2019 IEEE International Conference on Multimedia and Expo (ICME)

STOCHASTIC VIDEO GENERATION WITH DISENTANGLED REPRESENTATIONS

Maomao Li[1], *Chun Yuan*[1], *Zhihui Lin*[12], *Zhuobin Zheng*[12], *Yangyang Cheng*[12]

[1]Graduate School at Shenzhen,Tsinghua University
[2]Department of Computer Science and Technologies, Tsinghua University
{mm-li17, lin-zh14, zhengzb16,cheng-yy13}@mails.tsinghua.edu.cn, yuanc@sz.tsinghua.edu.cn

ABSTRACT

Frame-to-frame uncertainty is a major challenge in video prediction. The use of the deterministic models always leads to averaging of future states. Some methods draw samples from a prior at each time step to deal with the uncertainty of the future states, such as the SVG model [1]. However, these models always use only one set of latent variables to represent the whole stochastic part in a video clip whereas sequential data often involves multiple independent factors. In this paper, we exploit the complex representation of information in video sequences by formulating it explicitly with a disentangled-representation stochastic video generation (DR-SVG) model that imposes sequence-dependent prior and sequence-independent prior to different sets of latent variables. Through a variational lower-bound and adversarial objective functions in latent space, our model can produce crisper frames with clear content and pose which indicate the sequence-dependent and sequence-independent component respectively.

Index Terms— stochastic video prediction, disentangled-representation, variational inference, adversarial learning

1. INTRODUCTION

Video prediction is one of the most fundamental and difficult tasks in computer vision. Most video prediction studies focus on using deterministic models to perform frame-to-frame transitions. The limitation of a deterministic network is apparent when Villegas et al. [2] applied their prediction network to the Moving-MNIST dataset [3]. When a digit hits image edges, the future trajectory is inherently random for the velocity and direction are under an uncertain effect. Recently, some stochastic video prediction approaches have sought to overcome this limitation. Babaeizadeh et al. [4] proposed a latent variable model which introduce a vector of latent variables $\mathbf{z}$ accounting for the complex stochastic phenomena of videos. Denton et al. [1] proposed a stochastic video generation model (SVG) that combines a deterministic frame predictor with time-dependent stochastic latent variables which follow a learned prior or a fixed prior.

Despite successes with the aforementioned models, video generation often involves multiple independent factors operating in the latent space. This behavior results in the fact that some attributes tend to have a smaller amount of variation within a video sequence, while other attributes tend to have a smaller amount of variation between video sequences. We refer to the first type attributes as sequence-level attributes, and the other as image-level attributes. However, existing methods use only one set of latent variables to represent stochastic phenomena of the entire video sequence. In this context, there is always no guarantee that inference on stochastic models with tangled factors will maintain the sequence-level attributes. This will make these models suffer from the problem of changing content when they predict frames in a specific video clip. Take, for example, when the SVG [1] model is applied to predict numbers in a video sequence of the Moving MNIST dataset [3], the numbers themselves usually change. How can we build a stochastic generation model that can learn disentangled representations from video sequences?

Learning disentangled representations aims to model the factors of data variations. β-VAE [5] automatically discovered interpretable factorised latent representations via a modification of the variational autoencoder (VAE) [6] framework. InfoGAN [7] maximized the mutual information between a small subset of the latent variables and the observation to achieve disentanglement. Unlike the above methods, we are interested in learning disentangled representations that encode distinct aspects of the stochastic part of video sequences into separate variables. Specifically, we encode sequence-level attributes and image-level attributes into content variables $\mathbf{z}^c$ and pose variables $\mathbf{z}^p$ separately. Moreover, the former follows a sequence-dependent prior, and the latter obeys a sequence-independent prior. To preserve the sequence-level attributes of content variables, we argue that content prior should be consistent at every time step. Therefore, we take the content posterior of the last time step $q_{\phi_c}(\mathbf{z}^c_{t-1}|\mathbf{x}_{1:t-1})$ as the content prior at current time step. On the other hand, inspired by [2], we exploit an adversarial framework in the latent space to encourage the pose variables carries *no* content information, obtaining a sequence-independent and a sequence-dependent latent variable space. To sum up, the contributions of the paper are highlighted as follows:

- We propose DR-SVG, which factorizes stochastic phenomena in video sequences into a sequence-dependent

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00047

component (content) and a sequence-independent part (pose) for stochastic video generation.

- DR-SVG achieves disentanglement by driving the content representations of adjacent time steps close to each other and introducing adversarial loss terms in the latent space to eliminate pose information from content representations.
- Our approach is easily trained end-to-end on a range of synthetic and real videos, and compare favorably to the baseline model SVG, generating video clips with more consistent content.

2. BACKGROUND

In this section, we review the SVG model, which is able to generate a number of different sequences into the future. It has two variants: one with a fixed prior over the latent variables (SVG-FP) and another with a learned prior (SVG-LP). These two variants share the same framework but differ in how to obtain the prior. In order to explain the two models clearly, we adopt the formalism of variational auto-encoders. SVG-FP is trained by optimizing the following variational lower bound:

$$\begin{aligned}\mathcal{L}_{FP}(\mathbf{x}_{1:T}) = & \sum_{t=1}^{T}[\mathbb{E}_{q_\phi(\mathbf{z}_{1:t}|\mathbf{x}_{1:t})}\log p_\theta(\mathbf{x}_t|\mathbf{x}_{1:t-1},\mathbf{z}_{1:t}) \\ & -\beta D_{KL}(q_\phi(\mathbf{z}_t|\mathbf{x}_{1:t})||p(\mathbf{z}))].\end{aligned} \tag{1}$$

$p_\theta(\mathbf{x}_t|\mathbf{x}_{1:t-1},\mathbf{z}_{1:t})$ denotes a frame predictor (generative model) that is specified by a fixed-variance conditional Gaussian distribution $\mathcal{N}(\mu_\theta(\mathbf{x}_{1:t-1},\mathbf{z}_{1:t}),\sigma)$. $q_\phi(\mathbf{z}_t|\mathbf{x}_{1:t})$ indicates an inference model that is usesd to approximate the posterior during training time and outputs a conditional Gaussian distribution $\mathcal{N}(\mu_\phi(\mathbf{x}_{1:t}),\sigma_\phi(\mathbf{x}_{1:t}))$. The prior $p(\mathbf{z})$ is distributed according to the standard Gaussian $\mathcal{N}(0,\mathbf{I})$ at each time step,as assumed in many other papers [8, 4, 9, 10, 11]. β is a hyper-parameter that controls the trade-off between prediction error and regularization term between the variational posterior and the prior. Similarly, SVG-LP model is trained by maximizing:

$$\begin{aligned}\mathcal{L}_{LP}(\mathbf{x}_{1:T}) = & \sum_{t=1}^{T}[\mathbb{E}_{q_\phi(\mathbf{z}_{1:t}|\mathbf{x}_{1:t})}\log p_\theta(\mathbf{x}_t|\mathbf{x}_{1:t-1},\mathbf{z}_{1:t}) \\ & -\beta D_{KL}(q_\phi(\mathbf{z}_t|\mathbf{x}_{1:t})||p_\psi(\mathbf{z}_t|\mathbf{x}_{1:t-1})],\end{aligned} \tag{2}$$

where $p_\psi(\mathbf{z}_t|\mathbf{x}_{1:t-1})$ is a learned prior based on the observed frames, which is specified by a conditional Gaussian distribution $\mathcal{N}(\mu_\psi(\mathbf{x}_{1:t-1}),\sigma_\psi(\mathbf{x}_{1:t-1}))$. Note that in the SVG model, there is only one set of latent variables to represent the stochastic phenomena of videos whereas this representation should be tangled and involves multiple independent factors.

3. DISENTANGLED-REPRESENTATION STOCHASTIC VIDEO GENERATION MODEL

In this section, we present a novel disentangled-representation stochastic video generation (DR-SVG) model, which follows the line of research of combining deterministic frame predictor with random latent variables to deal with the inherent uncertainty in videos. Our model factorizes the stochastic part in each video clip into content and pose via a variational lower bound and an adversarial learning process in the latent space.

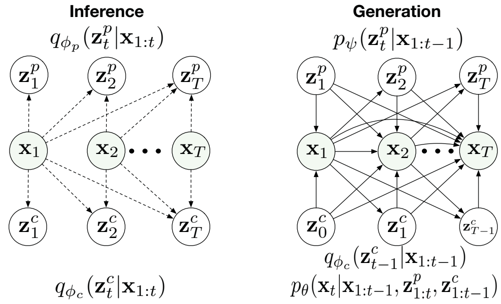

Fig. 1. The probabilistic graphical model of DR-SVG. The variational inference model (left) approximates the posterior given the observed frames (dotted lines). The generative model (right) predicts the next frame conditioned on the previous frames and two set of latent variables (solid lines).

3.1. Disentanglement via a Variational Lower-bound

In order to construct our DR-SVG model, we formulate a probabilistic graphical model, explaining the inference and generative process, as shown in Fig.1. To preserve the sequence-level attributes of content variables, we argue that content prior should be consistent at every time step. Therefore, we take the content posterior of the last time step $q_{\phi_c}(\mathbf{z}_{t-1}^c|\mathbf{x}_{1:t-1})$ as the content prior at current time step. For sequence-independent attributes, we impose an adversarial framework in the latent space to eliminate pose information from content representation, which will be detailed in the next subsection. Besides, similar to the SVG-LP [1], we adopt a sophisticated approach to get the pose prior. Specifically, instead of applying the standard Gaussian distribution $\mathcal{N}(0,\mathbf{I})$, our pose prior $p_\psi(\mathbf{z}_t^p|\mathbf{x}_{1:t-1})$ follows a learnable Gaussian $\mathcal{N}(\mu_\psi(\mathbf{x}_{1:t-1}),\sigma_\psi(\mathbf{x}_{1:t-1}))$, being a function of all past frames up to but not including the frame being predicted.

Since the exact posterior distribution is intractable, an inference model $q_\phi(\mathbf{z}_t^p,\mathbf{z}_t^c|\mathbf{x}_{1:t})$ is used to approximate the ture posterior. We consider the following inference model as in Fig.1 (left):

$$q_\phi(\mathbf{z}_t^p,\mathbf{z}_t^c|\mathbf{x}_{1:t}) = q_{\phi_p}(\mathbf{z}_t^p|\mathbf{x}_{1:t})q_{\phi_c}(\mathbf{z}_t^c|\mathbf{x}_{1:t}), \tag{3}$$

where the pose-inference model $q_{\phi_p}(\mathbf{z}_t^p|\mathbf{x}_{1:t})$ is specified by a conditional Gaussian distribution $\mathcal{N}(u_{\phi_p}(\mathbf{x}_{1:t}),\sigma_{\phi_p}(\mathbf{x}_{1:t}))$ and content-inference model $q_{\phi_c}(\mathbf{z}_t^c|\mathbf{x}_{1:t})$ is distributed according to another conditional Gausssian distribution $\mathcal{N}(u_{\phi_c}(\mathbf{x}_{1:t}),\sigma_{\phi_c}(\mathbf{x}_{1:t}))$. Besides, we force $q_{\phi_p}(\mathbf{z}_t^p|\mathbf{x}_{1:t})$ to close the pose prior $p_\psi(\mathbf{z}_t^p|\mathbf{x}_{1:t-1})$ and drive $q_{\phi_c}(\mathbf{z}_t^c|\mathbf{x}_{1:t})$ to approach the content prior $q_{\phi_c}(\mathbf{z}_{t-1}^c|\mathbf{x}_{1:t-1})$ as regularization terms. The variational lower bound for this inference model on the marginal likelihood is as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{x}_{1:T}) \; = \sum_{t=1}^{T}[\mathbb{E}_{q_{\phi}(\mathbf{z}^c_{1:t},\mathbf{z}^p_{1:t}|\mathbf{x}_{1:t})}\log p_{\theta}(\mathbf{x}_t|\mathbf{x}_{1:t-1},\mathbf{z}^c_{1:t},\mathbf{z}^p_{1:t}) \\ -\beta D_{KL}(q_{\phi_c}(\mathbf{z}^c_t|\mathbf{x}_{1:t})||q_{\phi_c}(\mathbf{z}^c_{t-1}|\mathbf{x}_{1:t-1})) \\ -\beta D_{KL}(q_{\phi_p}(\mathbf{z}^p_t|\mathbf{x}_{1:t})||p_{\psi}(\mathbf{z}^p_t|\mathbf{x}_{1:t-1}))],\end{aligned} \tag{4}$$

where the first term on the RHS represents the negative prediction loss. D_{KL} indicates the Kullback-Leibler divergence between the approximated posterior and assumed prior which in our case are the content posterior of the last time step $q_{\phi}(\mathbf{z}^c_{t-1}|\mathbf{x}_{1:t-1})$ and the learnable Gaussian $p_{\psi}(\mathbf{z}^p_t|\mathbf{x}_{1:t-1})$ respectively. The hyper-parameter β is less than 1 for a good reconstruction quality [1, 12].

In terms of the generative model, we argue that each frame is generated from some random process which involves the pose variables $\mathbf{z}_p$, the content variables $\mathbf{z}_c$, and the deterministic features derived from observed frames. The recurrent frame predictor $p_{\theta}(\mathbf{x}_t|\mathbf{x}_{1:t-1},\mathbf{z}^c_{1:t},\mathbf{z}^p_{1:t})$ is specified by a fixed-variance conditional Gaussian distribution $\mathcal{N}(\mu_{\theta}(\mathbf{x}_{1:t-1},\mathbf{z}^c_{1:t},\mathbf{z}^p_{1:t}),\sigma)$ during training time. That is, we set $\hat{\mathbf{x}}_t = \mu_{\theta}(\mathbf{x}_{1:t-1},\mathbf{z}^c_{1:t},\mathbf{z}^p_{1:t})$ in practice, where both variables are sampled using re-parameterization trick [6]. Note that the frame predictor receives $\mathbf{x}_{t-1}$ and the concatenation of $\mathbf{z}^c_t$ and $\mathbf{z}^p_t$ at time step t. The dependencies on the previous $\mathbf{x}_{1:t-2}$, $\mathbf{z}^c_{1:t-1}$ and $\mathbf{z}^p_{1:t-1}$ derive from the model recurrence.

As illustrated in Fig.1 (right), at test time, instead of sampling the latent variables from the approxiamated posterior distributions $q_{\phi_c}(\mathbf{z}^c_t|\mathbf{x}_{1:t})$ and $q_{\phi_p}(\mathbf{z}^p_t|\mathbf{x}_{1:t})$, we sample them from the learned pose prior $p_{\psi}(\mathbf{z}^p_t|\mathbf{x}_{1:t-1})$ and the content posterior of the last time step $q_{\phi_c}(\mathbf{z}^c_{t-1}|\mathbf{x}_{1:t-1})$, respectively. Then the frame $\hat{\mathbf{x}}_t$ is generated by $\hat{\mathbf{x}}_t = \mu_{\theta}(\mathbf{x}_{1:t-1},\mathbf{z}^c_{1:t-1},\mathbf{z}^p_{1:t})$. Note that we discard the inference model during test time.

3.2. Adversarial Learning in the Latent Space

To further improve the disentangled ability of the proposed DR-SVG model, we utilize adversarial learning in the latent space to separate pose representation $\mathbf{z}_p$ from content representation $\mathbf{z}_c$. More formally, let $\mathbf{x}^i = (\mathbf{x}^i_1, \mathbf{x}^i_2, ..., \mathbf{x}^i_T)$ denotes a sequence of T frames from video i. Our desired disentanglement should have the same content within a video clip, but distinct between them. Inspired by [2], we impose this via an adversarial framework between the pose-level discriminator D_p and the pose-inference model q_{ϕ_p}, as illustrated in Fig.2. The latter provides pairs of pose vectors, either from the same video clip $(\mathbf{z}^{p,(i)}_1, \mathbf{z}^{p,(i)}_t)$ or from different ones $(\mathbf{z}^{p,(i)}_1, \mathbf{z}^{p,(j)}_t)$. Then D_p tries to output a scalar probability via a cross-entropy loss, which indicates whether pose-vector pair have the same content. With the above discussions, we define the objective function $\mathcal{L}^{adv}_{D_p}$ as follows:

$$-\mathcal{L}^{adv}_{D_p} = \log[D_p(\mathbf{z}^{p,(i)}_1, \mathbf{z}^{p,(i)}_t)] + \log[1 - D_p(\mathbf{z}^{p,(i)}_1, \mathbf{z}^{p,(j)}_t)]. \tag{5}$$

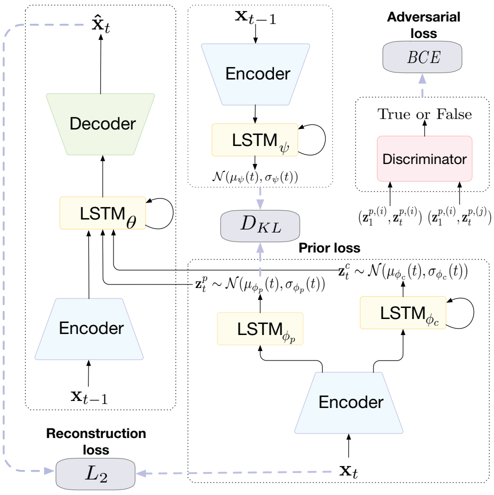
$\hat{\mathbf{x}}_t$
Decoder
LSTM$_\theta$
Encoder
$\mathbf{x}_{t-1}$
$\mathbf{x}_{t-1}$
Encoder
LSTM$_\psi$
$\mathcal{N}(\mu_\psi(t), \sigma_\psi(t))$
D_{KL}
Prior loss
Adversarial loss
BCE
True or False
Discriminator
$(\mathbf{z}^{p,(i)}_1, \mathbf{z}^{p,(i)}_t)$ $(\mathbf{z}^{p,(i)}_1, \mathbf{z}^{p,(j)}_t)$
$\mathbf{z}^c_t \sim \mathcal{N}(\mu_{\phi_c}(t), \sigma_{\phi_c}(t))$
$\mathbf{z}^p_t \sim \mathcal{N}(\mu_{\phi_p}(t), \sigma_{\phi_p}(t))$
LSTM$_{\phi_p}$
LSTM$_{\phi_c}$
Encoder
$\mathbf{x}_t$
Reconstruction loss
L_2

Fig. 2. The architecture of DR-SVG. Note that the model is trained with a variational lower-bound and adversarial loss terms. The whole pipeline is easily trained end-to-end.

The other half of the adversarial framework encourages pose-inference model q_{ϕ_p} to maximize the uncertainty of the discriminator D_p output on pairs of frames from the same clip:

$$-\mathcal{L}^{adv}_{q_{\phi_p}} = \frac{1}{2}\log[D_p(\mathbf{z}^{p,(i)}_1, \mathbf{z}^{p,(i)}_t)] + \frac{1}{2}\log[1 - D_p(\mathbf{z}^{p,(i)}_1, \mathbf{z}^{p,(i)}_t)]. \tag{6}$$

Through this loss term, the pose-inference model is driven to produce the pose pairs that do not contain the content information, confusing the discriminator D_p. Note that this does assume that the object's pose is not distinctive to a particular video sequence.

3.3. Architecture

In this subsection, we describe the architecture of the proposed DR-SVG. As shown in Fig.2, we apply LSTM networks as our p_θ, q_{ϕ_c}, q_{ϕ_p}, and p_ψ. Frames are input to the frame encoder Enc and then passed to the LSTMs, shared across all four parts. Besides, the frame decoder Dec brings the outputs of frame predictor p_θ back to the pixel space. In this context, we focus on how our model reconstructs frame $\hat{\mathbf{x}}_t$ during the training time, where we omit the reconstruction at other time steps for the recurrent nature of LSTMs:

$$\begin{aligned} h_{t-1} &= Enc(\mathbf{x}_{t-1}) \qquad h_t = Enc(\mathbf{x}_t) \\ [\mu_{\phi_c}(t), \sigma_{\phi_c}(t)] &= LSTM_{\phi_c}(h_t) \\ [\mu_{\phi_p}(t), \sigma_{\phi_p}(t)] &= LSTM_{\phi_p}(h_t) \\ \mathbf{z}^t_c &\sim \mathcal{N}(\mu_{\phi_c}(t), \sigma_{\phi_c}(t)) \\ \mathbf{z}^t_p &\sim \mathcal{N}(\mu_{\phi_p}(t), \sigma_{\phi_p}(t)) \\ g_t &= LSTM_\theta(h_{t-1}, \mathbf{z}^t_c, \mathbf{z}^t_p) \\ \mu_\theta(t) &= Dec(g_t). \end{aligned} \tag{7}$$

During training, the frame decoder Dec reconstructs $\hat{\mathbf{x}}_t$ by the concatenation of the deterministic component (encoded feature h_{t-1}) and the stochastic part (latent variables $\mathbf{z}_p^t$ and $\mathbf{z}_c^t$). Besides, since the sequence-independent prior p_ψ is learned across time, and the parameters of this distribution at time step t are generated as follows:

$$[\mu_\psi(t), \sigma_\psi(t)] = LSTM_\psi(h_{t-1}). \tag{8}$$

Here, in order to perform a disentanglement of content and pose in the latent space, we argue that the distribution of content prior should be consistent within a video clip. Therefore, we take the content posterior of the last time step $q_{\phi_c}(\mathbf{z}_{t-1}^c|\mathbf{x}_{1:t-1})$ as the content prior at current time step. Besides, we leverage an adversarial framework between the discriminator D_p and pose-inference model q_{ϕ_p} in the latent space, eliminating the pose information from the clip content.

During test time, the content variables $\mathbf{z}_t^c$ and the pose variables $\mathbf{z}_t^p$ are drawn from the sequence-dependent prior $q_{\phi_c}(\mathbf{z}_{t-1}^c|\mathbf{x}_{1:t-1})$ and the sequence-independent prior $p_\psi(\mathbf{z}_t^p|\mathbf{x}_{1:t-1})$ respectively. The prediction process can be expressed as follows:

$$\begin{aligned} g_t &= LSTM_\theta(h_{t-1}, \mathbf{z}_c^{t-1}, \mathbf{z}_p^t) \\ \mu_\theta(t) &= Dec(g_t). \end{aligned} \tag{9}$$

4. EXPERIMENTS

To evaluate our model, we test it on two datasets, one synthetic video dataset (Moving MNIST [3]) and one real-world video (KTH action [13]). We perform stochastic video prediction on the Moving-MNIST, showing the quantitative and qualitative comparison between our DR-SVG and the competitive baseline model SVG. Moreover, we carry out feature swapping on the KTH action, demonstrating the disentanglement ability of the proposed DR-SVG.

4.1. Model Details

$LSTM_\theta$ is a two-layer LSTMs with 256 cells in each layer while $LSTM_{\phi_c}$, $LSTM_{\phi_p}$ and $LSTM_\psi$ are single-layer LSTMs with 256 cells in each layer. For Moving-MNIST, we adopt the architecture of DCGAN discriminator [14] as our frame encoder *Enc*, where the output dimensionality $|h| = 128$. Similarly, the frame decoder *Dec* applies a DCGAN generator architecture and a sigmoid output layer. The output dimensions of LSTMs are $|g| = 128, |u_{\phi_c}| = 64, |u_{\phi_p}| = 8, |u_\psi| = 8$. For KTH action dataset,the frame encoder has the same architecture with the VGG16 [15] whose output dimensionality $|h| = 128$. The decoder is a mirror structure of the frame predictor, except that spatial up-sampling and a sigmoid output layer take place with the pooling layer. The output dimensions of LSTMs are $|g| = 128, |u_{\phi_c}| = |64, |u_{\phi_p}| = 16, |u_\psi| = 16$. Besides, the pose-level discriminator D_p is a fully connected neural network with 5 hidden layers.

We also train the SVG-LP model as a competitive baseline, since it is superior to the SVG-FP model. For brevity, we abbreviate SVG-LP to SVG in our experiments. Note that SVG has the same encoder, decoder, and LSTMs with our DR-SVG, but with only one set of latent variables. Both models are trained using the ADAM optimizer [16] with a starting learning rate of 0.002. We use β = 1e-4 for Moving-MNIST and β = 1e-6 for KTH dataset. Also, we add skip connections between the frame encoder at the last ground-truth frame and the decoder at t, enabling both networks to maintain and improve their performance even when they become deeper.

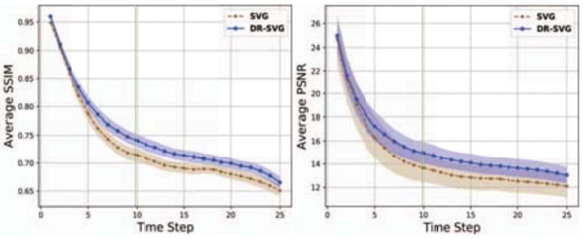

Fig. 3. Quantitative comparison of our DR-SVG and the competitive baseline SVG on the Moving-MNIST dataset. Both models have been sampled 100 times, showing the average SSIM and PSNR. Besides, they are trained up to the frame marked by a green vertical separator, and the part beyond the separator shows the generalization ability of the model. Note that the graph is plotted with 95% confidence interval shaded.

4.2. Moving MNIST

We introduce the Moving MNIST dataset which contains a variety of sequences generated by the method mentioned in [3]. Each frame in the sequences contains two digits bouncing around the 64x64 image. The digits were randomly chosen from the MNIST dataset, and they would bounce back at the edge of the image, which brings the difficulties for frame prediction. Note that we train both models on the Moving-MNIST dataset by conditioning on 5 frames and predicting the next 10 frames.

Quantitative Comparison. In the quantitative evaluation, we use SSIM [17] and PSNR [18] between the generated video sequence and the real sequence as evaluation indicators. For both of SSIM and PSNR, the higher, the better. Empirically, we construct five hundred video sequences with DR-SVG and SVG respectively. For DR-SVG, we calculate SSIM and PSNR by drawing 100 content samples and 100 pose samples from the sequence-dependent prior and the sequence-independent prior separately at each time step for each test sequence. Similarly, in SVG, we sample latent variables $\mathbf{z}$ 100 times to predict future frames. The average SSIM and PSNR on both models are plotted in Fig.3.

Qualitative Comparison. Besides the quantitative comparison, we provide a visual examination of the qualitative results on the Moving-MNIST and in Fig.4. As mentioned before, we sample latent variables 100 times both in both models to perform video prediction, picking the one with the highest SSIM with respect to the ground-truth sequence. We highlight some of the most significant differences in predictions

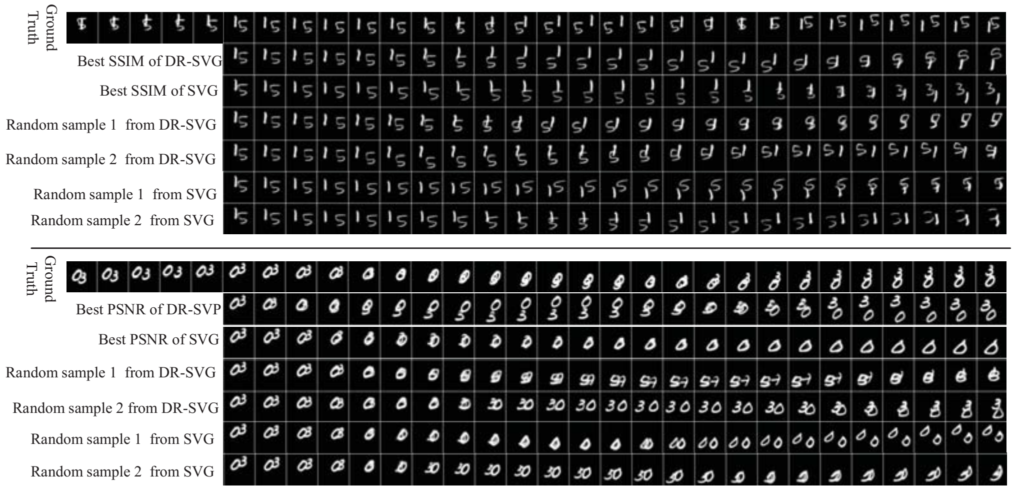

Fig. 4. Qualitative comparison between our DR-SVG model and SVG model on the Moving-MNIST dataset. Both models observe the first 5 frames and then predict the next 25 frames on the test sequences. The first row indicates the ground-truth sequence while the second row and the third row represent prediction results with best SSIM out of 100 random outputs of DR-SVG model and SVG model, respectively. The fourth and fifth rows are random predicted outcomes of the proposed DR-SVG. The last two rows indicate random predicted results of the SVG model. Compared with SVG, DR-SVG can generate frames which shows more stability of content.

by different models. Fig.4 shows the SVG model always produces sequences with changed content (which numbers are presented). From the top example, we can see the number in the best-SSIM prediction of SVG model is changed from '5' to '3'. In the bottom example, the number '3' disappears in the best-SSIM prediction when using SVG model. In contrast, our model can generate video sequences with consistent content.

4.3. KTH action

KTH action dataset includes a total of 2391 video samples of six type of human actions (walking, jogging, running, boxing, hand waving, and hand clamping) completed by 25 people in 4 different scenes. Although the human motion in this dataset is regular, there is still uncertainty in terms of the locations of the persons' joints in a continuous period of time. We use the first 20 people for training and 21-25 for testing. Both models, including DR-SVG and the baseline, are trained on the training set of KTH action dataset by generating subsequent 10 frames from the last 10 observations.

Feature swapping. One might also generate a new video sequence using DR-SVG model with content and pose from different sequences. Given two sequences $\mathbf{x}^i$ and $\mathbf{x}^j$ from KTH action dataset, we first sample content variables $\mathbf{z}_t^{c,(i)} \sim q_{\phi_c}(\mathbf{z}_t^{c,(i)}|\mathbf{x}_{1:t}^{(i)})$ from video i and sample pose variables $\mathbf{z}_t^{p,(j)} \sim q_{\phi_p}(\mathbf{z}_t^{p,(j)}|\mathbf{x}_{1:t}^{(j)})$ from video j, then generate a new squence by $\mathbf{x}_t^{new,(i)} \sim p_\theta(\mathbf{x}_t|\mathbf{x}_{t-1}^{new,(i)}, \mathbf{z}_t^{c,(i)}, \mathbf{z}_t^{p,(j)})$. This allow us th control both the content and pose of the stochastic part in the generated sequence. Similarly, the feature swapping with SVG model can be expressed by $\mathbf{x}_t^{new,(i)} \sim p_\theta(\mathbf{x}_t|\mathbf{x}_{t-1}^{new,(i)}, \mathbf{z}_t^{(j)})$. As detailed in Fig.5, when using SVG mdoel to perform feature swapping, the generated images are always fuzzy. In contrast, when we sample $\mathbf{z}^{p,(j)}$ from the pose posterior $q_{\phi_p}^{(j)}$ in the video j, the generated new sequences are crisp, with a clear content and pose. This indicates the stochastic part in video sequences inherently involves multiple factors while our DR-SVG is able to learn disentangled representations, factorizing content and pose for video sequences.

5. CONCLUSION

In this paper, we presented a disentangled-representation stochastic video generation model (DR-SVG) which factories stochastic part of video sequences into content information and pose information via a variational lower-bound and an adversarial framework. Our model performs video prediction by the concatenation of content variables and pose variables and deterministic features. The proposed DR-SVG compare favorably with the leading model SVG, generating frames with more consistent content. Besides, we perform feature swapping with DR-SVG and SVG model respectively, showing the disentanglement ability of the proposed DR-SVG.

6. ACKNOWLEDGMENT

This work is supported by NSFC project Grant No. U1833101, Shenzhen Science and Technologies project under Grant No. JCYJ20160428182137473 and the Joint Research Center of Tencent and Tsinghua.

7. REFERENCES

[1] Emily Denton and Rob Fergus, "Stochastic video generation with a learned prior," *arXiv preprint arXiv:1802.07687*, 2018.

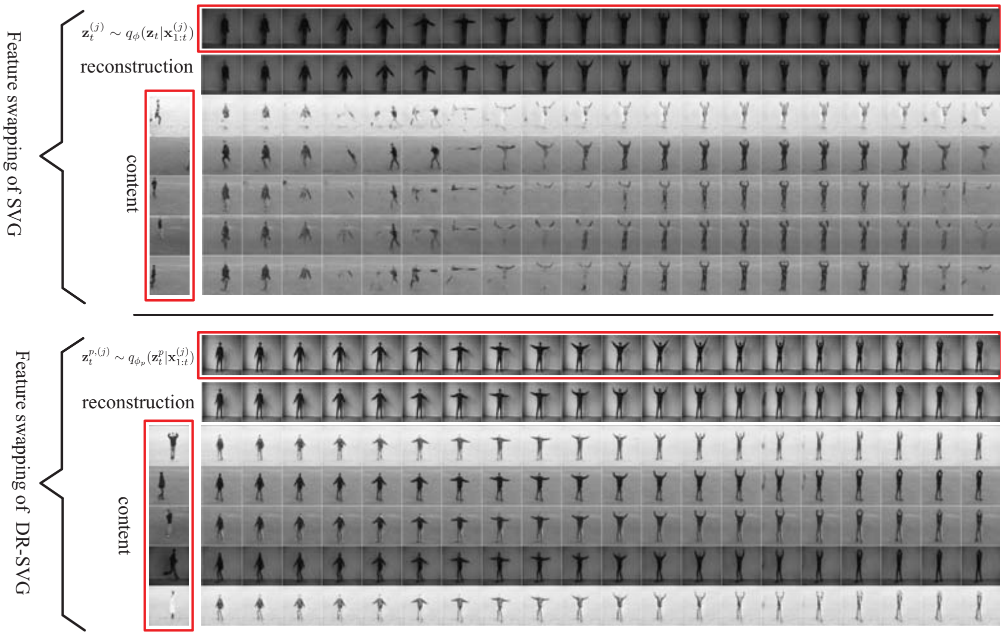
Feature swapping of SVG
$\mathbf{z}_t^{(j)} \sim q_\phi(\mathbf{z}_t|\mathbf{x}_{1:t}^{(j)})$
reconstruction
content
Feature swapping of DR-SVG
$\mathbf{z}_t^{p,(j)} \sim q_{\phi_p}(\mathbf{z}_t^p|\mathbf{x}_{1:t}^{(j)})$
reconstruction
content

Fig. 5. Feature swapping using SVG and DR-SVG model on the KTH action dataset. In DR-SVG, the new sequences are generated by $\mathbf{x}_t^{new,(i)} \sim p_\theta(\mathbf{x}_t|\mathbf{x}_{t-1}^{new,(i)}, \mathbf{z}_t^{c,(i)}, \mathbf{z}_t^{p,(j)})$ while the feature swapping with SVG can be derived from $\mathbf{x}_t^{new,(i)} \sim p_\theta(\mathbf{x}_t|\mathbf{x}_{t-1}^{new,(i)}, \mathbf{z}_t^{(j)})$. Note that the images in the second row of both examples represents the reconstructed images with respect to the ground-truth images from the first row.

[2] Emily L Denton et al., "Unsupervised learning of disentangled representations from video," in *NIPS*, 2017, pp. 4414–4423.

[3] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov, "Unsupervised learning of video representations using lstms," in *ICML*, 2015, pp. 843–852.

[4] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine, "Stochastic variational video prediction," *arXiv preprint arXiv:1710.11252*, 2017.

[5] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *ICLR*, 2017.

[6] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[7] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *NIPS*, 2016, pp. 2172–2180.

[8] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov, "Importance weighted autoencoders," *arXiv preprint arXiv:1509.00519*, 2015.

[9] Kihyuk Sohn, Honglak Lee, and Xinchen Yan, "Learning structured output representation using deep conditional generative models," in *NIPS*, 2015, pp. 3483–3491.

[10] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert, "An uncertain future: Forecasting from static images using variational autoencoders," in *European Conference on Computer Vision*. Springer, 2016, pp. 835–851.

[11] Rui Shu, James Brofos, Frank Zhang, Hung Hai Bui, Mohammad Ghavamzadeh, and Mykel Kochenderfer, "Stochastic video prediction with conditional density estimation," in *ECCV Workshop on Action and Anticipation for Visual Learning*, 2016, vol. 2.

[12] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," 2016.

[13] Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: a local svm approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. IEEE, 2004, vol. 3, pp. 32–36.

[14] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[15] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[17] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[18] Quan Huynh-Thu and Mohammed Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008.