Pose Guided Human Video Generation

Ceyuan Yang[1], Zhe Wang[2], Xinge Zhu[1], Chen Huang[3],
Jianping Shi[2], and Dahua Lin[1]

[1] CUHK-SenseTime Joint Lab, CUHK, Hong Kong S.A.R.
[2] SenseTime Research, Beijing, China
[3] Carnegie Mellon University
yangceyuan@sensetime.com

Abstract. Due to the emergence of Generative Adversarial Networks, video synthesis has witnessed exceptional breakthroughs. However, existing methods lack a proper representation to explicitly control the dynamics in videos. Human pose, on the other hand, can represent motion patterns intrinsically and interpretably, and impose the geometric constraints regardless of appearance. In this paper, we propose a pose guided method to synthesize human videos in a disentangled way: *plausible motion prediction* and *coherent appearance generation*. In the first stage, a Pose Sequence Generative Adversarial Network (PSGAN) learns in an adversarial manner to yield pose sequences conditioned on the class label. In the second stage, a Semantic Consistent Generative Adversarial Network (SCGAN) generates video frames from the poses while preserving coherent appearances in the input image. By enforcing semantic consistency between the generated and ground-truth poses at a high feature level, our SCGAN is robust to noisy or abnormal poses. Extensive experiments on both human action and human face datasets manifest the superiority of the proposed method over other state-of-the-arts.

Keywords: Human video generation, Pose synthesis, Generation adversarial network

1 Introduction

With the emergence of deep convolution networks, a large amount of generative models have been proposed to synthesize images, such as Variational Auto-Encoders [1] and Generative Adversarial Networks [2]. Meanwhile, video generation and video prediction tasks [3–6] have found big progress as well. Among them, the task of human video generation attracts increasing attention lately. One reason is that human video synthesis allows for many human-centric applications like avatar animation. On the other hand, generation of human videos/frames can act as a data augmentation method which largely relieves the burden of manual annotations. This will speed up the development of a wide range of video understanding tasks such as action recognition.

Human video generation is a non-trivial problem itself. Unlike static image synthesis, the human video generation task not only needs to take care of the

temporal smoothness constraint but also the uncertainty of human motions. Therefore, a proper representation of human pose and its dynamics plays an important role in the considered problem. Recent works attempted to model video dynamics separately from appearances. For instance, in [7] each frame is factorized into a stationary part and a temporally varying component. Vondrick *et al.* [8] untangled the foreground scene dynamics from background with a two-stream generative model. Saito *et al.* [9] generated a set of latent variables (each corresponds to an image frame) and learned to transform them into a video. Tulyakov *et al.* [10] generated a sequence of video frames from a sequence of random vectors, each consists of a content part and a motion part. All these methods show the promise of separate modeling of motion dynamics and appearances. However, motions cannot be controlled explicitly in these methods. The motion code is usually sampled from a random latent space, with no physical meaning about the targeted motion pattern.

Here we argue that for human video generation, to model human dynamics effectively and control motions explicitly, the motion representation should be interpretable and accessible. Inspired by the action recognition literature [11–14], human body skeletons are favorable in that they characterize the geometric body configuration regardless of the appearance difference, and their dynamics capture motion patterns interpretably. It is also worth noting that human skeletons can be easily obtained by many state-of-the-art human pose estimators (*e.g.* [15]).

Therefore, we propose a pose guided method to synthesize human videos. The method consists of two stages: *plausible motion prediction* and *coherent appearance generation*, generating the pose dynamics and corresponding human appearances separately. In the first stage, human pose is used to model various motion patterns. The Pose Sequence Generative Adversarial Network (PSGAN) is proposed to learn such patterns explicitly, conditioned on different action labels. In the second stage, a Semantic Consistent Generative Adversarial Network (SCGAN) is proposed to generate video frames given the generated pose sequence in the first stage and input image. Meanwhile, the semantic consistency between the generated and ground-truth poses is also enforced at a high feature level, in order to alleviate the influence of some noisy or abnormal poses. Experiments will show the efficacy and robustness of our method when generating a wide variety of human action and facial expression videos. Fig. 1 illustrates the overall framework. We summarize the major contributions of as follows:

- We propose a Pose Sequence Generative Adversarial Network (PSGAN) for plausible motion prediction based on human pose, which allows us to model the dynamics of human motion explicitly.
- Semantic Consistent Generative Adversarial Network (SCGAN) is designed to synthesize coherent video frames given the generated pose and input image, with an effective mechanism for handling abnormal poses.
- Qualitative and quantitative results on human action and facial expression datasets show the superiority of the proposed method over prior arts. The controlled experiments also show our flexibility to manipulate human motions as well as appearances. Codes will be made publicly available.
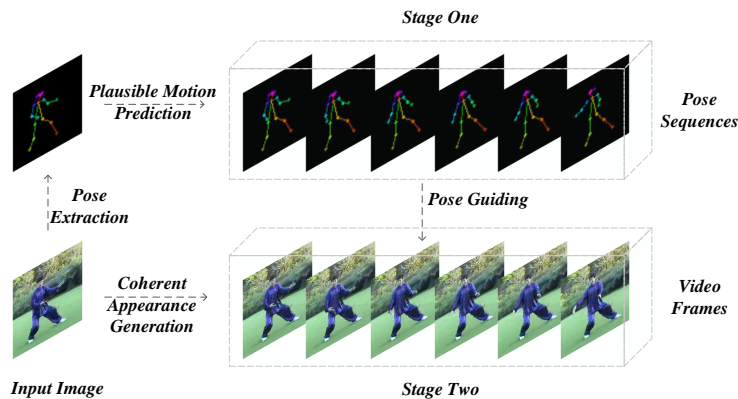
Fig. 1. The framework of our method. In the first stage, we extract the corresponding pose for an input image and feed the pose into our PSGAN to generate a pose sequence. In the second stage, SCGAN synthesizes photo-realistic video frames given the generated poses and input image

2   Related work

Deep generative models have been extensively studied to synthesize natural images, typically using Variational Auto-Encoders (VAEs) [1] and Generative Adversarial Networks (GANs) [2]. Many follow-up works aim to improve the training of GANs [2] and thus enhance the generated image quality. The authors of [16] noted that the uncertainty of data distribution is likely to cause model collapse, and proposed to use convolutional networks to stabilize training. The works in [17,18] handle the problem of GAN training instability as well.

Another direction is to explore image generation in a conditioned manner. The pioneer work in [19] proposes a conditional GAN to generate images controlled by class labels or attributes. More recently, Ma *et al.* [20] proposed a pose guided person generation network to synthesize person images in arbitrary new poses. StackGAN [21] is able to generate photo-realistic images from some text descriptions. The works in [22–24] further generalize to learn to translate data from one domain to another in an unsupervised fashion. StarGAN [25] even allows to perform image-to-image translation among multiple domains with only a single model. Our proposed PSGAN and SCGAN are also designed conditional, generating a human pose sequence given an action label and then generating video frames given the pose sequence and input image. Our two conditional models are able to generate a continuous human video at one time rather than static images only. The SCGAN can also alleviate the impact of abnormal poses by learning semantic pose representations.

The task of video generation is intrinsically a much more challenging one than the image synthesis task, due to the requirement of foreground dynamics modeling and temporal smoothness constraint. During the past few years, it is

with the access to powerful GPUs and the advent of deep convolution networks that video generation and video prediction [3–6] have gained large momentum as well. As an example, one GAN model with a spatio-temporal convolutional architecture is proposed in [8] to model the foreground scene dynamics in videos. Tulyakov *et al.* [10] also decomposed motion and content for video generation. In [26] future-frame predictions are made consistent with the pixel-wise flows in videos through a dual learning mechanism. Other works introduce recurrent networks into video generation (*e.g.* [27, 28]). In line with these works, our method separately models the motion and appearance as well, using the PSGAN and SCGAN respectively. This enables us to control the motion patterns explicitly and interpretably, which to the best of our knowledge, is the first attempt in human video generation.

3   Methodology

3.1   Framework Overview

Given an input image of human body or face and a target action class (*e.g.*, *Skip*, *TaiChi*, *Jump*), our goal is to synthesize a video of human action or facial expression belonging to the target category and starting with the input image. We wish to explicitly control the motion patterns in the generated video while maintaining appearance coherence with the input. We here propose to generate human videos in a disentangled way: *plausible motion prediction* and *coherent appearance generation*. Fig. 1 illustrates the overall framework of our method.

Similar to the action recognition literature [11–14], we use the human skeletons or human pose for representations of motion dynamics. Our method consists of two stages. In the first stage, we extract the pose from input image and the Pose Sequence GAN (PSGAN) is proposed to generate a temporally smooth pose sequence conditioned on the pose of input image and the target action class. In the second stage, we focus on appearance modeling and propose a Semantic Consistent GAN (SCGAN) to generate realistic and coherent video frames conditioned on the input image and pose sequence from stage one. The impact of noisy/abnormal poses is alleviated by maintaining semantic consistency between generated and ground-truth poses in high-level representation space. Details will be elaborated in the following sections.

3.2   Plausible Motion Prediction

In the first stage, the human pose extracted from input image together with the target action label is fed into our PSGAN to generate a sequence of poses. Obviously this is an ill-posed one-to-many problem with infinite possibilities. Our PSGAN learns from example pose sequences in the training set to mimic plausible motion patterns. Thus our learning objective is the modeling of rich motion patterns instead of the precise pose coordinates.
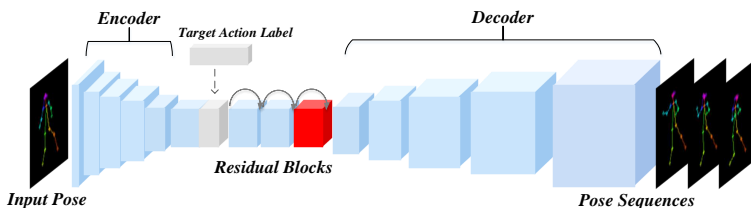
Fig. 2. Network architecture of our Pose Sequence GAN (PSGAN). PSGAN takes the input pose and target action label as input, and synthesizes pose sequences in an encoder-decoder manner. After the last residual block (red), the feature map is extended with a time dimension and then fed into the decoder which is composed of a series of fractionally-strided spatio-temporal convolution layers

Pose extraction. To extract the initial pose from input image, a state-of-the-art pose estimator in [15] is adopted to produce the coordinates of 18 key points. The pose is encoded by 18 heatmaps rather than coordinate vectors of the key points. Each heatmap is filled with 1 within a radius of 4 pixels around the corresponding key point and 0 elsewhere. Consequently, pose is actually represented as a $C = 18$ channel tensor. In this way, there is no need to learn how to map the key points into body part locations.

Pose Sequence GAN. Given the initial pose and the target action label, our PSGAN aims to synthesize a meaningful pose sequence at one time. As shown in Fig. 2, PSGAN adopts an encoder-decoder architecture. The $C \times W \times H$-sized pose is first encoded through several convolutional layers. The target action label is also inputted in the form of n-dimensional one-hot vector where n denotes the number of action types. After a few residual blocks, the two signals are embedded into common feature maps in the latent space. These feature maps will finally go through the decoder with a extended time dimension. The output is a $C \times T \times W \times H$-sized tensor via a series of fractionally-strided spatio-temporal convolution layers, where T denotes the number of time steps in the sequence. For the sake of better temporal modeling, the LSTM module [29] is integrated into our network as well. In summary, we define a generator G that transforms an input pose p into a pose sequence $\hat{P}$ conditioned on the target action label a, *i.e.*, $G(p, a) \Rightarrow \hat{P}$. We train the generator G in an adversarial way - it competes with a discriminator D as the PatchGAN [30] which classifies local patches from the ground-truth and generated poses as real or fake.

LSTM embedding. As is mentioned above, the decoder outputs a $C \times T \times W \times H$ tensor. It can be regarded as T tensors with size $C \times W \times H$, all of which are fed into a one-layer LSTM module for temporal pose modeling. Our experiments will demonstrate that the LSTM module stabilizes training and improves the quality of generated pose sequences.

Objective function. As in [2], the objective functions of our PSGAN can be formulated as follows:

$$\mathcal{L}_{adv}^{D} = \mathbb{E}_{P}[\log D(P)] + \mathbb{E}_{p,a}[\log(1 - D(G(p, a)))], \tag{1}$$

$$\mathcal{L}_{adv}^{G} = \mathbb{E}_{p,a}[\log(D(G(p, a)))], \tag{2}$$

where $\mathcal{L}_{adv}^{D}$ and $\mathcal{L}_{adv}^{G}$ denote the adversarial loss terms for discriminator D and generator G, respectively. The discriminator D aims to distinguish between the generated pose sequence $G(p,a)$ and ground-truth P. Moreover, we find adding a reconstruction loss term can stabilize the training process. The reconstruction loss is given below:

$$\mathcal{L}_{rec} = \lambda_{rec}||(P - G(p, a)) \odot (\alpha M + 1)||_1, \tag{3}$$

where M denotes the mask for each of the key point heatmap, $\odot$ denotes pixels-wise multiplication and λ_{rec} is the weight for this $L1$ loss. The introduction of mask M is due to the heatmap sparsity and imbalance of each key point, which makes learning difficult. We use the ground-truth P as mask M to mask out the small region around each key point for loss computation. Note when the scaling factor $\alpha = 0$, this loss term is reduced to unweighted $L1$ loss.

Abnormal poses. Fig. 3 shows some bad pose generation results where some key points seem bigger/smaller (b) than the ground-truth (a), or some key points seem missing (c) because of their weak responses. We call such cases as abnormal poses. For human beings however, abnormal poses might look weird at the first glance, but would hardly prevent us from imagining what the "true" pose is. This requires our network to grasp the semantic implication of human pose, and to alleviate the influence of small numerical differences.
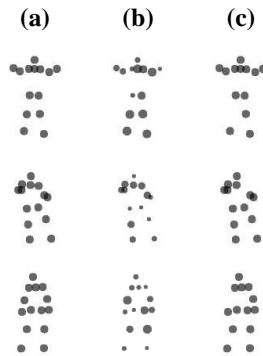

Fig. 3. Examples of abnormal poses. (a-c) show the ground-truth pose, generated poses with bigger/smaller key point responses, and with missing key points respectively
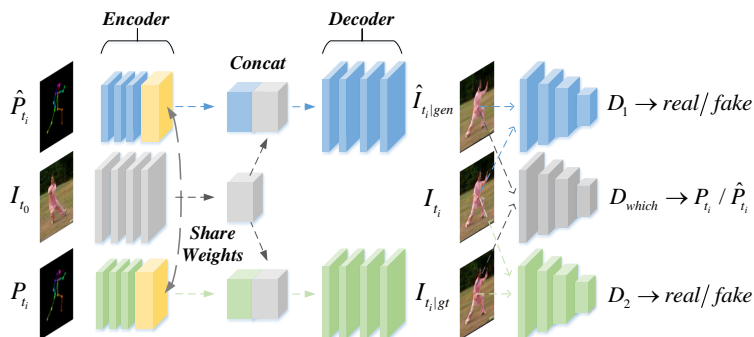
Fig. 4. Network architecture of our SCGAN in the second stage, where $\hat{P}_{t_i}, P_{t_i}, I_{t_i}$ respectively denote the pose generated by our method in stage one, ground-truth pose and the original image. Our generator has an encoder-decoder architecture and generates video frames conditioned on human poses P and the input image I_{t_0}. Discriminators D_1 and D_2 aim to distinguish whether the generated images are real, while D_{which} aims to tell which pose the frame is generated from

3.3   Coherent Appearance Generation

In the second stage, we aim to synthesize coherent video frames conditioned on the input image as well as the pose sequence from stage one. Since the noisy or abnormal poses will affect image generation in this stage, those methods (*e.g.* [20]) to directly generate images from pose input may be unstable or even fail. Therefore, we propose a Semantic Consistent GAN (SCGAN) to impose semantic consistency between generated pose and ground-truth at a high feature level. By only imposing the consistency at high feature level, SCGAN can be robust to noisy pose inputs.

Conditional generation. Our conditional image generation process is actually similar to that in recent work [20] which can generate person images controlled by pose. However, we have a major difference with this work: in [20] images are generated in two stages by synthesizing a coarse image first and then refining it; while our SCGAN generates results in one step, for all video frames over time at once. Specifically, given the input image I_{t_0} at time t_0 and the target pose P_{t_i} at time t_i, our generator $G(I_{t_0}, P_{t_i}) \Rightarrow \hat{I}_{t_i}$ is supposed to generate image $\hat{I}_{t_i}$ to keep the same appearance in I_{t_0} but on the new pose P_{t_i}. We design the discriminator D again to tell real image from fake to improve the image generation quality.

Semantic consistency. As discussed before, the noisy or abnormal pose prediction $\hat{P}_{t_i}$ from the first stage will affect image generation in the second stage. Unfortunately, the ground-truth pose P_{t_i} does not exist during inference for

pose correct purposes - it is only available for training. Therefore, it is necessary to teach training to properly handle abnormal poses with the guidance of ground-truth pose, in order to generalize to testing scenarios.

Through the observation of heatmaps of those abnormal poses, we find that they are often due to the small differences in corresponding key point responses, which will not contribute to large loss and thus incur small back-propagation gradients. As a matter of fact, there is no need to push the limit of the pose generation accuracy by PSGAN since the small errors should not affect how people interpret pose globally. Considering the pose prediction difference is inevitably noisy at the input layer or low-level feature layer, we propose to enforce the semantic consistency between abnormal poses and the ground-truth at the high-level feature layer.

Fig. 4 shows our Semantic Consistent GAN that encapsulates this idea. We share weights in the last convolutional layer of two pose encoder networks (the yellow block), aiming to impose semantic consistent in the high-level feature space. Moreover, we generate video frames from both predicted pose and ground-truth pose to gain tolerance to pose noise. A new discriminator D_{which} is used to distinguish which pose the generated video frame is conditioned on. We further utilize the $L1$ reconstruction loss to stabilize the training process.

Full objective function. As shown in Fig. 4, our final objective is to generate video frames from two pose streams and keep their semantic consistency in an adversarial way. Specifically, G_1 generates the image $I_{t_i|gen}$ at time t_i conditioned on input image I_{t_0} and the pose $\hat{P}_{t_i}$ generated by PSGAN. G_2 generates $I_{t_i|gt}$ in the same way but uses the ground-truth pose for image generation.

$$G_1(I_{t_0}, \hat{P}_{t_i}) \Rightarrow I_{t_i|gen}, \tag{4}$$

$$G_2(I_{t_0}, P_{t_i}) \Rightarrow I_{t_i|gt}. \tag{5}$$

There are three discriminators defined as follows: D_1 and D_2 aim to distinguish between real image and fake when using predicted pose and ground-truth pose respectively; D_{which} aims to judge which pose the generated image is conditioned on. Then we easily arrive at the full objective function for our model training as follows:

$$\mathcal{L}_{D_{which}} = \mathbb{E}[\log(D_{which}(I_{t_i|gt}))] + \mathbb{E}[\log(1 - D_{which}(I_{t_i|gen}))], \tag{6}$$

$$\mathcal{L}_{D_1} = \mathbb{E}[\log(D_1(I_{t_i}))] + \mathbb{E}[\log(1 - D_1(I_{t_i|gen}))], \tag{7}$$

$$\mathcal{L}_{D_2} = \mathbb{E}[\log(D_2(I_{t_i}))] + \mathbb{E}[\log(1 - D_2(I_{t_i|gt}))], \tag{8}$$

$$\mathcal{L}_{G_1} = \mathbb{E}[\log(D_1(I_{t_i|gen}))] + \mathbb{E}[\log(D_{which}(I_{t_i|gen}))], \tag{9}$$

$$\mathcal{L}_{G_2} = \mathbb{E}[\log(D_2(I_{t_i|gt}))]. \tag{10}$$

Since the ground-truth pose guided images $I_{t_i|gt}$ is *real* for D_{which}, the gradient of D_{which} is not propagated back to G_2 in Eq. (10).

3.4   Implementation Details

For our detailed network architecture, all of the generators (G, G_1, G_2) apply 4 convolution layers with a kernel size of 4 and the stride of 2 for downsampling. In the decoding step of stage one, transposed convolution layers with stride of 2 are adopted for upsampling, while normal convolutions layers together with interpolation operations take place of transposed convolution layers in the second stage. The feature map of the red block in Fig. 2 is extended with a time dimension $(C \times W \times H \Rightarrow C \times 1 \times W \times H)$ for the decoder of PSGAN. The discriminators (D, D_1, D_2, D_{which}) are PatchGANs [30] to classify whether local image patches are real or fake. Besides, ReLU [31] serves as the activation function after each layer and the instance normalization [32] is used in all networks. Several residual blocks [33] are leveraged to encode the concatenated feature representations jointly. For the last layer we apply tanh activation function. In addition, we use standard GRU in our PSGAN, without further investigating how the different structures of LSTM can improve pose sequence generation.

We implement all models using PyTorch, and use an ADAM [34] optimizer with a learning rate of 0.001 in all experiments. The batch size in stage one is 64, and 128 in the second stage. All reconstruction loss weights are empirically set to 10. The scaling factor α in Eq.(3) is chosen from 0 to 100, which only affects the convergence speed. We empirically set the scaling factor as 10 and 20 on the human action and facial expression dataset respectively. The PSGAN is trained to generate pose sequences. In the second stage, both the generated and ground-truth poses are utilized to train the SCGAN to learn robust handling of noisy poses. Only the generated pose is fed into SCGAN for inference.

4   Experiments

In this section, we present video generation results on both human action and facial datasets. Qualitative and quantitative comparisons are provided to show our superiority over other baselines and state-of-the-arts. User study is also conducted with a total of 50 volunteers to support our improvements. Ablation study for our two generation stages (for pose and video) is further included to show their efficacy.

4.1   Datasets

Our experiments are conducted not only on the human action dataset but also on facial expression dataset, where facial landmarks act as the pose to guide the generation of facial expression videos. Accordingly, we collected the Human Action Dataset and Human Facial Dataset as detailed below. For all experiments, the RGB images are scaled to 128×128 pixels while pose images are scaled to 64×64 pixels.

- Human Action Dataset comes from the UCF101 [35] and Weizmann Action database [36], including 198848 video frames of 90 persons performing 22 actions. Human pose is extracted by the method in [15] with 18 key points.
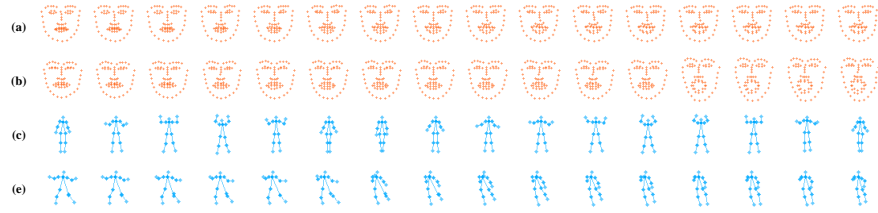
Fig. 5. Example pose sequences generated by our PSGAN with class labels of *Happy* (a), *Surprise* (b), *Wave* (c) and *TaiChi* (d), respecitvely

- Human Facial Dataset is from the $CK+$ dataset [37]. We consider 6 facial expressions: *angry*, *disgust*, *fear*, *happy*, *sadness* and *surprise*, corresponding to 60 persons and 60000 frames. The facial pose is annotated with 68 key points.

4.2   Evaluation of Pose Generation

Qualitative evaluation. As mentioned in Section 3.3, our PSGAN focuses on the generation of various pose motions. For qualitative comparisons, we follow the post-processing step in [15] to locate the maximum response region in each pose heatmap. Note such pose-processing is only used for visualization purposes. Fig. 5 shows some examples of the generated pose sequences for both human face and body. We can see that the pose sequences change in a smooth and typical way under each action scenario.

Quantitative comparison. Recall that our final video generator is tolerant to the tiny pose difference in stage one. Therefore, we measure the quality of generated pose sequences by calculating the average pairwise $L2$ distance rather than Euclidean norm between the generated and the ground-truth poses. Smaller distance indicates better pose quality.

   We compare three of our PSGAN variants: 1) PSGAN trained with the $L1$-norm loss rather than adversarial loss, 2) PSGAN trained without the LSTM module, and 3) the full PSGAN model with a GRU module. Table 1 indicates that it is better to train our pose generator with the adversarial loss than with the simple $L1$-norm loss. Also important is the temporal modeling by the GRU or LSTM module that improves the quality of pose sequences.

User study. Table 2 includes the user study results for our three PSGAN variants on human action dataset. For each variant, we generate 25 pose sequences with 20 actions and the time step of 32. All generated pose sequences are shown to 50 users in a random order. Users are then asked to rank the baselines based on their quality from 1 to 4 (best to worst). The distribution for the ranks of

Table 1. Quantitative comparison of pose generation baselines

Average $L2$	Action	Facial Exp.
Ground-truth	0	0
PSGAN-$L1$	0.0124	0.0078
PSGAN w/o LSTM	0.0072	0.0062
PSGAN	**0.0064**	**0.0051**

Table 2. User study of pose generation baselines on human action dataset

Distribution of ranks	1	2	3	4
Ground-truth	0.38	0.36	0.12	0.14
PSGAN-$L1$	0.09	0.08	0.32	0.51
PSGAN w/o LSTM	0.21	0.16	0.43	0.20
PSGAN	0.32	0.40	0.13	0.15

each baseline is calculated for comparison. As shown in Table 2, our full PS-GAN model has the highest chance to rank top. While the variants of PSGAN w/o LSTM and PSGAN-$L1$ tend to rank lower, indicating the importance of temporal and adversarial pose modeling again.

4.3 Evaluation of Video Generation

Qualitative comparison. Given the generated pose sequence in the first stage and the input image, our SCGAN is responsible for the generation of photo-realistic video frames. We mainly compare our method with state-of-the-art video generation methods VGAN [8] and MoCoGAN [10]. They are trained on the same human action and facial datasets with hyper-parameters tuned to their best performance. The visual results for example action and facial expression classes *Wave*, *Taichi* and *Superised* are shown in Fig. 6.

It is clear that our method generates much sharper and more realistic video frames than VGAN and MoCoGAN. For the simple action *Wave*, our method performs better or equally well with the strong competitors. For the difficult action *TaiChi* with complex motion patterns, our advantage is evident - the pose dynamics are accurately captured and rendered to visually pleasing images. This confirms the necessity of our pose-guided video generation which benefits from explicit pose motion modeling rather than using a noise vector in VGAN and MoCoGAN. Our supplementary material provides more visual results.

Quantitative comparison. Table. 3 shows the measures of Inception Score [38] (IS) (and its variance) for different methods. Larger IS value indicates better performance. Such quantitative results are in line with our visual evaluations where our method outperforms others by a large margin.

Table 3. Comparison of IS for video generation baselines

IS	Action	Facial Exp.
VGAN [8]	2.73 ± 0.21	1.68 ± 0.17
MoCoGAN [10]	4.02 ± 0.27	1.83 ± 0.08
Ours	**5.70 ± 0.19**	**1.92 ± 0.12**

Table 4. User Study for video generation baselines

Winning percentage	Action	Facial Exp.
Ours/MCGAN [10]	0.83/0.17	0.86/0.14
Ours/VGAN [8]	0.88/0.12	0.93/0.07

VGAN [8]

MoCoGAN [10]

Ours

Fig. 6. Generated video frames for example action and facial expression classes *Wave*, *Taichi* and *Superised* by VGAN [8], MoCoGAN [10] and our method

User study. We further conduct a user study where each method generates 50 videos for comparisons. Results are provided to users in pairs and in random order. The users are then asked to select the winner (looks more realistic) from the paired methods and we calculate the winning percentage for each method. Table 4 demonstrates that most of the time users would choose our method as the winner over MoCoGAN and VGAN.

Controlled generation results. Fig. 7 validates our capability of explicit pose modeling and good generalization ability by controlled tests: generate different action videos with a fixed human appearance, and generate videos with a fixed action for different humans. Successes are found for both human action (a-c) and facial expression (d-f) cases, showing the benefits of separate modeling for pose and appearance.
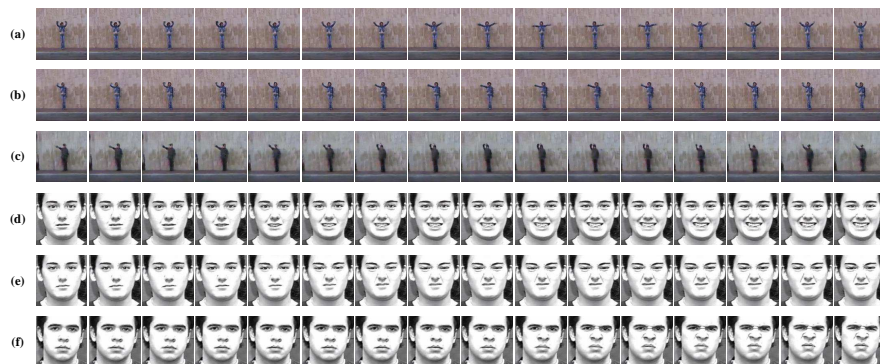
Fig. 7. Controlled video generation with pose and appearance: different poses for the same human (a-b for body, d-e for face), and same pose on different humans (b-c for body, e-f for face)

4.4   Ablation Study

One major feature of our human video generator is its reliance on generated pose $\hat{P}_{t_i}$. It is worth noting that there is no ground-truth pose P_{t_i} during inference. Only for training, we use the available P_{t_i} to enforce the semantic consistency with respect to generated pose $\hat{P}_{t_i}$. To highlight the impact of the semantic consistency constraint, we compare several training alternatives as follows:

- Static generator: video generation by repeating the first frame
- SCGAN-gen: video generation guided by generated pose $\hat{P}_{t_i}$ only.
- SCGAN-gt: video generation guided by ground-truth pose P_{t_i} only.
- SCGAN: video generation guided by both $\hat{P}_{t_i}$ and P_{t_i} as shown in Fig. 4.

The baseline of static generator simply constructs a video by repeating the first frame thus involves no predictions. It acts as a performance lower bound here. Sometimes its performance can be not too bad in those short videos or videos with little change (generating a static video will not be heavily penalized in these cases).

Table 5. SSIM and LPIPS measures for our training alternatives

SSIM/LPIPS	Action	Facial Exp.
Static	0.66/0.063	0.77/0.025
SCGAN-gen	0.73/0.083	0.89/0.038
SCGAN-gt	0.89/0.040	0.92/0.024
SCGAN	0.87/0.041	0.91/0.026

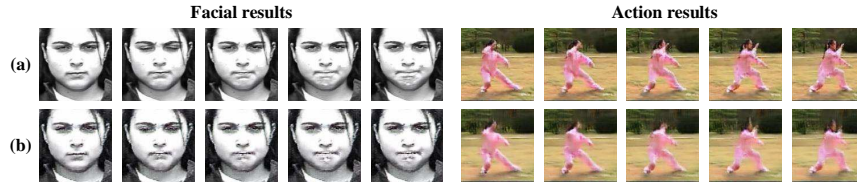**Facial results**   **Action results**

Fig. 8. Visual comparisons of SCGAN (a) and SCGAN-gen (b) on facial expression and human action datasets

Fig. 8 visually compares the baselines of SCGAN-gen and SCGAN. The full SCGAN model can generate sharper and more photo-realistic results, especially in the **mouth** (facial expression) and **waist** (human action) regions. This suggests the efficacy of enforcing semantic consistency between the generated and ground-truth poses. Using the generated pose only can be noisy and thus hinder the final video quality.

We also evaluate performance by computing the SSIM (structural similarity index measure) [39] and LPIPS (Learned Perceptual Image Patch Similarity) scores [40]. The SSIM score focuses on structural similarity between the generated image and ground-truth, while the LPIPS score cares more about perceptual similarity. Higher SSIM score and smaller LPIPS score indicates better performance.

Table 5 shows that SCGAN indeed outperforms SCGAN-gen quantitatively and stands close to the SCGAN-gt using ground-truth pose. The semantic consistency constraint plays a key role in this improvement, because it can alleviate the influence of abnormal poses during the pose guided image generation process. When compared to the static video generator, our method outperforms by generating a variety of motion patterns.

5   Conclusion and Future Work

This paper presents a novel method to generate human videos in a disentangled way. We show the important role of human pose for this task, and propose a pose-guided method to generate realistic human videos in two stages. Quantitative and qualitative results on human action and face datasets demonstrate the superiority of our method, which is also shown to be able to manipulate human pose and appearance explicitly. Currently, our method is limited to cropped human or face images since the detectors are missing. In the future, we will integrate detectors as an automatic pre-processing step which will enable multi-person video generation.

References

1. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. ICLR (2014) 1, 3
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS. (2014) 1, 3, 6
3. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using lstms. In: ICML. (2015) 1, 4
4. Finn, C., Goodfellow, I., Levine, S.: Unsupervised learning for physical interaction through video prediction. In: NIPS. (2016) 1, 4
5. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. ICLR (2016) 1, 4
6. Oh, J., Guo, X., Lee, H., Lewis, R.L., Singh, S.: Action-conditional video prediction using deep networks in atari games. In: NIPS. (2015) 1, 4
7. Denton, E.L., et al.: Unsupervised learning of disentangled representations from video. In: NIPS. (2017) 2
8. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: NIPS. (2016) 2, 4, 11, 12
9. Saito, M., Matsumoto, E., Saito, S.: Temporal generative adversarial nets with singular value clipping. In: ICCV. (2017) 2
10. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. arXiv preprint arXiv:1707.04993 (2017) 2, 4, 11, 12
11. Hodgins, J.K., O'Brien, J.F., Tumblin, J.: Perception of human motion with different geometric models. IEEE Transactions on Visualization and Computer Graphics (1998) 2, 4
12. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. AAAI (2018) 2, 4
13. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: CVPR. (2015) 2, 4
14. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: CVPR. (2014) 2, 4
15. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. (2017) 2, 5, 9, 10
16. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. ICLR (2015) 3
17. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv preprint arXiv:1701.07875 (2017) 3
18. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: NIPS. (2017) 3
19. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014) 3
20. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. In: NIPS. (2017) 3, 7
21. Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., Metaxas, D.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV. (2017) 3
22. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. ICCV (2017) 3
23. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. ICCV (2017) 3

24. Kim, T., Cha, M., Kim, H., Lee, J., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. arXiv preprint arXiv:1703.05192 (2017) 3

25. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. CVPR (2018) 3

26. Liang, X., Lee, L., Dai, W., Xing, E.P.: Dual motion gan for future-flow embedded video prediction. ICCV (2017) 4

27. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: ICCV, IEEE (2015) 4

28. Zhou, Y., Berg, T.L.: Learning temporal transformations from time-lapse videos. In: ECCV, Springer (2016) 4

29. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation (1997) 5

30. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CVPR (2017) 5, 9

31. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML. (2010) 9

32. TDmitry Ulyanov, Andrea Vedaldi, V.L.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016) 9

33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016) 9

34. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. ICLR (2014) 9

35. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012) 9

36. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV, IEEE (2005) 9

37. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, IEEE (2010) 10

38. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: NIPS. (2016) 11

39. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP (2004) 14

40. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. CVPR (2018) 14