecting it with a plug-and-play P2I network. Significantly, our framework is flexible to generate various action types from single input, via controlling action labels.

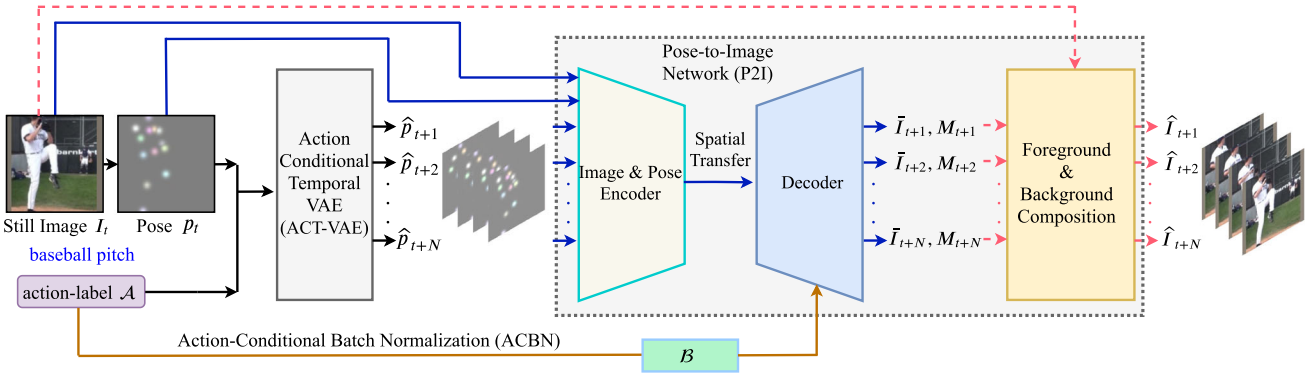


Fig. 2 Our framework for human action video prediction from a still image. It consists of two modules: ACT-VAE and P2I networks. ACT-VAE can generate future pose sequences with novel temporal modeling on latent variables to achieve the individual temporal coherence (its structure is shown in Fig. 3). ACT-VAE can further synthesize image sequences by connecting it with the plug-and-play P2I network

2 Related Work

2.1 Human Image Synthesis

We summarize existing works for human action video prediction,i.e., generating image sequences with the image and other conditions.

I) The target is single image synthesis, i.e., given the source image and conditions, the target is to generate the desired image. The conditions mainly contain the pose-guided (Cui et al., 2021; Zhu et al., 2019; Tang et al., 2020; Li et al., 2021; Ge et al., 2021a, b), text-guided (Jiang et al., 2022; Zhou et al., 2019; Liu et al., 2022), sketch-based image synthesis (Ho et al., 2020; Chen & Hays, 2018; Ghosh et al., 2019). And there are also some human image synthesis works without conditions, so-called unconditional human image synthesis (Frühstück et al., 2022; Fu et al., 2022; Karras et al., 2020).

II) Different from the single image synthesis, another target is image sequence/video synthesis, which is also our work's target. The most common topics include animation (Yoon et al., 2021), pose retargeting (Zhu et al., 2022), future forecast (Yang et al., 2018), etc. Conditions mainly contain pose-guided (Yoon et al., 2021; Zhu et al., 2022; Yang et al., 2020; Gafni et al., 2021; Wu et al., 2020; Kappel et al., 2021), text-guided (Han et al., 2022; Li et al., 2021; Balaji et al., 2019), speech-guided (Siyao et al., 2022; Ren et al., 2020; Guo et al., 2021), actioon-label-based (Kim et al., 2019; Yang et al., 2018). *Especially, our works belong to the category of forecast, and consider the challenge and practical setting where we aim to generate image sequences with the source image and the action label as the guidance. And our framework allows users to control the action type in the generated sequences.*

III) There are also some works whose targets are to predict the pose sequences. The first category is the pose prediction in the current two-stage-based image sequence synthesis framework. They use different conditions to generate sequences, including pose-guided (Yang et al., 2020; Zhu et al., 2022; Kappel et al., 2021; Gafni et al., 2021), text-guided, speech-guided (Ren et al., 2020; Guo et al., 2021), actioon-label-based (Kim et al., 2019; Yang et al., 2018) conditions. There are some works whose targets are only human pose sequence synthesis with different conditions, including pose (Mao et al., 2020, 2019), text (Guo et al., 2022; Ahuja & Morency, 2019), and speech (Lee et al., 2019; Zhuang et al., 2022).

2.2 Action Video Prediction

Some existing works for human action video prediction adopt deterministic models that directly minimize the distance between the synthesized action frames and the real frames, to produce deterministic image sequences (Finn et al., 2016; Jia et al., 2016; Yoo et al., 2017; Kalchbrenner et al., 2017; Villegas et al., 2017; Zhao et al., 2018; Wang et al., 2019; Kwon & Park, 2019; Guen & Thome, 2020) or future pose (Li et al., 2018; Zhao et al., 2018; Wang et al., 2019; Guo & Choi, 2019; Gopalakrishnan et al., 2019; Mao et al., 2019; Cai et al., 2020; Mao et al., 2020; Piergiovanni et al., 2020). Their corresponding performances are generally limited since their results may converge to the average of possible outcomes. To achieve more realistic and dynamic predictions, recent methods employ deep generative models, including VAE (Kingma & Welling, 2014), VRNN (Chung et al., 2015) and GAN (Goodfellow et al., 2014).

GAN-based approaches extend the structure of vanilla GAN into the sequential one (Wichers et al., 2018; Lee et al., 2018; Cai et al., 2018; Mathieu et al., 2015). Cai et al. (2018) predicted human pose sequences with a latent variable in adversarial learning. The basic idea is to construct a discriminator to classify the realness of synthesized sequences and corresponding real ones. It updates the generator to pass

the discriminator with good-quality generated sequences. On the other hand, using VAE (Li et al., 2018; Kim et al., 2019; Lee et al., 2018; Yan et al., 2018; Babaeizadeh et al., 2017; Kumar et al., 2019; Razavi et al., 2019; Aliakbarian et al., 2020) can also achieve promising performance. Kim et al. (2019) extended VAE with RNN structure, and set a common latent variable for predicting overall time steps. Lee et al. (2018) utilized multiple latent variables with identical distribution for prediction at each time step during inference. Babaeizadeh et al. (2017) modeled motion patterns and diversity with a single set of fixed latent variables for prediction. These strategies do not consider temporal coherence during inference.

Video prediction can also use VRNN (Castrejon et al., 2019; Denton & Fergus, 2018; Minderer et al., 2019). Denton and Fergus (2018) proposed a VRNN framework with a learned prior for inference. Castrejon et al. (2019) improved the performance by extending hierarchical structures for latent variables of VRNN. These approaches do not maintain individual temporal coherence for each action category, and thus are different from our work. Besides, our framework's modeling on latent variables varies from the current VRNN.

One crucial issue about human action video prediction is the use of structural information. Some existing works (Srivastava et al., 2015; Xu et al., 2018; Oliu et al., 2018; Tulyakov et al., 2018; Wichers et al., 2018; Lee et al., 2018; Cai et al., 2018) directly synthesized action frames from networks and achieved success on simple datasets with low motion variance and image resolution. With the advance in pose-guided image generation (Zhu et al., 2019; Neverova et al., 2018; Ma et al., 2017; Siarohin et al., 2019b, a), recent methods favored a two-stage strategy to generate pose sequences firstly and then use them as conditions to hallucinate image sequences (Li et al., 2018; Kim et al., 2019; Zhao et al., 2018; Villegas et al., 2017; Wang et al., 2018; Yang et al., 2018; Walker et al., 2017).

Moreover, besides human motion prediction, there are also some works about human trajectory prediction (Huang et al., 2019; Kothari et al., 2021; Chen et al., 2021; Duan et al., 2022; Adeli et al., 2021). These works are mainly built on the image or spatial position conditions, and aim to estimate the future state of trajectory. *Trajectory prediction is a different task from motion prediction and is not our target*. Meanwhile, our ACT-VAE is different from theirs in terms of temporal modeling, model structure, and optimization strategy, since our approach can simultaneously guarantee temporal coherence and high accuracy within each class, and has high controllability for generation types.

The ill-posed property of previous works' solutions is severe since they have not individually achieved temporal coherence for each action category, resulting in the solution space, which is extremely larger than the distribution of real image sequences. To reduce the ill-posed degree, we are the first to formulate the individual temporal coherence for the outputs of ACT-VAE. The distribution of image sequences with the individual temporal coherence is close to the distribution of real ones. As a result, our ACT-VAE weakens the ill-posed problem, improving the model accuracy and preserving diversity compared with existing methods.

3 Method

Following the task setting of Kim et al. (2019), our model predicts human actions by synthesizing N future RGB frames $\{\widehat{I}_{t+i}\}_{i=1,\ldots,N}$ for an initial image $I_t \in \mathbb{R}^{H\times W\times 3}$ with its target action label $\mathcal{A} \in \mathbb{R}^C$ (one-hot vector form) as

$$\{\widehat{I}_{t+i}\}_{i=1,\ldots,N} = \mathcal{F}(I_t|\mathcal{A}), \tag{1}$$

where $\mathcal{F}$ is the desired action video prediction model, I denotes a real image with pose p, and $\widehat{I}$ denotes a synthesized image with pose $\widehat{p}$. t and $t+i$ index time in a video. H and W denote image height and width respectively. N and C are the length of synthesized frames and the number of action categories to be modeled.

Sequential action modeling should be independent of object appearance and background. To this end, we propose a framework consisting of two modules, which are ACT-VAE and P2I networks, as shown in Fig. 2. With the pose of the initial input image I_t and the target action label $\mathcal{A}$, ACT-VAE generates pose sequences in key point form. ACT-VAE further produces realistic videos by connecting it with the P2I network that is a plug-and-play module.

3.1 Action Conditional Temporal VAE

Given an initial image I_t, its initial pose p_t (we employ the setting of Villegas et al. (2017); Zhao et al. (2018) to set p_t for I_t), and target action label $\mathcal{A}$, our proposed ACT-VAE predicts future pose sequence as

$$\{\widehat{p}_{t+i}\}_{i=1,\ldots,N} = \mathcal{T}(p_t|\mathcal{A}), \tag{2}$$

where $\mathcal{T}$ is ACT-VAE. Both p_t and $\widehat{p}_t$ can be represented in the form of key point coordinate values (Kim et al., 2019).

The design of ACT-VAE is motivated by the observation that a particular kind of action should have a distinctive primary motion pattern. Meanwhile, it may exhibit diverse local details for different persons as the diversity of motion. For example, batting is a standard motion pattern in baseball while everybody's batting differs from each other a bit. Such motion pattern and regional diversity are supposed to be temporally correlated for realism, and each action category

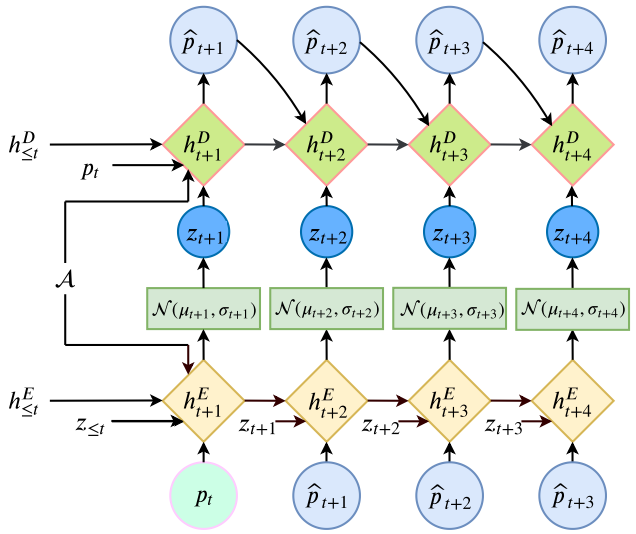


Fig. 3 The architecture of ACT-VAE to synthesize future pose sequence based on an initial human pose p_t and an action label $\mathcal{A}$. ACT-VAE consists of the encoder E and the decoder D, which are both implemented with LSTM networks. Four future frames from the time index t are generated here, and ACT-VAE can indeed synthesize the pose sequence with arbitrary length. Moreover, $h^E_{\leq t}$ and $h^D_{\leq t}$ are set as the initial hidden state of E and D. In addition, we feed action labels into ACT-VAE by concatenating them with other input variables of LSTM at each time step

should have its individual temporal coherence. The individual temporal coherence for human action video prediction of diverse action types in our ACT-VAE is achieved not only by the novel model structure, but also the temporal modeling of latent variables and the corresponding new training strategy for ACT-VAE.

Thus, unlike previous VAE-based video prediction methods (Kim et al., 2019; Li et al., 2018) that take identical latent variables as the condition for the generation across all time steps, we equip VAE with the property of temporal coherence for various action categories. As shown in Fig. 3, ACT-VAE is different from conventional VAE by modeling the temporally correlated latent variable $\{z_{t+i}\}_{i=1,\ldots,N}$ and pose $\{\widehat{p}_{t+i}\}_{i=1,\ldots,N}$ with a distinctive recurrent structure, and using different action categories as the condition. Especially, ACT-VAE models the latent variable at each time step with features of both previous pose and latent variables. For variable y, $y_{<t}$ denotes the sequence $\{y_{t'}\}_{t'<t}$, and $y_{\leq t}$ represents $\{y_{t'}\}_{t'\leq t}$.

3.1.1 Structure of ACT-VAE

As shown in Fig. 3, ACT-VAE has an encoder E and a decoder D both in recurrent manner using LSTM (Greff et al., 2016). The encoder E is to sample latent variable $\{z_{t+i}\}_{i=1,\ldots,N}$, if the initial input pose of ACT-VAE is denoted as p_t. In such process, the sampling of z_{t+i} is implemented by modeling joint posterior distribution of $z_{<t+i}$ and $\widehat{p}_{<t+i}$ conditioned by the action label $\mathcal{A}$ as

$$\begin{aligned}(\mu_{t+i}, \sigma_{t+i}, h^E_{t+i}) &= E(h^E_{t+i-1}, z_{t+i-1}, \widehat{p}_{t+i-1}|\mathcal{A}),\\ z_{t+i} &\sim \mathcal{N}(\mu_{t+i}, \sigma_{t+i}),\end{aligned} \tag{3}$$

where $\mathcal{N}(\mu, \sigma)$ is normal distribution with mean value μ and standard deviation σ, $\sim$ is the operation of sampling, and h^E_{t+i-1} is the hidden state of the encoder E which contains information from $z_{<t+i-1}$ and $\widehat{p}_{<t+i-1}$.

The decoder D can synthesize pose sequence recurrently according to the joint posterior distribution of $z_{\leq t+i}$ and $\widehat{p}_{<t+i}$ conditioned by $\mathcal{A}$ as

$$(\widehat{p}_{t+i}, h^D_{t+i}) = D(\widehat{p}_{t+i-1}, h^D_{t+i-1}, z_{t+i}|\mathcal{A}), \tag{4}$$

where h^D_{t+i-1} is the hidden state of D, which involves information from $\widehat{p}_{<t+i-1}$ and $z_{<t+i}$.

Our structure is the first framework that can achieve controllable action video prediction with "individual temporal coherence", individually maintaining temporal coherence for each action category. Different from previous RNN, VAE models, and their combinations, in ACT-VAE, the priors of the action category are explicitly led to the input and intermediate states of RNN with VAE's latent modeling, using the fusion strategy of the neural forwarding module. Also, novel temporal modeling on latent variables is designed in ACT-VAE, where the latent variable at each time step is queried based on previous action features and latent variables. Such a temporal modeling manner is distinct from existing approaches. VRNNs (Castrejon et al., 2019; Denton & Fergus, 2018; Minderer et al., 2019) mainly formulate latent variables z as $z_t \sim \mathcal{Q}_\phi(z_t|x_{<t})$ where $x_{<t}$ is the previous observation (e.g., pose sequences). ACT-VAE will model z as $z_t \sim \mathcal{Q}_\phi(z_t|z_{<t}, x_{<t})$. Moreover, our temporal modeling is also varying from the VAE networks (Li et al., 2018; Kim et al., 2019; Lee et al., 2018; Yan et al., 2018; Babaeizadeh et al., 2017; Kumar et al., 2019; Razavi et al., 2019; Aliakbarian et al., 2020), which implement temporal coherence with temporal-varying conditions, i.e., they model z as $z_t \sim \mathcal{Q}_\phi(z_t|z_{<t})$.

3.1.2 Learning of ACT-VAE

Suppose the initial input image is denoted as I_0 and its pose is p_0, ACT-VAE predicts pose sequence for action category $\mathcal{A}$ as $\{\widehat{p}_i\}_{i=1,\ldots,N}$ (i.e., $\widehat{p}_{\leq N}$). Thus, ACT-VAE is to synthesize future pose by optimizing conditional posterior probability $\mathcal{P}_\theta(\widehat{p}_{\leq N}|p_0, \mathcal{A})$, which is approximated by a network with parameters θ. Directly computing $\mathcal{P}_\theta(\widehat{p}_{\leq N}|p_0, \mathcal{A})$ is intractable since it is difficult to compute its probability density function. In VAE (Kingma & Welling, 2014), regarding the probability distribution of $\mathcal{P}_\theta(\widehat{p}_{\leq N}|p_0, \mathcal{A})$, we maximize

its lower bound instead. And this lower bound can be obtained with Jensen's inequality as

$$\begin{aligned}&\ln(\mathcal{P}_\theta(\widehat{p}_{\leq N}|p_0,\mathcal{A}))\\&\geq \mathbb{E}_{z\sim\mathcal{Q}_\phi(z|\widehat{p},\mathcal{A})}\left[\ln\frac{\mathcal{P}_\theta(\widehat{p}_{\leq N},z|p_0,\mathcal{A})}{\mathcal{Q}_\phi(z|\widehat{p},\mathcal{A})}\right],\end{aligned}\tag{5}$$

where $\mathcal{Q}_\phi(z|\widehat{p},\mathcal{A})$ is a posterior distribution and ln is the operation of computing natural logarithm. We further notice that $\mathcal{P}_\theta(\widehat{p}_{\leq N},z|p_0,\mathcal{A})$ and $\mathcal{Q}_\phi(z|\widehat{p},\mathcal{A})$ can be decomposed as:

$$\begin{aligned}&\ln(\mathcal{P}_\theta(\widehat{p}_{\leq N},z|p_0,\mathcal{A}))\\&=\ln(\prod_{t=1:N}\mathcal{P}_\theta(\widehat{p}_t|z_{\leq t},\widehat{p}_{<t},\mathcal{A})\mathcal{P}_\phi(z_t)),\end{aligned}\tag{6}$$

$$\begin{aligned}&\ln(\mathcal{Q}_\phi(z|\widehat{p},\mathcal{A}))\\&=\ln(\prod_{t=1:N}\mathcal{Q}_\phi(z_t|z_{<t},\widehat{p}_{<t},\mathcal{A})),\end{aligned}\tag{7}$$

where $\mathcal{P}_\theta(\widehat{p}_t|z_{\leq t},\widehat{p}_{<t},\mathcal{A})$ and $\mathcal{Q}_\phi(z_t|z_{<t},\widehat{p}_{<t},\mathcal{A})$ are two posterior distributions that are approximated by D and E in ACT-VAE. $\mathcal{P}_\phi(z_t)$ is the prior distribution for z_t. According to Eqs. (5), (6), and (7), we can obtain the lower bound of $\ln(\mathcal{P}_\theta(\widehat{p}_{\leq N}|p_0,\mathcal{A}))$ as

$$\begin{aligned}&\mathbb{E}_{z\sim\mathcal{Q}_\phi(z|\widehat{p},\mathcal{A})}\Bigg[\sum_{t=1:N}(\ln(\mathcal{P}_\theta(\widehat{p}_t|z_{\leq t},\widehat{p}_{<t},\mathcal{A}))\\&+\ln(\mathcal{P}_\phi(z_t))-\ln(\mathcal{Q}_\phi(z_t|z_{<t},\widehat{p}_{<t},\mathcal{A}))\Bigg].\end{aligned}\tag{8}$$

Moreover, it is trivial to obtain the expression of an objective to optimize when the initial input pose is denoted as p_t, by replacing the corresponding time index. We will show the superiority of this novel optimization objective for the accuracy of prediction with experiments in Sect. 4. Note

$$\begin{aligned}&\mathbb{E}_{z\sim\mathcal{Q}_\phi}\Big[\ln(\mathcal{P}_\phi(z_t))-\ln(\mathcal{Q}_\phi(z_t|z_{<t},\widehat{p}_{<t},\mathcal{A}))\Big]\\&=-\mathbb{E}_{z\sim\mathcal{Q}_\phi}\left[\ln\frac{\mathcal{Q}_\phi(z_t|z_{<t},\widehat{p}_{<t},\mathcal{A})}{\mathcal{P}_\phi(z_t)}\right],\end{aligned}\tag{9}$$

which is the negative KL-divergence between two distributions of $\mathcal{Q}_\phi(z_t|z_{<t},\widehat{p}_{<t},\mathcal{A})$ and $\mathcal{P}_\phi(z_t)$. This is the cost function of the encoder in ACT-VAE.

For the decoder, its objective is $\mathcal{P}_\theta(\widehat{p}_t|z_{\leq t},\widehat{p}_{<t},\mathcal{A})$ in Eq. (8). Maximizing it leads the predicted pose sequence to be close to its ground truth. It is trivial to obtain the expression of an objective to optimize when the initial input image is denoted as I_t, by replacing the corresponding time index.

3.1.3 Training Objective

If the input pose of ACT-VAE is p_t and we predict N frames, then the optimization target of ACT-VAE is to minimize the distance $\mathcal{L}_{dis}$ and KL-divergence $\mathcal{L}_{div}$, as

$$\begin{aligned}\mathcal{L}_{dis}&=\sum_{t'=t+1}^{t+N}||\widehat{p}_{t'}-p_{t'}||_1,\\\mathcal{L}_{div}&=\sum_{t'=t+1}^{t+N}\mathrm{KL}(\mathcal{N}(\mu_{t'},\sigma_{t'}),\mathcal{P}_\phi(z_{t'})),\\\mathcal{L}_{vae}&=\lambda_1\mathcal{L}_{dis}+\lambda_2\mathcal{L}_{div},\end{aligned}\tag{10}$$

where $\mathcal{L}_{vae}$ is the loss to optimize for ACT-VAE, $\widehat{p}_{t'}$ is the generated pose from ACT-VAE, and $p_{t'}$ is its corresponding ground truth at time t'. The prior distribution $\mathcal{P}_\phi(z_t)$ is assumed to be the standard normal distribution $\mathcal{N}(0,\mathrm{I})$. KL is to compute the KL-divergence between two distributions. λ_1 and λ_2 are loss weights that are obtained by using the grid search on the validation set. Note that the action label $\mathcal{A}$ should be consistent with the input image/pose during training, while can be inconsistent with the input during inference to control which action type to generate.

3.1.4 Inference

Given the pose of input image p_t and the target action label $\mathcal{A}$, we aim to generate pose sequence $\{\widehat{p}_{t+i}\}_{i=1,\ldots,N}$ during inference. To obtain $\widehat{p}_{t+i}$, we first sample latent variable z_{t+i} with Eq. (3) and then use it to compute $\widehat{p}_{t+i}$ with Eq. (4). Obviously, the process to sample $\{z_{t+i}\}_{i=1,\ldots,N}$ and generate the pose sequence $\{\widehat{p}_{t+i}\}_{i=1,\ldots,N}$ is same for both training and inference.

Our modeling on latent variables differs from the current VRNN works (Castrejon et al., 2019; Denton & Fergus, 2018): VRNN models z_t with the posterior $z_t\sim\mathcal{Q}_\phi(z_t|z_{<t},\widehat{p}_{\leq t})$ during training and the prior $z_t\sim\mathcal{P}_\phi(z_t|z_{<t},\widehat{p}_{<t})$ during inference; ACT-VAE models z_t with $z_t\sim\mathcal{Q}_\phi(z_t|z_{<t},\widehat{p}_{<t})$ for both training and inference, resulting in higher accuracy and diversity as proved in Sect. 4.5.2. Besides, ACT-VAE also differs from SVG-FP (Denton & Fergus, 2018) that models z_t with a fixed prior during inference.

The novel temporal modeling during training and inference has several significant benefits compared with previous methods. First, although previous approaches have considered the temporal consistency among different frames, our new modeling strategy can maintain temporal coherence for each action category, i.e., the synthesized image sequence will be in accord with the given input action label to control what action to generate. Second, our temporal modeling allows the same sampling pipeline for training and inference, decreasing the gap between the training and inference phase

and thus improving the generalization. Third, our temporal modeling has new temporal conditional modeling for latent variables (different from existing approaches including VAE, RNN, and others, as indicated in Sects. 2 and 3.1.1), which can enhance the prediction accuracy as proved by experiments in Sect. 4.

Moreover, as a general approach, ACT-VAE can synthesize image sequences, by connecting it with a plug-and-play network that maps pose sequences to image sequences. To this, we design an effective Pose-to-Image network.

3.2 Pose-to-Image Network

P2I network (denoted as $\mathcal{G}$) predicts a realistic image sequence $\{\widehat{I}_{t+i}\}_{i=1,\ldots,N}$ by taking input of a pose sequence $\{p_{t+i}\}_{i=1,\ldots,N}$ (or $\{\widehat{p}_{t+i}\}_{i=1,\ldots,N}$), a still image I_t with its pose p_t, and an action label $\mathcal{A}$. We employ the encoder-decoder structure in Zhu et al. (2019) as the backbone, with our attention mechanism and conditional batch normalization.

3.2.1 Foreground Attention

Considering the elusive variance in the background of human videos, directly generating an image sequence tends to yield severe artifacts in the background. For the nearly static background in I_t, we exploit foreground-background composition with an attention mechanism. It makes the generator concentrate on foreground synthesis, which is our main focus in this paper. Other background synthesis methods will be our future work. Generally, given I_t and p_t and p_{t+i}, the procedure to synthesize the target frame is

$$\begin{aligned}(\bar{I}_{t+i}, M_{t+i}) &= \mathcal{G}(I_t, p_t, p_{t+i}),\\ \widehat{I}_{t+i} &= \bar{I}_{t+i} \odot M_{t+i} + I_t \odot (1 - M_{t+i}),\end{aligned} \tag{11}$$

where $\widehat{I}_{t+i}$ is the generated frame, M_{t+i} is a soft mask indicating foreground, and $\odot$ refers to Hadamard product. This procedure is denoted as $\widehat{I}_{t+i} = \mathcal{G}_{\mathcal{M}}(I_t, p_t, p_{t+i})$.

3.2.2 Action Conditional Batch Normalization

We utilize task-related conditions into normalization operations to improve results (Perez et al., 2018; Clark et al., 2019) through incorporating action conditional batch normalization (ACBN) into $\mathcal{G}$. This design is based on the assumption that statistics of intermediate feature maps in $\mathcal{G}$ for each action category should be distinctive. From this perspective, we assign affine transformation parameters $\gamma_{\mathcal{A}}$ and $\tau_{\mathcal{A}}$ for BN operations in the decoder of $\mathcal{G}$ with the condition of the action label $\mathcal{A}$ as

$$\widehat{x} = \gamma_{\mathcal{A}} \frac{x - \nu_x}{\sqrt{\rho_x^2 + \epsilon}} + \tau_{\mathcal{A}}, \ (\gamma_{\mathcal{A}}, \tau_{\mathcal{A}}) = \mathcal{B}(\mathcal{A}), \tag{12}$$

where ν_x and ρ_x are mean and variance computed from input feature map x, and ϵ is a small positive constant for numerical stability. $\gamma_{\mathcal{A}}$ and $\tau_{\mathcal{A}}$ for each BN operation in decoder are predicted by $\mathcal{A}$ from a network $\mathcal{B}$ as shown in Fig. 2. $\mathcal{B}$ is an embedding layer with the input of action labels.

3.2.3 Training P2I Network

The training objective of the P2I network consists of reconstruction and adversarial loss. We employ both pixel-level and perceptual-level (Johnson et al., 2016; Zhu et al., 2017) reconstruction loss as

$$\mathcal{L}_{rec} = \sum_{i=1}^{N} \sum_{m=0}^{5} \mathbb{E}(||\Phi_m(\widehat{I}_{t+i}) - \Phi_m(I_{t+i})||_1), \tag{13}$$

where $\widehat{I}_{t+i} = \mathcal{G}_{\mathcal{M}}(I_t, p_t, p_{t+i})$, $\mathbb{E}$ is the operation to compute mean value, $\Phi_0(\cdot)$ is the raw pixel space, $\Phi_1(\cdot)$ to $\Phi_5(\cdot)$ are five feature spaces of an ImageNet-pretrained VGG-16 network (Johnson et al., 2016). Further, adversarial learning is proved to be effective in human video synthesis (Yang et al., 2018; Wang et al., 2018). Thus, we use the form of LS-GAN (Mao et al., 2017) for adversarial learning, as

$$\begin{aligned}\mathcal{L}_{gan_d} &= \mathbb{E}((\mathcal{D}(I_{t+i} \dagger p_{t+i}) - 1)^2)\\ &\quad + \mathbb{E}((\mathcal{D}(\widehat{I}_{t+i} \dagger p_{t+i}))^2),\\ \mathcal{L}_{gan_g} &= \mathbb{E}((\mathcal{D}(\widehat{I}_{t+i} \dagger p_{t+i}) - 1)^2),\end{aligned} \tag{14}$$

where $i \in [1, N]$, $\dagger$ is the operation of channel concatenation, $\mathcal{D}$ is the discriminator and $\mathcal{L}_{gan_d}$ is its loss term, $\mathcal{L}_{gan_g}$ is the loss term for the P2I network. To stabilize adversarial learning, we utilize feature match loss $\mathcal{L}_{aux}$ (Wang et al., 2018) as an auxiliary part of the adversarial loss, which is the distance between real and fake samples in the feature space of $\mathcal{D}$. Compared with the generation with $\{p_{t+i}\}_{i=1,\ldots,N}$, the generation with $\{\widehat{p}_{t+i}\}_{i=1,\ldots,N}$ is harder since there is no ground truth. To address this issue, we adopt a term of adversarial loss, similar to Eq. (14), as

$$\begin{aligned}\mathcal{L}_{gan_{\widehat{d}}} &= \mathbb{E}((\mathcal{D}(\mathcal{G}_{\mathcal{M}}(I_t, p_t, \widehat{p}_{t+i}) \dagger \widehat{p}_{t+i}) - 0)^2),\\ \mathcal{L}_{gan_{\widehat{g}}} &= \mathbb{E}((\mathcal{D}(\mathcal{G}_{\mathcal{M}}(I_t, p_t, \widehat{p}_{t+i}) \dagger \widehat{p}_{t+i}) - 1)^2).\end{aligned} \tag{15}$$

In summary, the overall loss terms for $\mathcal{G}$ and $\mathcal{D}$ are

$$\begin{aligned}\mathcal{L}_d &= \lambda_3 (\mathcal{L}_{gan_d} + \mathcal{L}_{gan_{\widehat{d}}}),\\ \mathcal{L}_g &= \lambda_4 (\mathcal{L}_{gan_g} + \mathcal{L}_{gan_{\widehat{g}}}) + \lambda_5 \mathcal{L}_{aux} + \lambda_6 \mathcal{L}_{rec},\end{aligned} \tag{16}$$

Algorithm 1 Training algorithm for ACT-VAE

Input: training image sequences, with their pose sequences and action labels. Initialized ACT-VAE including the encoder E and the decoder D;

for $i = 0$ to $maxIters$ **do**
  for each training tuple $(p_t, ..., p_{t+N}, \mathcal{A})$ in dataset **do**
    Initialize hidden state of E as $\mathbf{0}$ and set it as $h^E_{\leq t}$;
    Initialize $z_{\leq t} \sim \mathcal{N}(0, I)$;
    $(\mu_{t+1}, \sigma_{t+1}, h^E_{t+1}) \leftarrow E(p_t, h^E_{\leq t}, z_{\leq t}|\mathcal{A})$;
    $z_{t+1} \sim \mathcal{N}(\mu_{t+1}, \sigma_{t+1})$;
    Initialize hidden state of D as $\mathbf{0}$ and set it as $h^D_{\leq t}$;
    $(\widehat{p}_{t+1}, h^D_{t+1}) \leftarrow D(p_t, h^D_{\leq t}, z_{t+1}|\mathcal{A})$;
    for $k = 1$ to $N - 1$ **do**
      $(\mu_{t+k+1}, \sigma_{t+k+1}, h^E_{t+k+1}) \leftarrow E(\widehat{p}_{t+k}, h^E_{t+k}, z_{t+k}|\mathcal{A})$;
      $z_{t+k+1} \sim \mathcal{N}(\mu_{t+k+1}, \sigma_{t+k+1})$;
      $(\widehat{p}_{t+k+1}, h^D_{t+k+1}) \leftarrow D(\widehat{p}_{t+k}, h^D_{t+k}, z_{t+k+1}|\mathcal{A})$;
    end for
    Compute $\mathcal{L}_{vae}$ in Eq. (10) using $(\widehat{p}_{t+1}, ..., \widehat{p}_{t+N})$, $(p_{t+1}, ..., p_{t+N})$, $(z_{t+1}, ..., z_{t+N})$, λ_1 and λ_2;
    Update the weights of E and D;
  end for
end for
return trained encoder E and decoder D of ACT-VAE;

Algorithm 2 Training algorithm for P2I network

Input: training image sequences, with their pose sequences and action labels.

for $i = 0$ to $maxIters$ **do**
  for each tuple $[(I_t, ..., I_{t+N}), (p_t, ..., p_{t+N}), \mathcal{A}]$ in dataset **do**
    for $k = 1$ to N **do**
      $\widehat{I}_{t+k} \leftarrow \mathcal{G}_{\mathcal{M}}(I_t, p_t, p_{t+k})$;
    end for
    Compute $\mathcal{L}_{rec}$ by Eq. (13);
    Compute $\mathcal{L}_{gan_g}$ and $\mathcal{L}_{aux}$ by Eq. (14);
    Compute $\lambda_4 \mathcal{L}_{gan_g} + \lambda_5 \mathcal{L}_{aux} + \lambda_6 \mathcal{L}_{rec}$, update $\mathcal{G}$;
    Compute $\lambda_3 \mathcal{L}_{gan_d}$ by Eq. (14), update $\mathcal{D}$;
    Take p_t, $\mathcal{A}$ as input, ACT-VAE outputs $(\widehat{p}_{t+1}, ..., \widehat{p}_{t+N})$;
    for $k = 1$ to N **do**
      $\widehat{I}_{t+k} \leftarrow \mathcal{G}_{\mathcal{M}}(I_t, p_t, \widehat{p}_{t+k})$;
      Compute $\lambda_4 \mathcal{L}_{gan_{\widehat{g}}}$ by Eq. (15) to update $\mathcal{G}$;
      Compute $\lambda_3 \mathcal{L}_{gan_{\widehat{d}}}$ by Eq. (15) to update $\mathcal{D}$;
    end for
  end for
end for
return trained P2I network $\mathcal{G}$;

where λ_3 to λ_6 are loss weights that are set based on parameters of pose-guided image generation methods (Zhu et al., 2019; Wang et al., 2018) and grid search. $\mathcal{L}_d$ is the loss for $\mathcal{D}$ and $\mathcal{L}_g$ is the loss for $\mathcal{G}$. In our experiments, $\lambda_1 = 200$, $\lambda_2 = 0.002$, $\lambda_3 = 5$, $\lambda_4 = 5$, $\lambda_5 = 1$, $\lambda_6 = 30$.

3.3 Training and Inference Algorithm

To train our framework, ACT-VAE and P2I networks are trained separately. We denote the operation of sampling as $\sim$, the variable with all zero value as $\mathbf{0}$. The operation $A \leftarrow B$ means that we set the value of A as the output of B. Training algorithms are given below.

Algorithm 3 Inference algorithm for action prediction

Input: Trained ACT-VAE which includes encoder E and decoder D, trained P2I network $\mathcal{G}$, input still image I_t with its pose p_t and action label $\mathcal{A}$;
Initialize the hidden state of E as $\mathbf{0}$ and set it as $h^E_{\leq t}$;
Initialize $z_{\leq t} \sim \mathcal{N}(0, \mathrm{I})$;
$(\mu_{t+1}, \sigma_{t+1}, h^E_{t+1}) \leftarrow E(p_t, h^E_{\leq t}, z_{\leq t}|\mathcal{A})$;
$z_{t+1} \sim \mathcal{N}(\mu_{t+1}, \sigma_{t+1})$;
Initialize the hidden state of D as $\mathbf{0}$ and set it as $h^D_{\leq t}$;
$(\widehat{p}_{t+1}, h^D_{t+1}) \leftarrow D(p_t, h^D_{\leq t}, z_{t+1}|\mathcal{A})$;
for $k = 1$ to $N - 1$ **do**
  $(\mu_{t+k+1}, \sigma_{t+k+1}, h^E_{t+k+1}) \leftarrow E(\widehat{p}_{t+k}, h^E_{t+k}, z_{t+k}|\mathcal{A})$;
  $z_{t+k+1} \sim \mathcal{N}(\mu_{t+k+1}, \sigma_{t+k+1})$;
  $(\widehat{p}_{t+k+1}, h^D_{t+k+1}) \leftarrow D(\widehat{p}_{t+k}, h^D_{t+k}, z_{t+k+1}|\mathcal{A})$;
end for
for $k = 1$ to N **do**
  Compute $\widehat{I}_{t+k} = \mathcal{G}_{\mathcal{M}}(I_t, p_t, \widehat{p}_{t+k})$ by using $\mathcal{A}$ to compute parameters of ACBN for $\mathcal{G}$;
end for
return $\{\widehat{I}_{t+k}\}_{k=1,...,N}$ as predicted image sequence;

- To train ACT-VAE, we have training data as the pose sequences with their corresponding action labels. A detailed training algorithm for ACT-VAE is shown in Algorithm 1.
- To train P2I network, we need image sequences with their pose sequences and action labels. Its training procedure is shown in Algorithm 2.

One feature of our ACT-VAE is that the process of sampling latent variables and generating pose sequences are identical in training and inference. After we have trained the ACT-VAE and P2I networks, we can synthesize image sequence $\{\widehat{I}_{t+k}\}_{k=1,...,N}$ by using Algorithm 3, with the input as a still image I_t with action label $\mathcal{A}$.

3.4 Network Details

In experiments, the network configuration for each component (ACT-VAE and P2I network) in our framework is summarized as the following.

3.4.1 ACT-VAE

ACT-VAE is structured like an encoder-decoder, and both encoder and decoder are implemented as one-layer LSTM. The input size of encoder E is $J \times 2 + C + 512$ (J is the joint number in pose and C is the number of action categories), the hidden size is 1024, the output size is 512; the input size of decoder D is $J \times 2 + C + 512$, the hidden size is 26, the output size is $J \times 2$.

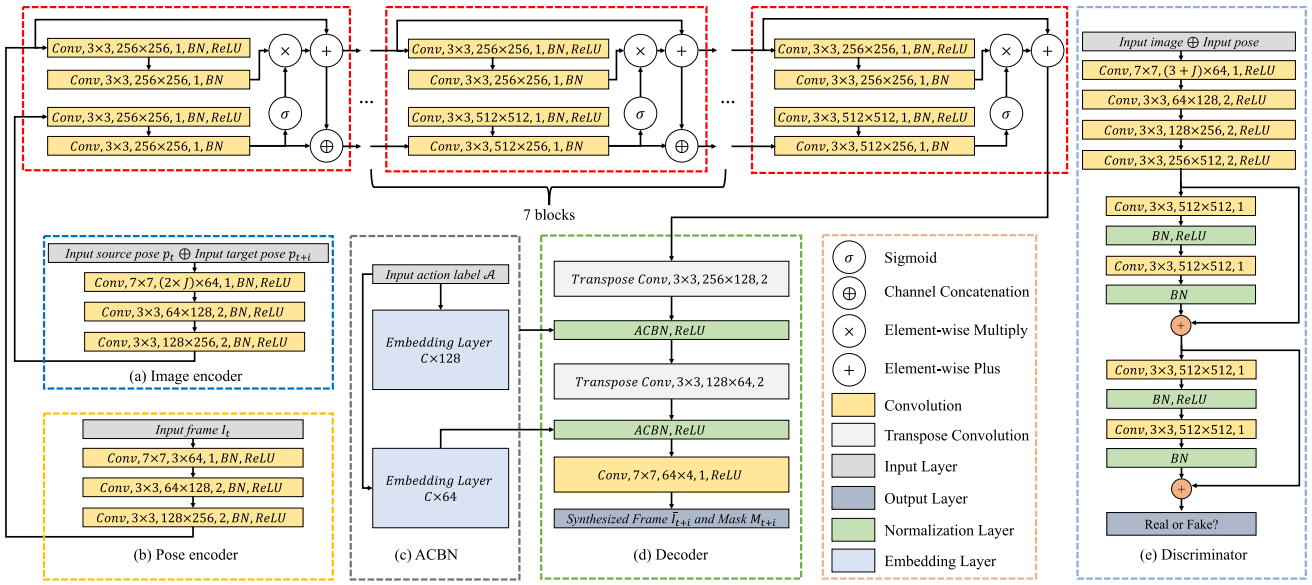


Fig. 4 Detailed structure of the P2I network

The fusion mechanisms of temporal variables and action labels can be summarized as the following. First, the action label $\mathcal{A}$ is sent through the embedding layer to obtain its vector representation, and its embedding is fused with the temporal latent variables with a multi-layer forwarding module, building the corresponding conditional probability distribution. And the fusion of h, z, $\widehat{p}$ in Eqs. 3 and 4 is completed by concatenation with position embedding to distinguish different input parts. And the fusion with the embedding of $\mathcal{A}$ is also implemented with a multi-layer forwarding module.

3.4.2 P2I Network

The P2I network has two encoder heads as image encoder E_I (Fig. 4a) and pose encoder E_P (Fig. 4b), and it contains an image decoder D_I (Fig. 4d). Besides, it is trained in adversarial manner, hence it is attached with a discriminator for training (Fig. 4e). In Fig. 4, "$Conv, 7 \times 7, 3 \times 64, 1, BN, ReLU$" means that this convolution layer adopts kernel size of 7×7 with stride size of 1, and has 3 input feature channels and 64 output feature channels. A batch normalization layer and an activation function ReLU is applied to the output of this convolution layer. Meaning of other convolution layers can be interpreted in the same way.

For the synthesis with input as I_t, p_t and p_{t+i}, the image encoder E_I and pose encoder E_P will transform I_t into image feature, and transform pair (p_t, p_{t+i}) into spatial feature. These features are then sent into pose-attentional transfer blocks (Zhu et al., 2019) (the red rectangles in Fig. 4) and the image decoder D_I to obtain two types of outputs: the produced image and the mask to distinguish foreground and background. To compute the parameters of ACBN in D_I, we use two 1D embedding layers $\mathcal{B}$ (Fig. 4c) with the input dimension as the number of action categories.

4 Experiments

4.1 Datasets

To verify our method's generality, we employ datasets containing various action categories, which are Penn-action (Zhang et al., 2013) and Human3.6M (Ionescu et al., 2013). Penn-action contains videos of humans in 15 sport action categories. The total number of videos is 2,326. For each video, 13 human joint annotations are provided as the ground truth of pose. We adopt the experimental setting of Kim et al. (2019) with 9 action categories for experiments, including baseball pitch, baseball swing, clean and jerk, pull ups, golf swing, tennis forehand, tennis serve, jumping jacks and squats. Besides, we follow the same train/test split of Kim et al. (2019) for a fair comparison.

Human3.6M contains various daily human actions, and this dataset provides 17 human joint annotations as the ground truth of pose. Moreover, to conduct experiments on the action category with obvious motion patterns, we choose 8 action categories from this dataset for experiments: directions, greeting, phoning, posing, purchases, walking, walking dog, and walking together. Moreover, we follow the same train/test split of Mao et al. (2019).

This paper focuses on the modeling of human action, thus we experiment on action datasets with static or simple backgrounds to minimize disturbances from backgrounds. This

is the reason why we choose Penn-action & Human3.6M. The synthesis with dynamic backgrounds and the modeling of general object moving will be our future work. We set the resolution of both input images and output videos as 128×128, since it is the maximal resolution adopted in existing methods.

4.2 Implementation Details

To train ACT-VAE and P2I networks, we employ Adam optimizer (Kingma & Ba, 2014) with β_1 and β_2 set as 0.5 and 0.999 respectively. The learning rate is 10^{-4} and the batch size is 24. Our approach is implemented in PyTorch 1.0.1 (Paszke et al., 2019), and runs on an Intel 2.60GHz CPU and TITAN X GPU. On average, our framework can create 4 frames in resolution 128×128 with a single input image and an action label within 34.93 ms, where 2.57 ms is spent for ACT-VAE and 32.36ms for P2I network. The model capacity for ACT-VAE and P2I networks are 6.503M and 41.352M respectively.

4.3 Metrics

4.3.1 Key Points Generation Evaluation

To evaluate the accuracy of estimated future key points, we adopt L_2 distance between coordinates of predicted pose sequence and the ground truth, the same as in that of Yang et al. (2018). The coordinate range is $[0, 127] \times [0, 127]$ for 128×128 images. Especially, for each sample, we synthesize 100 sequences to compute their distances with the ground truth and report the average value of 10 smallest distances. Besides, similar to Li et al. (2018), we compute the standard deviation of the predicted key point coordinates from the estimated sequences (these sequences are obtained through the repeated sampling of latent variables for identical input pose) as the indication of how diverse the model predictions are. Specifically, we compute the standard deviation for the coordinate value of each joint, and take the average value of these standard deviations on all joints as the metric. This is also computed with coordinate range as $[0, 127] \times [0, 127]$.

4.3.2 Image Sequence Generation Evaluation

For the evaluation of generated image sequences, we adhere to the protocols in Kim et al. (2019), using FVD (Unterthiner et al., 2018), L_2 distance, action recognition accuracy, and user study. FVD refers to the Fréchet distance between the deep features of real and generated videos. Such features are gained from the I3D model (Carreira & Zisserman, 2017), as used in Kim et al. (2019). FVD is set to compare the visual quality, temporal coherence, and diversity for generated videos. We also employ metrics utilized in Castrejon et al. (2019), which are SSIM (Wang et al., 2004) and LPIPS (Zhang et al., 2018). Especially, LPIPS computes perceptual similarity with deep features.

4.4 Ablation Study

4.4.1 Key Point Generation Evaluation

We first set ablation studies to explore how extra conditions (action labels) contribute to the prediction in ACT-VAE. In ACT-VAE, $z_t \sim \mathcal{Q}_\phi(z_t|z_{<t}, \widehat{p}_{<t}, \mathcal{A})$, and we conduct experiments without action labels for the encoder and decoder of ACT-VAE to analyze the role of $\mathcal{A}$. This setting is called "w/o $\mathcal{A}$" and $z_t \sim \mathcal{Q}_\phi(z_t|z_{<t}, \widehat{p}_{<t})$. The corresponding results are shown in Table 1 (*We use ↓ to denote "the lower the better"; use ↑ to denote "the higher the better"*). Though its performance is inferior to that with action labels, it is still better than existing methods (as shown in Table 3). Thus, the modeling of ACT-VAE does improve pose prediction accuracy even without action labels.

Moreover, it is inevitable to prove the contribution of the condition $z_{<t}$: we build ACT-VAE without action labels, remove the input condition $z_{<t}$, and keep the condition of the past pose for the encoder of ACT-VAE. This setting is called "w/o $\mathcal{AZ}$", directly sampling $z_t \sim \mathcal{Q}_\phi(z_t|\widehat{p}_{<t})$ like (Denton & Fergus, 2018). Compared with "w/o A" in Table 1, "w/o AZ" has worse results for prediction. Thus, the structural novelty of setting $z_{<t}$ as the condition to sample z_t has a great contribution to the accuracy and diversity.

We also verify the significance of temporal coherence by conducting ACT-VAE without action labels and modeling z_t as independent Gaussian distribution. The results ("w/o $\mathcal{AC}$") are worse than "w/o $\mathcal{A}$" and "w/o $\mathcal{AZ}$". Thus, removing temporal coherence reduces performance. The "temporal coherence" principle should be different within various actions, which is called "individual temporal coherence". The superiority of "Ours" over "w/o $\mathcal{A}$" in Table 1 proves its positive impact.

In addition, we provide the visual samples for each ablation setting in Fig. 5, where "w/o $\mathcal{AC}$ I/II", "w/o $\mathcal{AZ}$ I/II", "w/o $\mathcal{A}$ I/II" and "Ours I/II" means two diverse predictions for identical input, derived from the ablation setting of "w/o $\mathcal{AC}$", "w/o $\mathcal{AZ}$", "w/o $\mathcal{A}$" and our full setting, respectively. We should note that our full setting leads to the most outstanding visual accuracy and diversity.

4.4.2 Image Sequence Generation Evaluation

There are two significant parts in our P2I network, i.e., the foreground-background composition strategy and ACBN. We conduct an ablation study to illustrate their respective importance by deleting them from our framework respectively. When remove the foreground-background composi-

Table 1 Quantitative results of ablation study on the key point generation evaluation

	Penn-action				Human3.6M			
	w/o $\mathcal{A}$	w/o $\mathcal{AZ}$	w/o $\mathcal{AC}$	Ours	w/o $\mathcal{A}$	w/o $\mathcal{AZ}$	w/o $\mathcal{AC}$	Ours
L_2 (↓)	29.59	30.31	34.74	**28.32**	30.88	31.01	31.37	**30.41**
Std (↑)	1.584	1.462	0.732	**1.663**	0.730	0.672	0.564	**0.838**

"L_2" is the L_2 distance to measure accuracy, and "Std" is the standard deviation to measure diversity

tion strategy, the P2I network directly synthesizes images without mask prediction. We call this setting "w/o mask". Removing ACBN from the P2I network, and replacing it with normal BN is denoted as "w/o ACBN". The quantitative results are reported in Table 2, and the qualitative samples are shown in Fig. 6. Clearly, removing any of them leads to the degeneration of performance.

4.5 Comparison with Existing Methods

4.5.1 Baselines

For the keypoint generation evaluation, we choose current state-of-the-art strategies which include the pose generation module of KL-VP (Kim et al., 2019), IVRNN (Castrejon et al., 2019), SVG-FP (Denton & Fergus, 2018), SVG-LP (Denton & Fergus, 2018), Dlow (Yuan & Kitani, 2020), MT-VAE (Yan et al., 2018), pose prediction network of LF-VP (Zhao et al., 2018), Traj (Mao et al., 2019) and Rep (Mao et al.,

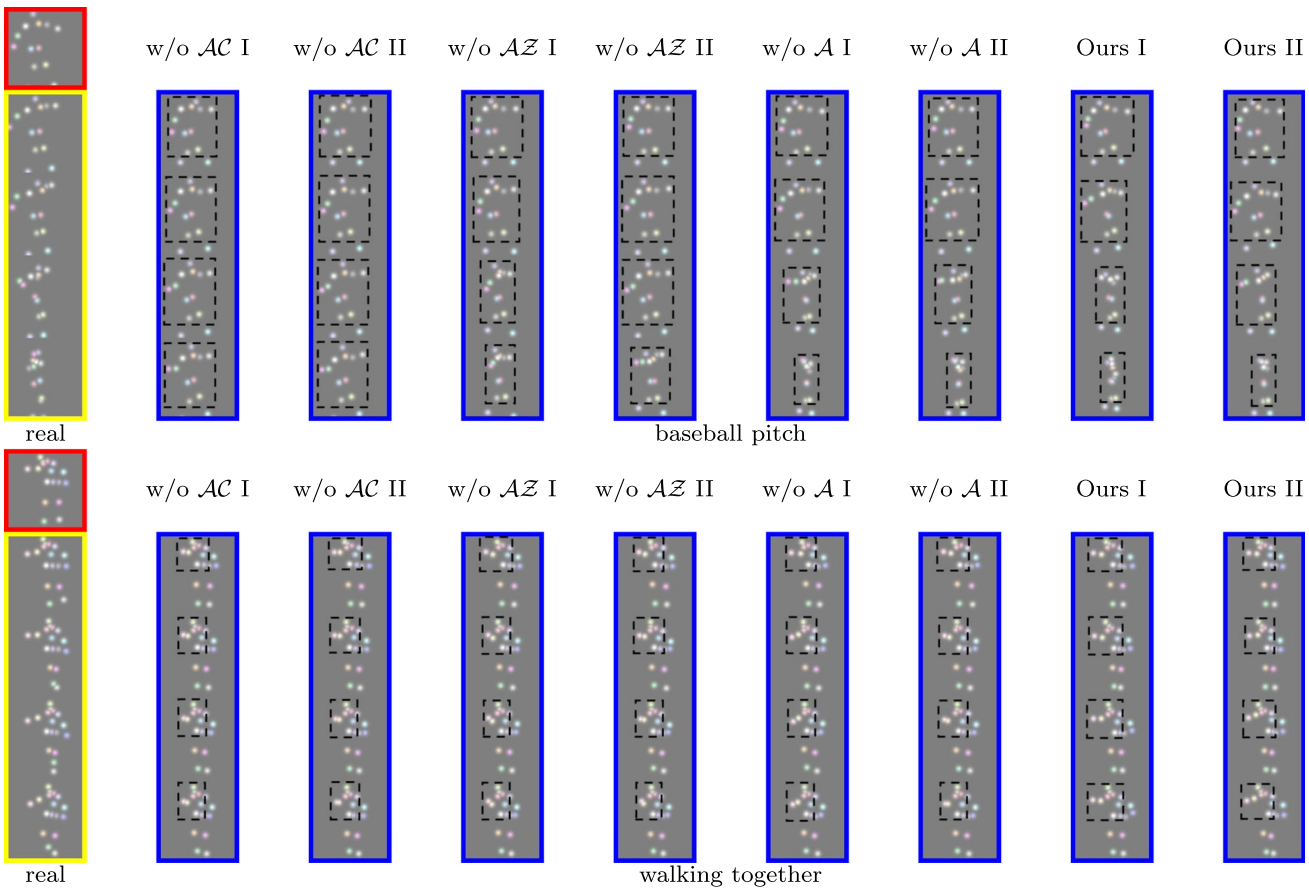


Fig. 5 Visual comparison for ablation study on key point generation evaluation in the testing set. The column of "real" is the input image and real future pose sequence. The red rectangles are input conditional images, the yellow rectangles are real future pose sequences, and the blue rectangles are the predictions. "I" and "II" are two diverse predictions for identical input, derived from each ablation setting. And the black dotted rectangles are the parts that can notably reflect the accuracy and diversity of our results (Color figure online)

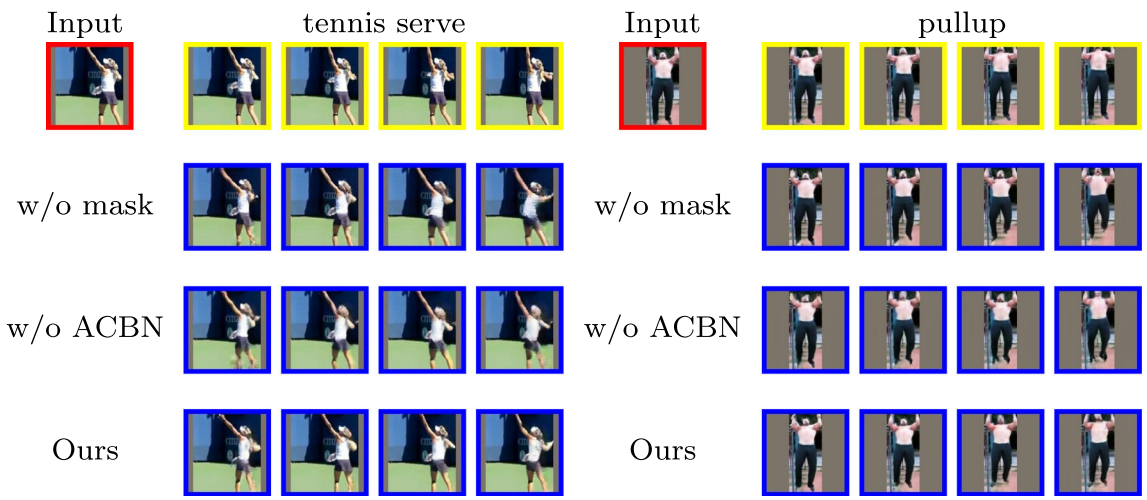


Fig. 6 Visual comparison for ablation study on image sequence generation evaluation in the testing set. Input images, real future pose sequences, and the corresponding predictions are marked by red, yellow, and blue rectangles, respectively (Color figure online)

Table 2 Ablation study on image sequence generation evaluation

Method	FVD (↓)	Accuracy (↑)	L_2 (↓)	SSIM (↑)	LPIPS (↓)
w/o mask	1455.2	66.03	50.04	0.7913	0.1132
w/o ACBN	1377.2	68.83	42.16	0.8028	0.1247
Ours	**1092.8**	**70.04**	**39.82**	**0.8248**	**0.0908**

Table 3 The results of ACT-VAE and chosen strategies on key point generation evaluation

Method	Penn-action		Human3.6M	
	L_2 (↓)	Std (↑)	L_2 (↓)	Std (↑)
VAE of Kim et al. (2019)	32.88	0.895	32.65	0.336
VRNN (Castrejon et al., 2019)	30.24	1.495	30.96	0.552
SVG-FP (Denton & Fergus, 2018)	31.44	1.543	32.88	0.720
SVG-LP (Denton & Fergus, 2018)	30.72	1.519	31.92	0.648
Dlow (Yuan & Kitani, 2020)	55.84	0.617	31.74	**1.082**
Mix-and-Match (Aliakbarian et al., 2020)	32.14	1.541	34.03	0.772
MT-VAE (Yan et al., 2018)	31.85	1.328	33.96	0.514
LSTM of Zhao et al. (2018)	33.43	0.000	38.22	0.000
Traj (Mao et al., 2019)	31.18	0.000	38.55	0.000
Rep (Mao et al., 2020)	29.03	0.000	32.35	0.000
ACT-VAE (w/o $\mathcal{A}$)	29.59	1.584	30.88	0.730
ACT-VAE	**28.32**	**1.663**	**30.41**	0.838

"L_2" is the L_2 distance to measure accuracy, and "Std" is the standard deviation to measure diversity

2020). KL-VP, MT-VAE and Dlow utilize VAE, IVRNN and SVG-LP employ architectures of VRNN, LF-VP and Traj and Rep are all deterministic models for prediction. For fairness, we change the input of all chosen approaches to our pose representation and concatenate it with the action label, and retrain their models with their released codes for comparison. Besides, we align their settings with ours, including training/testing split and training epochs.

Moreover, five representative methods are taken as baselines for image sequence generation evaluation, including HL-VP (Wichers et al., 2018), LG-VP (Villegas et al., 2017), LF-VP (Zhao et al., 2018), KL-VP (Kim et al., 2019) and IVRNN (Castrejon et al., 2019). HL-VP (Wichers et al., 2018) is a typical GAN-based approach. LG-VP (Villegas et al., 2017), LF-VP (Zhao et al., 2018) and KL-VP (Kim et al., 2019) all produce videos by first predicting pose sequences.

Table 4 Comparison between our method and baselines

Method	FVD (↓)	Acc (↑)	L_2 (↓)	SSIM (↑)	LPIPS (↓)	Ranking (↓)
KL-VP (Kim et al., 2019)	1431.2	67.88	44.47	0.7929	0.1127	3.225 ± 0.352
LG-VP (Villegas et al., 2017)	1982.1	52.18	61.39	0.7433	0.1428	6.150 ± 0.522
HL-VP (Wichers et al., 2018)	2814.3	45.21	60.92	0.7231	0.1459	6.300 ± 0.143
LF-VP (Zhao et al., 2018)	1736.9	56.85	58.54	0.7452	0.1368	4.290 ± 0.850
IVRNN (Castrejon et al., 2019)	1970.2	67.35	46.30	0.7548	0.1297	3.590 ± 0.892
KL-VP-Ours	1318.9	62.33	48.61	0.7677	0.1221	2.955 ± 0.241
Ours	**1092.8**	**70.04**	**39.82**	**0.8248**	**0.0908**	**1.490** ± 0.292

Metrics include FVD (Unterthiner et al., 2018), SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), recognition accuracy (Acc), L_2 distance (L_2), and ranking results from user study (written in the form of "mean $\pm$ standard deviation")

IVRNN (Castrejon et al., 2019) achieves the best results among current works adopting VRNN. We use the authors' released codes and unify the training/testing setting for fairness. Further, the comparison is conducted on Penn-action following the setting of Kim et al. (2019). Since our training setting is the same as that of Kim et al. (2019), we use its pre-trained model for comparison. Moreover, we retrain models of Zhao et al. (2018); Villegas et al. (2017); Wichers et al. (2018); Castrejon et al. (2019) with our task setting for evaluation.

4.5.2 Key Point Generation Evaluation

To illustrate the superiority of ACT-VAE in future pose sequence prediction, we compare ACT-VAE with the VAE proposed by Kim et al. (2019), the hierarchical VRNN designed by Castrejon et al. (2019), the straightforward LSTM network adopted in Zhao et al. (2018), SVG-FP and SVG-LP (Denton & Fergus, 2018), MT-VAE (Yan et al., 2018), Dlow (Yuan & Kitani, 2020), Mix-and-Match (Aliakbarian et al., 2020), Traj (Mao et al., 2019) and Rep (Mao et al., 2020). For fairness, we change the input of all chosen baselines to our pose representation and concatenate it with the action label. Significantly, all methods are implemented with their public source codes. As listed in Table 3, ACT-VAE yields the lowest L_2 distance. As for diversity, ACT-VAE achieves higher standard deviations compared with most of the baselines. Although Dlow results in greater diversity on Human3.6M, ACT-VAE has lower L_2 errors. Besides, ACT-VAE has superior results than Dlow on Penn-action in terms of accuracy and diversity. It is mainly the higher complexity of motion patterns in Penn-action over Human3.6M, which causes poor results of Dlow on Penn-action. Thus, ACT-VAE is of higher accuracy and diversity than these approaches on the whole.

4.5.3 Image Sequence Generation Evaluation

Quantitative results. We unify the training and testing settings of all methods, and the experiments are conducted on the Penn-action dataset. The comparison of FVD for different approaches is given in Table 4. "KL-VP-Ours" is produced by synthesizing pose sequences with the VAE structure of Kim et al. (2019) and using our P2I network to obtain image sequences. This table shows that our framework yields the lowest FVD. Next, as in Kim et al. (2019), we also use action recognition accuracy to evaluate the plausibility of synthesized videos. To this end, we train a network for action recognition with the structure of two-stream CNN (Simonyan et al., 2014) on the Penn-Action dataset, which achieves accuracy 82.33% on real testing videos. And it is clearly in Table 4 that the recognition accuracy on our synthesized results is higher than others. It proves that our synthesized motion is in accordance with the ground truth. Further, we compute the L_2 distance in pixel-level between the synthesized image sequence and the ground truth, and a lower L_2 distance suggests better performance. The results in Table 4 show that the L_2 distance between our prediction and the ground truth is the lowest. Moreover, it is also evident in Table 4 that our approach has the highest SSIM while the lowest LPIPS. This outcome further illustrates our superiority.

Finally, a user study is conducted to check the visual quality of generation following the strategy of Kim et al. (2019). For each question, there is a real video for reference and seven synthesized videos by diverse strategies. The order of these seven videos is randomly chosen, and we ask users to rank them based on quality and accuracy of prediction. We invite 20 participants using Google Form, and report the average ranking for each method. Each participant is required to answer 30 questions. Results reported in Table 4 clearly show that the ranking of our results is the highest with low variance. Thus, our results are the best in human perception compared with these baselines.

Qualitative results. Video prediction results of our approach and the baselines on several categories of action in Penn-

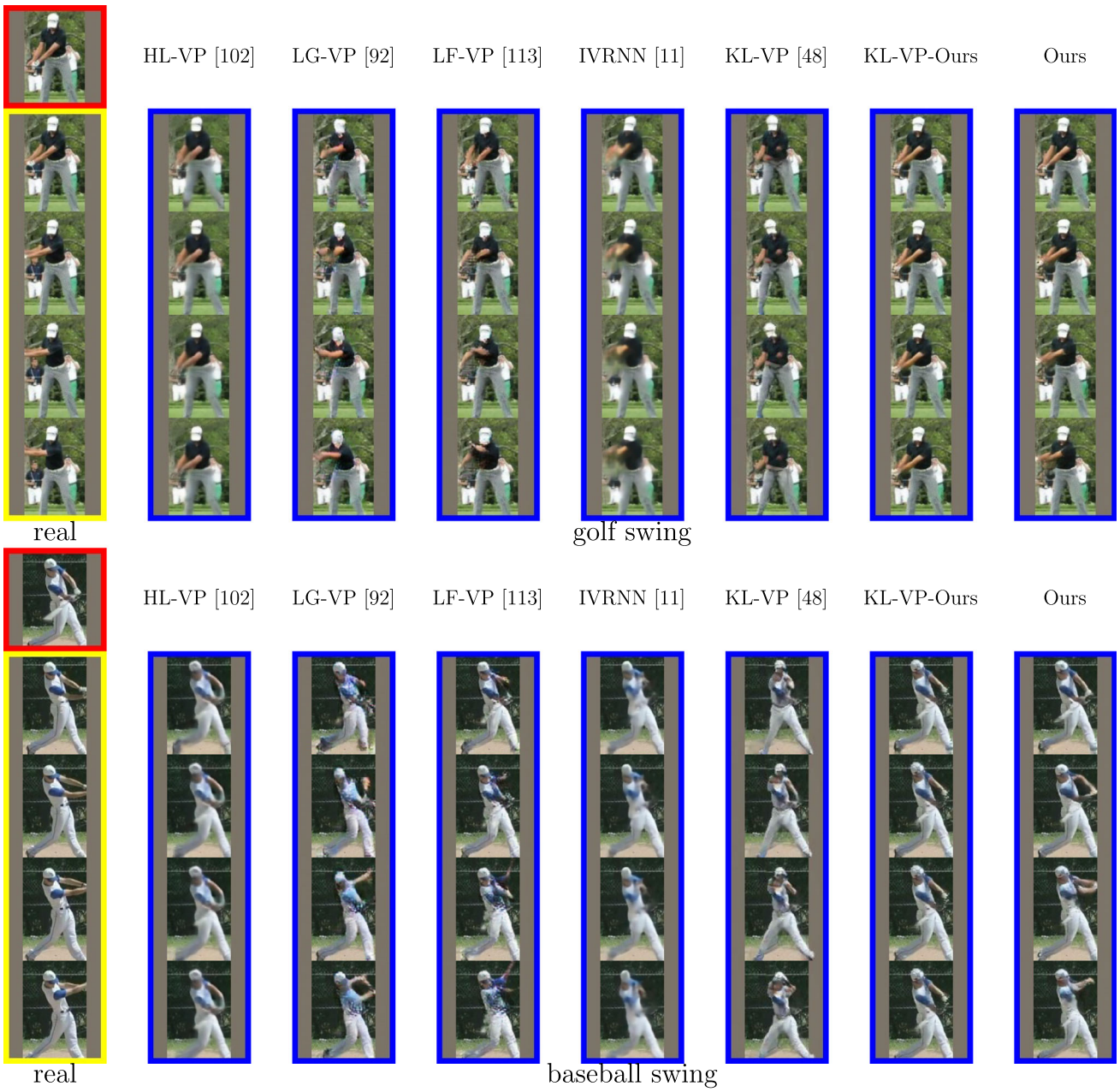


Fig. 7 Visual comparison between our framework and all baselines in the testing set. In the column of "real", the first row is the input image and the following rows are real future frames. The red rectangles are input conditional images, the yellow rectangles are real future frames, and the blue rectangles are the predicted frames from different methods (Color figure online)

Action, are shown in Fig. 7. Our synthesized videos give both the realistic image sequences and the plausible motion for the input action condition. It is also clear that, compared with these baselines, our method achieves improvements in both the visual and the dynamics quality. We note KL-VP and KL-VP-Ours are the strongest baselines, while our results are of higher visual realism. For Zhao et al. (2018); Villegas et al. (2017), the quality of appearance and accuracy of motion prediction have more room for improvement. And our results are also more dynamic and sharper than those of Wichers et al. (2018) and Castrejon et al. (2019).

4.6 The Control of Action Types in Synthesis

Our method can control the action types of the synthesized sequences like (Kim et al., 2019). Given an input image, we can provide a list of action labels $\mathcal{A}_1$, ..., $\mathcal{A}_h$, to synthesize a list of videos whose action types are consistent with the

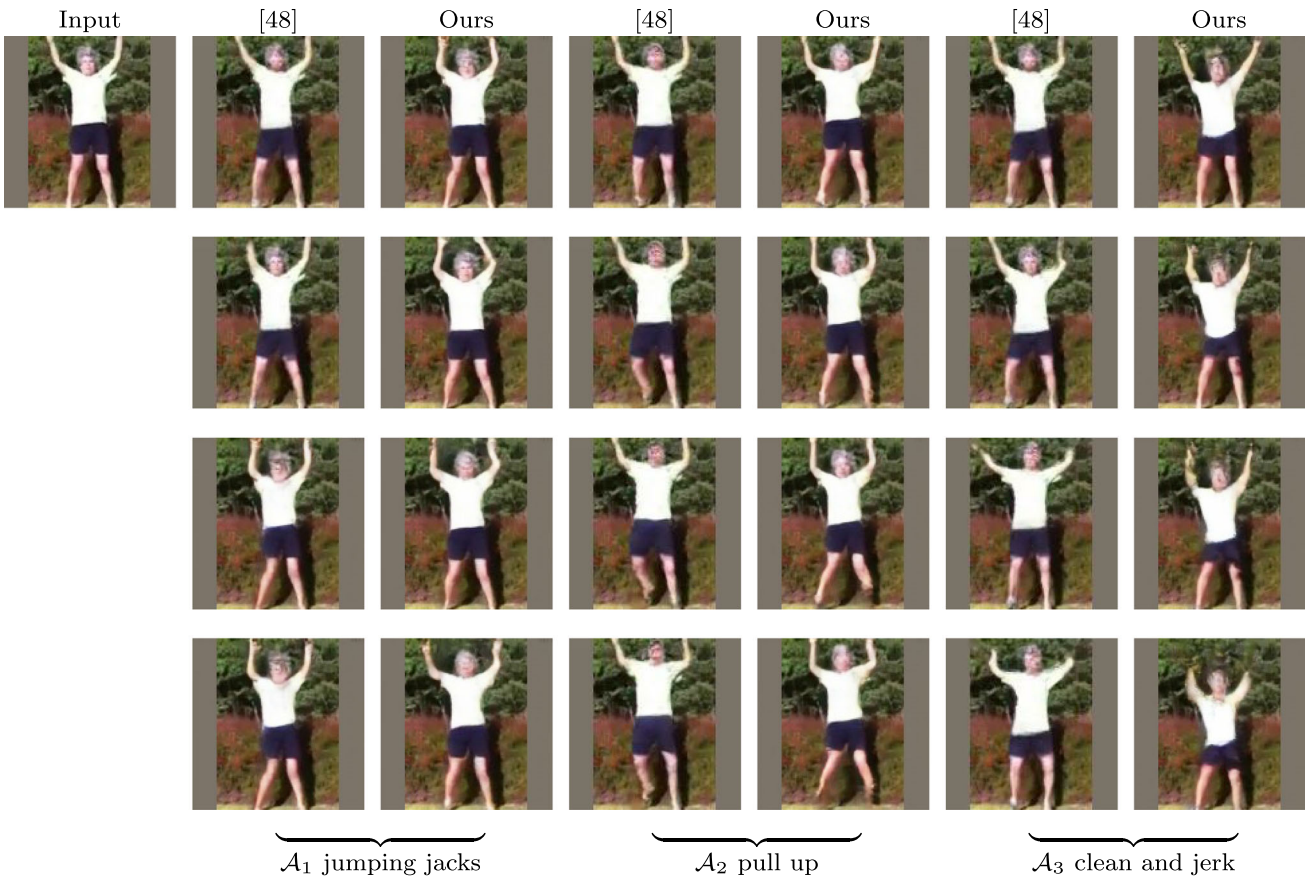


Fig. 8 Our framework can control action types of synthesized image sequences, via controlling the input of action labels

input action labels $\mathcal{A}_1, ..., \mathcal{A}_h$. To verify this, we conduct a quantitative analysis as follows. We synthesize all types of action sequences from each testing image via providing various action labels, and report the value of FVD and recognition accuracy for the synthesized image sequences. The FVD and recognition accuracy of our method are 1257.6 and 58.87; the results of Kim et al. (2019) are 1573.7 and 50.04. These results demonstrate that our approach can achieve the control of action types in the synthesis and outperforms the strong baseline (Kim et al., 2019). Qualitative comparison samples are displayed in Fig. 8.

4.6.1 User Study: AB-test

A user study with AB-test is conducted to evaluate the control of action types in synthesis: we invite 20 participants to sees two videos that are synthesized by our method and Kim et al. (2019), and they will choose which one is more consistent with the input action label. To demonstrate the performance of controlling action types, the action labels are inconsistent with the input conditional images for synthesis. Each participant is required to complete 30 pairs of AB-test and 83.7% of them prefer our method's results.

They are also invited to complete the AB-test for the evaluation of pose sequence: we synthesize pose sequences with our ACT-VAE and all other baselines whose implementations are reported in Sect. 4.5.1. All baselines can synthesize different types of the action sequences since their inputs include the action labels. To demonstrate the performance of controlling action types, the action labels are inconsistent with the input conditional images for synthesis. We invite 20 participants to sees two pose sequences that are synthesized by our ACT-VAE and one of the other baselines, and they will choose which one is more consistent with the input action label (they can also choose that they have no preference). Each participant is required to complete 100 pairs of AB-test (10 baselines and the comparison with each baseline contains 10 questions). The results are shown in Table 5. These results demonstrate that our ACT-VAE can better implement the control of action types in the synthesis.

4.6.2 User Study: Scores for Multiple Dimensions

To evaluate the results of image sequence synthesis with controllable action types comprehensively, we further conduct another user study with 80 participants (age ranges from 15–

Table 5 User preference comparison in the user study

Methods	Other (%)	Same (%)	Ours (%)
VAE of Kim et al. (2019)	15.5	8.5	76.0
VRNN (Castrejon et al., 2019)	24.0	4.5	71.5
SVG-FP (Denton & Fergus, 2018)	20.5	6.0	73.5
SVG-LP (Denton & Fergus, 2018)	18.0	13.5	68.5
Dlow (Yuan & Kitani, 2020)	20.5	12.5	67.0
Mix-and-Match (Aliakbarian et al., 2020)	17.5	10.5	72.0
MT-VAE (Yan et al., 2018)	8.0	17.5	74.5
LSTM of Zhao et al. (2018)	12.5	5.5	82.0
Traj (Mao et al., 2019)	19.5	10.0	70.5
Rep (Mao et al., 2020)	9.0	8.5	82.5

"Ours" is the percentage that our result is preferred, "Other" is the percentage that other method is preferred, "Same" is the percentage that the user has no preference

Table 6 The results of the user study to measure the performance of image sequence synthesis with controllable actions types

Methods	KL-VP (Kim et al., 2019)	LG-VP (Villegas et al., 2017)	HL-VP (Wichers et al., 2018)	LF-VP (Zhao et al., 2018)	IVRNN (Castrejon et al., 2019)	KL-VP-Ours	Ours
Q1: Are the video frames realistic?	3.79	2.54	2.62	3.16	3.35	3.97	**4.29**
Q2: Is the video frames temporally smooth?	3.63	2.89	3.06	3.25	3.14	3.82	**4.06**
Q3: Is the video in accord with the input action label?	4.10	3.52	3.40	4.08	3.96	4.24	**4.81**
Q4: How is video frames' visual quality?	3.57	3.11	2.89	3.29	3.48	3.73	**4.27**
Q5: Are videos with different action labels diverse?	4.05	3.73	3.65	4.02	3.82	4.16	**4.54**
Q6: What is your overall rating for videos?	3.84	3.16	3.02	3.31	3.50	4.03	**4.31**

67 and the ratio of male to female participants is about 2.31 to 1). Inspired by your questions, we evaluate the results from multiple aspects. During the user study, each participant will be shown the synthesis results with different input action labels (9 for Penn-action) from one method (ours and baselines). And the participant should evaluate the videos of one method by rating six questions, including

Q1: Are the video frames realistic?
Q2: Is the video frames temporally smooth?
Q3: Is the video in accord with input action label?
Q4: How is video frames' visual quality?
Q5: Are videos with different action labels diverse?
Q6: What is your overall rating for videos?

These questions measure the realistic degrees of videos, temporal smoothness of videos, the consistency between the generated videos and input action labels, and the diversity among videos with various input action labels. The rating ranges from 1 (worst) to 5 (best), with the real image sequences shown at the beginning of the subjective evaluation as the reference (the scores of the ground truths are all 5). The results are shown in Table 6, indicating that our generated image sequences with action control are better than baselines, and they are realistic as well as temporally smooth. Significantly, the videos generated by our approach are in accord with the given action labels, and they vary for different input action labels.

4.7 Hyper-Parameters Analysis

There are two important hyper-parameters for the training of ACT-VAE, which are λ_1 and λ_2 in Eq. (10). To analyze the influence of λ_1 and λ_2 in training, we conduct experiments with different values for λ_1 and λ_2. The value of the pair (λ_1, λ_2) in above experiments is (200, 0.002).

In this section, we set its value as (20, 0.002), (2000, 0.002), (200, 0.02), (200, 0.0002) respectively. The results on Penn-action dataset can be seen in Table 7, which shows that the value setting (200, 0.002) is rational. And users can adopt our value setting of λ_1 and λ_2, as an appropriate reference for different datasets. The settings of $\lambda_3 \sim \lambda_6$ are based on hyper-parameters of pose-guided image generation methods (i.e., (Zhu et al., 2019)).

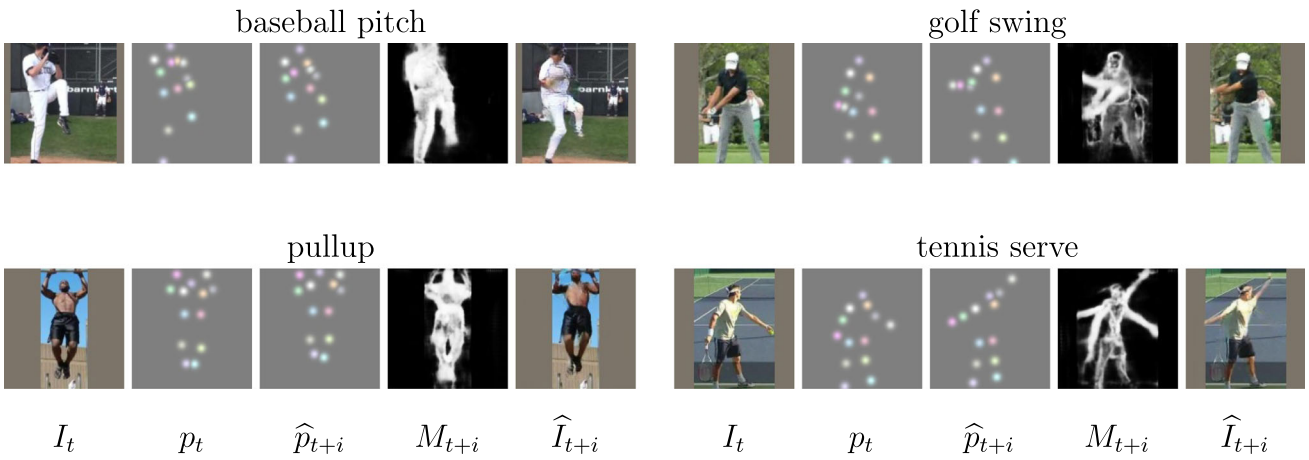


Fig. 9 The visualization of the mask output from the P2I network

Table 7 Results of ACT-VAE with different λ_1 and λ_2

(λ_1)	200	20	2000	200	200
(λ_2)	0.002	0.002	0.002	0.02	0.0002
L_2 (↓)	**28.32**	31.84	29.27	31.99	29.10
Std (↑)	**1.663**	0.720	1.577	0.749	1.582

4.8 Experiments on NTU RGB+D Dataset

We further conduct experiments on the dataset of the NTU RGB+D dataset (Shahroudy et al., 2016) that contains 60 action classes. Especially, we increase the capacity of our approach, by increasing the parameter number of ACT-VAE, enhancing the P2I network with more SOTA structure (Ren et al., 2022) while keeping "Foreground Attention" and "Action Conditional Batch Normalization". The image size for experiments is 128 × 128.

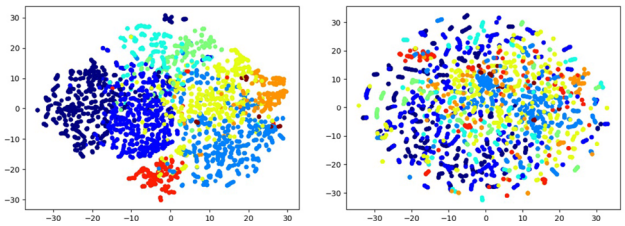
For the evaluation without the requirement of action control, the source image is given with its corresponding action label provided in the dataset. The evaluation metrics are the same as that in Table 4, and the results are shown in Table 8. Especially, we employ the SOTA action recognization approach (Duan et al., 2022) on NTU to train the recognization network for computing the recognition accuracy. The network with multi-modality fusion (RGB + Pose) can achieve an accuracy of 99.1% on the real data. For the methods that can synthesize both RGB and pose sequences (e.g., ours), the RGB and pose outputs are sent into the recognization network to compute the recognition accuracy. And we follow the current video synthesis method (Li et al., 2022) on NTU to evaluate the foreground region, focusing more on the generated action and computing the key-region-based FVD scores, L_2 error, SSIM, and LPIPS. The results in Table 8 show the effectiveness of our method and our superiority over other baselines.

Moreover, for the synthesis with action control (the evaluation setting in Sect. 4.6), The FVD score and recognition accuracy of our method are 36.74 and 82.85; the results of Kim et al. (2019) are 80.67 and 71.73, demonstrating that our method can also achieve the control of action types in the synthesis on the NTU dataset and outperforms the strong baseline (Kim et al., 2019).

Also, since no ground truth is given for the synthesis with action control, subjective results from the large-scale user study are still the reliable metric for us to employ. Thus, we also conduct a user study using the strategy of "Scores for Multiple Dimensions" in Sect. 4.6 The results are shown in Table 9. We can see that our approach's results on NTU RGB+D still obtain higher ratings on all six questions, demonstrating that our method can control the action types in the generation and that the generated realistic and diverse videos are consistent with the given action labels in the NTU RGB+D dataset.

4.9 Visual Analysis for Foreground Attention

Since we adopt a foreground-background composition strategy for the P2I network, the P2I network has two types of outputs which are $\bar{I}_{t+i}$ and M_{t+i} as shown in Eq. (11). We show several cases about the M_{t+i} in Fig. 9, which illustrates that P2I network can distinguish the foreground and background in unsupervised learning, without the ground truth of mask.



(a) P2I with ACBN (b) P2I without ACBN

Fig. 10 t-SNE visualizations in the feature space of the P2I network for different action categories

4.10 Visual Analysis for ACBN in Feature Space

The intuition behind the Action Conditional Batch Normalization (ACBN) can be interpreted from a statistical viewpoint. As indicated in Sec. 3.2.2, we assume that the statistics of intermediate feature maps for each action category are distinctive. Otherwise, the feature maps with different action kinds will be confused with each other, harming the performance of the P2I network. To prove this, we visualize the feature distribution of our framework with and without ACBN. The visualized results are shown in Fig. 10, indicating that the feature maps of different categories are distinct and easy to be individually handled by the generator to achieve high-quality results. On the other hand, for the P2I network without ACBN, the features from different action categories are mixed as shown in Fig. 10.

4.11 Our Performance with 2D/3D Keypoints

In our framework, we need the keypoints for the given single source image. We employ the 2D keypoints of the source image rather than 3D keypoints, since estimating 3D keypoints from a single image is a highly ill-posed problem and 3D keypoints are often estimated from videos (Pavllo et al., 2019; Cheng et al., 2019; Choi et al., 2021) or multi-view data (Kocabas et al., 2019; Rhodin et al., 2018; Iqbal et al., 2020). The approaches predicting 3D keypoints from the single image are few with limited accuracy. Thus, adopting the 3D keypoints into our framework would increase the error accumulation in the stage of ACT-VAE. On the other hand, estimating 2D keypoints from an input image has achieved great success and accuracy (Cao et al., 2019; Geng et al., 2021; Wang et al., 2022). Thus, we adopt the format of 2D keypoints, allowing our framework to be employed in practice.

To compare the effectiveness of using the predicted 2D and 3D keypoints in our framework, we adopt SOTA 2D (Geng et al., 2021) and 3D (Wandt et al., 2021) human pose prediction networks to obtain the pose of the source image, respectively. And then, we evaluate the quality of the final synthesized image sequences. The experimental settings with the predicted 2D and 3D keypoints are called "Ours (Pred. 2D)" and "Ours (Pred. 3D)". The results are shown in Table 8. From this, it is observed that adopting the performance of adopting the estimated 2D keypoints is slightly worse than the effects of using the ground truth of 2D images. And we

Table 8 The comparison for the results when our framework employ either 2D and 3D keypoints

Method	FVD (↓)	Acc (↑)	L_2 (↓)	SSIM (↑)	LPIPS (↓)
Ours (Pred. 3D)	1349.1	67.05	43.68	0.7875	0.1097
Ours (Pred. 2D)	*1167.3*	*68.92*	*41.47*	*0.8013*	*0.1084*
Ours	**1092.8**	**70.04**	**39.82**	**0.8248**	**0.0908**

Bold denotes the best results and *Italics* represents the second best results

Table 9 The comparison of our framework with separate and end-to-end training strategy

Method	FVD (↓)	Acc (↑)	L_2 (↓)	SSIM (↑)	LPIPS (↓)
Ours E2E	1309.2	68.41	42.93	0.7970	0.1104
Ours	**1092.8**	**70.04**	**39.82**	**0.8248**	**0.0908**

Table 10 The improvement of our framework by increasing the capacity of the P2I network

Method	FVD (↓)	Acc (↑)	L_2 (↓)	SSIM (↑)	LPIPS (↓)
Ours with (Zhu et al., 2019)	1092.8	70.04	39.82	0.8248	0.0908
Ours with (Ren et al., 2022) wo mask	1183.5	70.26	40.26	0.8093	0.0974
Ours with (Ren et al., 2022) wo ACBN	1079.2	73.81	35.18	0.8347	0.0915
Ours with (Ren et al., 2022)	**825.4**	**75.52**	**32.71**	**0.8626**	**0.0713**

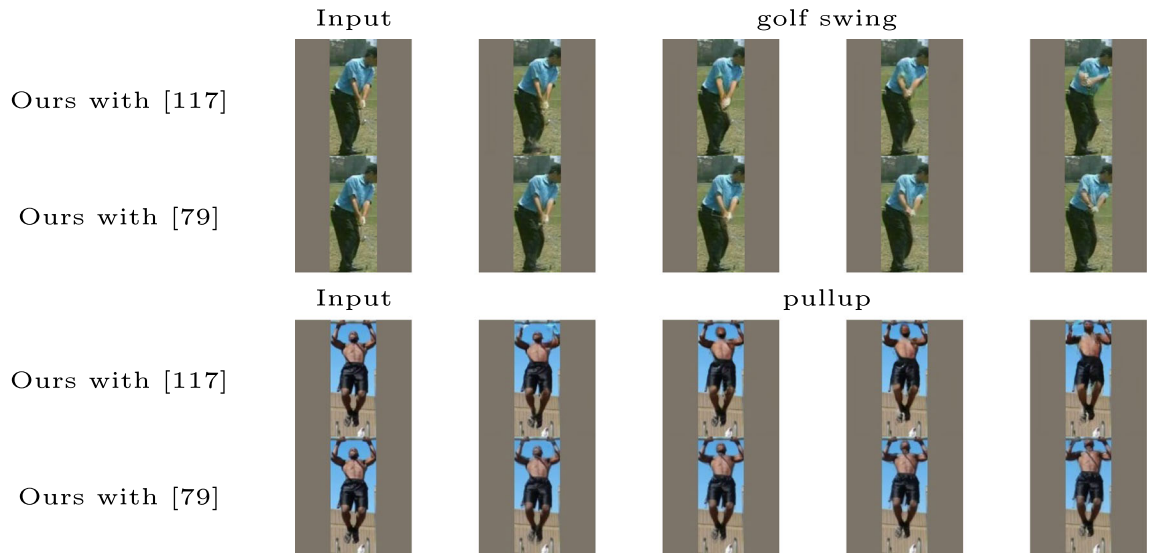


Fig. 11 Visual comparisons for the results before (Previous Result) and after (New Result) enhancing the backbone of the P2I network

Table 11 The results of keypoint generation evaluation following the evaluation protocol of Mao et al. (2020) for short-term prediction of 3D joint positions on H3.6M (Ionescu et al., 2013)

	Walking				Eating				Smoking				Discussion			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Traj (Mao et al., 2019)	11.1	21.4	37.3	42.9	7.0	14.8	29.8	37.3	7.5	15.5	30.7	37.5	10.8	24.0	52.7	65.8
Rep (Mao et al., 2020)	10.0	19.5	34.2	39.8	**6.4**	14.0	28.7	**36.2**	7.0	14.9	29.9	36.4	10.2	**23.4**	**52.1**	65.4
Ours	**9.6**	**17.4**	**33.5**	**37.9**	8.1	**13.5**	**27.2**	37.7	**6.7**	**13.6**	**28.6**	**36.1**	**10.0**	25.2	52.4	**64.7**

	Directions				Greeting				Phoning				Posing				Purchases				Sitting			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Traj (Mao et al., 2019)	8.0	18.8	43.7	**54.9**	14.8	31.4	65.3	79.7	9.3	19.1	39.8	49.7	10.9	25.1	59.1	75.9	13.9	30.3	62.2	75.9	9.8	20.5	44.2	55.9
Rep (Mao et al., 2020)	7.4	18.4	44.5	56.5	13.7	**30.1**	63.8	78.1	8.6	18.3	**39.0**	49.2	10.2	**24.2**	58.5	**75.8**	13.0	29.2	**60.4**	**73.9**	9.3	**20.1**	44.3	56.0
Ours	**7.0**	**17.9**	**43.2**	55.8	**13.3**	30.5	**63.2**	**77.4**	**8.2**	**18.1**	40.5	**49.0**	**9.8**	24.7	**58.2**	76.9	**12.5**	**28.8**	61.3	74.2	**9.1**	21.2	**42.7**	**54.6**

	Sitting Down				Taking Photo				Waiting				Walking Dog				Walking Together				Average			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Traj (Mao et al., 2019)	15.6	31.4	59.1	**71.7**	8.9	18.9	41.0	51.7	9.2	19.5	43.3	54.4	20.9	40.7	73.6	86.6	9.6	19.4	36.5	44.0	11.2	23.4	47.9	58.9
Rep (Mao et al., 2020)	14.9	30.7	59.1	72.0	**8.3**	18.4	40.7	51.5	8.7	**19.2**	43.4	54.9	**20.1**	**40.3**	73.3	86.3	8.9	18.4	35.1	41.9	10.4	22.6	47.1	58.3
Ours	**14.3**	**30.2**	**58.5**	72.4	8.6	**18.1**	**40.2**	**50.4**	**8.3**	20.8	**42.7**	**53.5**	20.7	40.5	**73.0**	**85.9**	**8.4**	**18.1**	**34.7**	**41.5**	**10.3**	**22.5**	**46.7**	**57.9**

can also see that the results of employing 3D keypoints are obviously worse than those of using 2D keypoints. Thus, in practice, 2D keypoints are preferred in our framework.

4.12 Our Performance with End-to-End Training

The main challenge in training our framework end-to-end is the pose error in the first stage (ACT-VAE) will influence the appearance modeling in the second stage (P2I network). Especially, in the phase where the predictions from ACT-VAE are not converged, the error accumulation from the ACT-VAE will cause the training instability of the P2I network that utilizes the adversarial training strategy. The instability leads to unsatisfactory generation effects.

Table 12 The results of keypoint generation evaluation following the evaluation protocol of Mao et al. (2020) for long-term prediction of 3D joint positions on H3.6M (Ionescu et al., 2013)

	Walking				Eating				Smoking				Discussion			
milliseconds	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
Traj (Mao et al., 2019)	53.1	59.9	66.2	70.7	51.1	62.5	72.9	78.6	49.4	59.2	66.9	71.8	88.1	104.5	115.5	121.6
Rep (Mao et al., 2020)	47.4	52.1	55.5	58.1	**50.0**	61.4	70.6	**75.7**	47.6	56.6	64.4	**69.5**	**86.6**	102.2	**113.2**	119.8
Ours	**45.3**	**51.8**	**57.2**	**56.4**	53.6	**60.5**	**70.1**	77.3	**45.1**	**55.9**	**62.7**	70.2	87.5	**100.8**	117.9	**117.1**

	Directions				Greeting				Phoning				Posing				Purchases				Sitting			
milliseconds	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
Traj (Mao et al., 2019)	72.2	**86.7**	98.5	**105.8**	103.7	120.6	134.7	140.9	67.8	83.0	**96.4**	105.1	107.6	136.1	**159.5**	175.0	98.3	115.1	130.1	139.3	**76.4**	**93.1**	106.9	115.7
Rep (Mao et al., 2020)	73.9	88.2	100.1	106.5	101.9	118.4	**132.7**	**138.8**	**67.4**	82.9	96.5	**105.0**	107.6	136.8	161.4	178.2	95.6	**110.9**	125.0	134.2	**76.4**	**93.1**	107.0	115.9
Ours	**71.9**	87.3	**96.5**	108.3	**100.2**	**116.5**	136.8	142.7	68.1	**81.4**	98.2	107.6	**105.4**	**135.8**	163.0	**172.7**	**92.1**	113.8	**122.5**	**131.9**	78.2	95.0	**103.7**	**114.6**

	Sitting Down				Taking Photo				Waiting				Walking Dog				Walking Together				Average			
milliseconds	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
Traj (Mao et al., 2019)	96.2	115.2	**130.8**	142.2	72.5	90.9	105.9	116.3	**73.4**	**88.2**	99.8	107.5	109.7	122.8	139.0	150.1	55.7	61.3	66.4	69.8	78.3	93.3	106.0	114.0
Rep (Mao et al., 2020)	97.0	116.1	132.1	143.6	**72.1**	**90.4**	105.5	115.9	74.5	89.0	100.3	108.2	**108.2**	120.6	135.9	146.9	**52.7**	57.8	62.0	64.9	77.3	91.8	104.1	112.1
Ours	**95.3**	**113.8**	134.2	**141.9**	74.5	92.3	**103.2**	**113.6**	76.0	90.6	**98.4**	**106.2**	111.2	**120.1**	**133.5**	**144.1**	53.8	**56.9**	**60.1**	**63.6**	**77.2**	**91.5**	**103.8**	**111.2**

Table 13 The evaluation on image sequence generation following the evaluation protocol of Kim et al. (2019)

Method	FVD (↓)	Acc (↑)
LG-VP (Villegas et al., 2017)	2187.5	47.14
HL-VP (Wichers et al., 2018)	3324.9	40.00
KL-VP (Kim et al., 2019)	1509.0	68.89
Ours	**1356.9**	**69.57**

To demonstrate the superiority of the separate training strategy, we set another experiment where the ACT-VAE and the P2I network are trained end-to-end. The setting is called "Ours E2E", and the comparisons are shown in Table 9, where the results of end-to-end training are obtained by repeatedly training the framework 5 times and taking the average results, reducing the influence of instability. The results show that the separate training strategy can lead to better performance.

4.13 Foreground Attention and ACBN for Other Types of P2I Backbone

We find our proposed foreground attention and ACBN can be generally applied for other types of P2I backbones, enhancing the corresponding performance. Different from the experiments above, we increase the capacity of the P2I network in this section, evaluate the performance of our framework with a new P2I network (Ren et al., 2022) and equip it with our "Foreground Attention" and "Action Conditional Batch Normalization" that are our contributions. The experiments are conducted with the machine with higher effi-

ciency than TITIAN X, i.e., RTX3090. The results (with name of "Ours with (Ren et al., 2022)") in Table 10 show that the performance of our framework is increased, and the visual results are also improved, as shown in Fig. 11. And the results with (Ren et al., 2022) will be decreased obviously without the foreground attention and ACBN, by comparing "Ours with (Ren et al., 2022)" and "Ours with (Ren et al., 2022) wo mask"/ "Ours with (Ren et al., 2022) wo ACBN" in Table 10. Thus, our proposed "Foreground Attention" and "Action Conditional Batch Normalization" are general network design principles for action video synthesis in various backbones to improve the corresponding performance.

4.14 Evaluation of Our Framework with Other Evaluation Protocols

In this section, we make a further comparison with existing SOTA methods, using their evaluation settings.

4.14.1 Key Point Generation Evaluation

For the keypoint generation evaluation, most baselines can not control the action types of the predicted pose sequences. In their settings, the target is to predict the future pose sequences only with the condition of the given image's content. We evaluate the performance of our framework with such a setting, by removing the action labels from the inputs. The results of keypoint generation evaluation are shown in Tables 11 and 12, where the experiments are conducted on H3.6M (Ionescu et al., 2013) that is one representative dataset. Also, we follow the evaluation protocol of Mao et al. (2020), which is one standard evaluation protocol to evaluate the short-term and long-term pose sequences prediction performance. We can see that our method still produces better results than two SOTA baselines, Traj (Mao et al., 2019) and Rep (Mao et al., 2020), demonstrating the accuracy of our approach in modeling real-world pose sequences.

4.14.2 Image Sequence Generation Evaluation

For the evaluation of image sequence generation, the difference between our setting and Kim et al. (2019) is that keypoints are not provided as the conditions while they are unsupervised learned by learning the keypoints detector with the image translator. We follow the task setting of Kim et al. (2019) by first training a keypoints detector with the P2I network, and then training the ACT-VAE for the motion generation. The experiments are conducted on Penn-Action since the representative baseline (Kim et al., 2019) utilizes this dataset for evaluation. The evaluation of image sequence generation is displayed in Table 13. The results demonstrate that our approach can still result in SOTA performance on image sequence generation under the standard evaluation protocol of Kim et al. (2019).

These results show that our approach can still result in SOTA performance on keypoint and image sequence generation under the standard evaluation protocol from other works.

5 Conclusion

We have proposed an effective framework for human action video prediction from a still image within various action categories. In our framework, ACT-VAE predicts pose by modeling the motion patterns and diversity in future videos with temporal coherence for each action category. The temporal coherence is ensured by sampling the latent variable at each time step based on both historical latent variables and pose during inference. When connected with a plug-and-play P2I network, ACT-VAE can synthesize image sequences and control action types in synthesis. Extensive experiments on datasets containing complicated action videos illustrate the superiority of our framework.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11263-023-01832-8.

Acknowledgements This work is supported by Key Research Project of Zhejiang Lab (No. K2022PG1BB01). This work is also supported by Research Project of Zhejiang Lab (No.2022PD0AC02).

Data Availability The data that support the results and analysis of this study is publicly available in a repository. The dataset of Penn-action is available at http://dreamdragon.github.io/PennAction. The dataset of Human3.6M is available at http://vision.imar.ro/human3.6m/description.php. The dataset of NTU RGB+D Dataset is available at https://rose1.ntu.edu.sg/dataset/actionRecognition.

References

Aberman, K., Wu, R., Lischinski, D., Chen, B., & Cohen-Or, D. (2019). *Learning character-agnostic motion for motion retargeting in 2d*. arXiv:1905.01680.

Adeli, V., Ehsanpour, M., Reid, I., Niebles, J. C., Savarese, S., Adeli, E., & Rezatofighi, H. (2021). Tripod: Human trajectory and pose dynamics forecasting in the wild. In *International conference on computer vision*.

Ahuja, C., & Morency, L. P. (2019). Language2pose: Natural language grounded pose forecasting. In *2019 International conference on 3D vision (3DV)*.

Aliakbarian, S., Saleh, F. S., Salzmann, M., Petersson, L., & Gould, S. (2020). A stochastic conditioning scheme for diverse human motion prediction. In *IEEE conference on computer vision and pattern recognition*.

Babaeizadeh, M., Finn, C., Erhan, D., Campbell, R. H., & Levine, S. (2017). *Stochastic variational video prediction*. arXiv:1710.11252.

Balaji, Y., Min, M. R., Bai, B., Chellappa, R., & Graf, H. P. (2019). Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*.

Cai, H., Bai, C., Tai, Y. W., & Tang, C. K. (2018). Deep video generation, prediction and completion of human action sequences. In *The European Conference on Computer Vision*.

Cai, Y., Huang, L., Wang, Y., Cham, T. J., Cai, J., Yuan, J., Liu, J., Yang, X., Zhu, Y., Shen, X., et al. (2020). Learning progressive joint propagation for human motion prediction. In *The european conference on computer vision*.

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., & Sheikh, Y. A. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Intell.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In *IEEE conference on computer vision and pattern recognition*.

Castrejon, L., Ballas, N., & Courville, A. (2019). Improved conditional vrnns for video prediction. In *International Conference on Computer Vision*.

Chen, G., Li, J., Lu, J., & Zhou, J. (2021). Human trajectory prediction via counterfactual analysis. In *International Conference on Computer Vision*.

Chen, W., & Hays, J. (2018). Sketchygan: Towards diverse and realistic sketch to image synthesis. In *IEEE conference on computer vision and pattern recognition*.

Cheng, Y., Yang, B., Wang, B., Yan, W., & Tan, R. T. (2019). Occlusion-aware networks for 3d human pose estimation in video. In *International conference on computer vision*.

Choi, H., Moon, G., Chang, J. Y., & Lee, K. M. (2021). Beyond static features for temporally consistent 3d human pose and shape from a video. In *IEEE conference on computer vision and pattern recognition*.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A.C., & Bengio, Y. (2015). A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*.

Clark, A., Donahue, J., & Simonyan, K. (2019). Adversarial video generation on complex datasets. arXiv:1907.06571.

Cui, A., McKee, D., & Lazebnik, S. (2021). Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *International Conference on Computer Vision*

Denton, E., & Fergus, R. (2018). Stochastic video generation with a learned prior. arXiv:1802.07687.

Duan, H., Zhao, Y., Chen, K., Lin, D., & Dai, B. (2022). Revisiting skeleton-based action recognition. In *IEEE conference on computer vision and pattern recognition*.

Duan, J., Wang, L., Long, C., Zhou, S., Zheng, F., Shi, L., & Hua, G. (2022). Complementary attention gated network for pedestrian trajectory prediction. In *AAAI*.

Finn, C., Goodfellow, I., & Levine, S. (2016). Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems*.

Frühstück, A., Singh, K. K., Shechtman, E., Mitra, N. J., Wonka, P., & Lu, J. (2022). Insetgan for full-body image generation. In *IEEE conference on computer vision and pattern recognition*.

Fu, J., Li, S., Jiang, Y., Lin, K. Y., Qian, C., Loy, C. C., Wu, W., & Liu, Z. (2022). Stylegan-human: A data-centric odyssey of human generation. In *The European Conference on Computer Vision*.

Gafni, O., Ashual, O., & Wolf, L. (2021). Single-shot freestyle dance reenactment. In *IEEE conference on computer vision and pattern recognition*.

Ge, C., Song, Y., Ge, Y., Yang, H., Liu, W., & Luo, P. (2021). Disentangled cycle consistency for highly-realistic virtual try-on. In *IEEE conference on computer vision and pattern recognition*.

Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., & Luo, P. (2021). Parser-free virtual try-on via distilling appearance flows. In IEEE conference on computer vision and pattern recognition.

Geng, Z., Sun, K., Xiao, B., Zhang, Z., & Wang, J. (2021). Bottom-up human pose estimation via disentangled keypoint regression. In *IEEE conference on computer vision and pattern recognition*.

Ghosh, A., Zhang, R., Dokania, P.K., Wang, O., Efros, A.A., Torr, P.H., & Shechtman, E. (2019). Interactive sketch & fill: Multiclass sketch-to-image translation. In *International conference on computer vision*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*.

Gopalakrishnan, A., Mali, A., Kifer, D., Giles, L., & Ororbia, A.G. (2019). A neural temporal model for human motion prediction. In *IEEE conference on computer vision and pattern recognition*.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*.

Guen, V. L., & Thome, N. (2020). Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *IEEE conference on computer vision and pattern recognition*.

Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., & Cheng, L. (2022). Generating diverse and natural 3d human motions from text. In *IEEE conference on computer vision and pattern recognition*.

Guo, X., & Choi, J. (2019). Human motion prediction via learning local structure representations and temporal dependencies. In *AAAI*.

Guo, X., Zhao, Y., & Li, J. (2021). *Danceit: Music-inspired dancing video synthesis*. IEEE Transactions on Image Process.

Han, L., Ren, J., Lee, H.Y., Barbieri, F., Olszewski, K., Minaee, S., Metaxas, D., & Tulyakov, S. (2022). Show me what and tell me how: Video synthesis via multimodal conditioning. In *IEEE conference on computer vision and pattern recognition*.

Ho, T.T., Virtusio, J.J., Chen, Y.Y., Hsu, C.M., & Hua, K.L. (2020). Sketch-guided deep portrait generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*.

Huang, Y., Bi, H., Li, Z., Mao, T., & Wang, Z. (2019). Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *International Conference on Computer Vision*.

Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2013). Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Iqbal, U., Molchanov, P., & Kautz, J. (2020). Weakly-supervised 3d human pose learning via multi-view images in the wild. In *IEEE conference on computer vision and pattern recognition*.

Jia, X., De Brabandere, B., Tuytelaars, T., & Gool, L. V. (2016). Dynamic filter networks. In *Advances in Neural Information Processing Systems*.

Jiang, Y., Yang, S., Qju, H., Wu, W., Loy, C. C., & Liu, Z. (2022). *Text2human: Text-driven controllable human image generation. ACM Transactions on Graph*.

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *The European Conference on Computer Vision*.

Kalchbrenner, N., van den Oord, A., Simonyan, K., Danihelka, I., Vinyals, O., Graves, A., & Kavukcuoglu, K. (2017). Video pixel networks. In *ICML*.

Kappel, M., Golyanik, V., Elgharib, M., Henningson, J. O., Seidel, H. P., Castillo, S., Theobalt, C., & Magnor, M. (2021). High-fidelity neural human motion transfer from monocular video. In *IEEE conference on computer vision and pattern recognition*.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan.

In *IEEE IEEE conference on computer vision and pattern recognition*.

Kim, Y., Nam, S., Cho, I., & Kim, S.J. (2019). Unsupervised keypoint learning for guiding class-conditional video prediction. In *Advances in Neural Information Processing Systems*.

Kim, Y., Nam, S., Cho, I., & Kim, S. J. (2019). Unsupervised keypoint learning for guiding class-conditional video prediction. In *Advances in Neural Information Processing Systems*.

Kingma, D.P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv:1412.6980.

Kingma, D.P., & Welling, M. (2014). Auto-encoding variational bayes. In *The International Conference on Learning Representations*.

Kocabas, M., Karagoz, S., & Akbas, E. (2019). Self-supervised learning of 3d human pose using multi-view geometry. In *IEEE conference on computer vision and pattern recognition*.

Kothari, P., Sifringer, B., & Alahi, A. (2021). Interpretable social anchors for human trajectory forecasting in crowds. In *IEEE conference on computer vision and pattern recognition*.

Kumar, M., Babaeizadeh, M., Erhan, D., Finn, C., Levine, S., Dinh, L., & Kingma, D. (2019). Videoflow: A flow-based generative model for video. arXiv:1903.01434

Kwon, Y.H., & Park, M.G. (2019). Predicting future frames using retrospective cycle gan. In *IEEE conference on computer vision and pattern recognition*.

Lee, A.X., Zhang, R., Ebert, F., Abbeel, P., Finn, C., & Levine, S. (2018). Stochastic adversarial video prediction. arXiv:1804.01523.

Lee, H. Y., Yang, X., Liu, M. Y., Wang, T. C., Lu, Y. D., Yang, M. H., & Kautz, J. (2019). Dancing to music. In *Advances in Neural Information Processing Systems*.

Li, C., Zhang, Z., Sun Lee, W., & Hee Lee, G. (2018). Convolutional sequence to sequence model for human dynamics. In *IEEE conference on computer vision and pattern recognition*.

Li, L., Wang, S., Zhang, Z., Ding, Y., Zheng, Y., Yu, X., & Fan, C. (2021). Write-a-speaker: Text-based emotional and rhythmic talking-head generation. In *AAAI*.

Li, X., Zhang, J., Li, K., Vyas, S., & Rawat, Y.S. (2022). Pose-guided generative adversarial net for novel view action synthesis. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*.

Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., & Yang, M.H. (2018). Flow-grounded spatial-temporal video prediction from still images. In *The European Conference on Computer Vision*.

Li, Y., Li, Y., Lu, J., Shechtman, E., Lee, Y. J., & Singh, K. K. (2021). Collaging class-specific gans for semantic image synthesis. In *International Conference on Computer Vision*.

Liu, D., Wu, L., Zheng, F., Liu, L., & Wang, M. (2022). Verbal-person nets: Pose-guided multi-granularity language-to-person generation. *IEEE Transactions on Neural Networks and Learning Systems*.

Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., & Van Gool, L. (2017). Pose guided person image generation. In *Advances in Neural Information Processing Systems*.

Mao, W., Liu, M., & Salzmann, M. (2020). History repeats itself: Human motion prediction via motion attention. In *The European Conference on Computer Vision*.

Mao, W., Liu, M., Salzmann, M., & Li, H. (2019). Learning trajectory dependencies for human motion prediction. In *The European Conference on Computer Vision*.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Paul Smolley, S. (2017). Least squares generative adversarial networks. In IEEE conference on computer vision and pattern recognition.

Mathieu, M., Couprie, C., & LeCun, Y. (2015). Deep multi-scale video prediction beyond mean square error. arXiv:1511.05440.

Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *11th Annual Conference of the International Speech Communication Association*.

Minderer, M., Sun, C., Villegas, R., Cole, F., Murphy, K. P., & Lee, H. (2019). Unsupervised learning of object structure and dynamics from videos. In *Advances in Neural Information Processing Systems*.

Neverova, N., Alp Guler, R., & Kokkinos, I. (2018). Dense pose transfer. In *The European Conference on Computer Vision*.

Oliu, M., Selva, J., & Escalera, S. (2018). Folded recurrent neural networks for future video prediction. In *The European Conference on Computer Vision*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*.

Pavllo, D., Feichtenhofer, C., Grangier, D., & Auli, M. (2019). 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *IEEE conference on computer vision and pattern recognition*.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., & Courville, A. (2018). Film: Visual reasoning with a general conditioning layer. In *AAAI*.

Piergiovanni, A., Angelova, A., Toshev, A., & Ryoo, M.S. (2020). *Adversarial generative grammars for human activity prediction*. arXiv:2008.04888.

Razavi, A., Oord, A. V. D., Poole, B., & Vinyals, O. (2019). Preventing posterior collapse with delta-vaes. In *ICML*

Ren, X., Li, H., Huang, Z., & Chen, Q. (2020). Self-supervised dance video synthesis conditioned on music. In *ACM International Conference on Multimedia*.

Ren, Y., Fan, X., Li, G., Liu, S., & Li, T.H. (2022). Neural texture extraction and distribution for controllable person image synthesis. In *IEEE conference on computer vision and pattern recognition*.

Rhodin, H., Spörri, J., Katircioglu, I., Constantin, V., Meyer, F., Müller, E., Salzmann, M., & Fua, P. (2018). Learning monocular 3d human pose estimation from multi-view images. In *IEEE conference on computer vision and pattern recognition*.

Shahroudy, A., Liu, J., Ng, T. T., & Wang, G. (2016). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *IEEE conference on computer vision and pattern recognition*.

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019). First order motion model for image animation. In *Advances in Neural Information Processing Systems*.

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019). Animating arbitrary objects via deep motion transfer. In *IEEE conference on computer vision and pattern recognition*.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*.

Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C. C., & Liu, Z. (2022). Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *IEEE conference on computer vision and pattern recognition*.

Srivastava, N., Mansimov, E., & Salakhudinov, R. (2015). Unsupervised learning of video representations using lstms. In *ICML*.

Tang, H., Bai, S., Zhang, L., Torr, P.H., & Sebe, N. (2020). Xinggan for person image generation. In *The European Conference on Computer Vision*.

Tulyakov, S., Liu, M.Y., Yang, X., & Kautz, J. (2018). Mocogan: Decomposing motion and content for video generation. In *IEEE conference on computer vision and pattern recognition*.

Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., & Gelly, S. (2018). Towards accurate generative models of video: A new metric & challenges. arXiv:1812.01717.

Villegas, R., Yang, J., Ceylan, D., & Lee, H. (2018). Neural kinematic networks for unsupervised motion retargetting. In *IEEE conference on computer vision and pattern recognition*.

Villegas, R., Yang, J., Hong, S., Lin, X., & Lee, H. (2017). Decomposing motion and content for natural video sequence prediction. arXiv:1706.08033.

Villegas, R., Yang, J., Zou, Y., Sohn, S., Lin, X., & Lee, H. (2017). Learning to generate long-term future via hierarchical prediction. In *ICML*.

Walker, J., Marino, K., Gupta, A., & Hebert, M. (2017). The pose knows: Video forecasting by generating pose futures. In *International Conference on Computer Vision*.

Wandt, B., Rudolph, M., Zell, P., Rhodin, H., & Rosenhahn, B. (2021). Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *IEEE conference on computer vision and pattern recognition*.

Wang, B., Adeli, E., Chiu, H. K., Huang, D. A., & Niebles, J. C. (2019). Imitation learning for human pose prediction. In *International Conference on Computer Vision*.

Wang, T. C., Liu, M. Y., Zhu, J. Y., Liu, G., Tao, A., Kautz, J., & Catanzaro, B. (2018). Video-to-video synthesis. arXiv:1808.06601.

Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE conference on computer vision and pattern recognition*.

Wang, W., Alameda-Pineda, X., Xu, D., Fua, P., Ricci, E., & Sebe, N. (2018). Every smile is unique: Landmark-guided diverse smile generation. In *IEEE conference on computer vision and pattern recognition*.

Wang, Y., Li, M., Cai, H., Chen, W.M., & Han, S. (2022). Lite pose: Efficient architecture design for 2d human pose estimation. In *IEEE conference on computer vision and pattern recognition*.

Wang, Y., Zhang, J., Zhu, H., Long, M., Wang, J., & Yu, P. S. (2019). Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In *IEEE conference on computer vision and pattern recognition*.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). *Image quality assessment: From error visibility to structural similarity*. *IEEE Transactions on Image Process*.

Wichers, N., Villegas, R., Erhan, D., & Lee, H. (2018). *Hierarchical long-term video prediction without supervision*. arXiv:1806.04768.

Wu, Q., Chen, X., Huang, Z., & Wang, W. (2020). Generating future frames with mask-guided prediction. In *The IEEE International Conference on Multimedia and Expo*.

Xu, J., Ni, B., Li, Z., Cheng, S., & Yang, X. (2018). Structure preserving video prediction. In *IEEE conference on computer vision and pattern recognition*.

Yan, X., Rastogi, A., Villegas, R., Sunkavalli, K., Shechtman, E., Hadap, S., Yumer, E., & Lee, H. (2018). Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *The European Conference on Computer Vision*.

Yang, C., Wang, Z., Zhu, X., Huang, C., Shi, J., & Lin, D. (2018). Pose guided human video generation. In *The European Conference on Computer Vision*.

Yang, Z., Zhu, W., Wu, W., Qian, C., Zhou, Q., Zhou, B., & Loy, C.C. (2020). Transmomo: Invariance-driven unsupervised video motion retargeting. In IEEE conference on computer vision and pattern recognition.

Yoo, Y., Yun, S., Jin Chang, H., Demiris, Y., & Young Choi, J. (2017). Variational autoencoded regression: high dimensional regression of visual data on complex manifold. In *IEEE conference on computer vision and pattern recognition*.

Yoon, J.S., Liu, L., Golyanik, V., Sarkar, K., Park, H.S., & Theobalt, C. (2021). Pose-guided human animation from a single image in the wild. In *IEEE conference on computer vision and pattern recognition*.

Yuan, Y., & Kitani, K. (2020). Dlow: Diversifying latent flows for diverse human motion prediction. In The European Conference on Computer Vision.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In IEEE conference on computer vision and pattern recognition.

Zhang, W., Zhu, M., & Derpanis, K.G. (2013). From actemes to action: A strongly-supervised representation for detailed action understanding. In *International Conference on Computer Vision*.

Zhao, L., Peng, X., Tian, Y., Kapadia, M., & Metaxas, D. (2018). Learning to forecast and refine residual motion for image-to-video generation. In *The European Conference on Computer Vision*.

Zhou, X., Huang, S., Li, B., Li, Y., Li, J., & Zhang, Z. (2019). Text guided person image synthesis. In *IEEE conference on computer vision and pattern recognition*.

Zhu, J.Y., Park, T., Isola, P., & Efros, A.A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision*.

Zhu, W., Yang, Z., Di, Z., Wu, W., Wang, Y., & Loy, C.C. (2022). Mocanet: Motion retargeting in-the-wild via canonicalization networks. In *AAAI*.

Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., & Bai, X. (2019). Progressive pose attention transfer for person image generation. In *IEEE conference on computer vision and pattern recognition*.

Zhuang, W., Wang, C., Chai, J., Wang, Y., Shao, M., & Xia, S. (2022). Music2dance: Dancenet for music-driven dance generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*