HAF-SVG: Hierarchical Stochastic Video Generation with Aligned Features

Zhihui Lin[1,2], **Chun Yuan**[2,3] and **Maomao Li**[2]

[1]Department of Computer Science and Technologies, Tsinghua University, Beijing, China
[2]Graduate School at Shenzhen, Tsinghua University, Shenzhen, China
[3]Peng Cheng Laboratory, Shenzhen, China
{lin-zh14, mm-li17}@mails.tsinghua.edu.cn, yuanc@sz.tsinghua.edu.cn,

Abstract

Stochastic video generation methods predict diverse videos based on observed frames, where the main challenge lies in modeling the complex future uncertainty and generating realistic frames. Numerous of Recurrent-VAE-based methods have achieved state-of-the-art results. However, on the one hand, the independence assumption of the variables of approximate posterior limits the inference performance [Zhao *et al.*, 2017; Blei *et al.*, 2017]. On the other hand, although these methods adopt skip connections between encoder and decoder to utilize multi-level features, they still produce blurry generation due to the spatial misalignment between encoder and decoder features at different time steps. In this paper, we propose a hierarchical recurrent VAE with a feature aligner, which can not only relax the independence assumption in typical VAE but also use a feature aligner to enable the decoder to obtain the aligned spatial information from the last observed frames. The proposed model is named **H**ierarchical **S**tochastic **V**ideo **G**eneration network with **A**ligned **F**eatures, referred to as HAF-SVG. Experiments on Moving-MNIST, BAIR, and KTH datasets demonstrate that hierarchical structure is helpful for modeling more accurate future uncertainty, and the feature aligner is beneficial to generate realistic frames. Besides, the HAF-SVG exceeds SVG on both prediction accuracy and the quality of generated frames.

1 Introduction

Video generation is a wide research area in computer vision which contains deterministic and stochastic video generation. Deterministic video generation methods learn to generate only one possible future for observed frames[Villegas *et al.*, 2017a; Wang *et al.*, 2017b; Wang *et al.*, 2018a; Wang *et al.*, 2018b]. In contrast, stochastic video generation methods focus on modeling future uncertainty and generating different possible future frames. Recently, a large amount of work has accomplished stochastic video generation within the framework of variational autoencoders (VAEs) [Kingma and Welling, 2014], which predicts diverse future by sampling latent variables and decoding them into multiple frames [Shu *et al.*, 2016; Babaeizadeh *et al.*, 2018]. Afterwards, [Denton and Fergus, 2018] proposed an effective **s**tochastic **v**ideo **g**eneration method SVG, which learns a prior model of uncertainty in a given environment. During training, SVG adopts an inference model to approximate the true posterior distributions and a generative model to produce frames by observing the past frames $x_{1:t-1}$ and sampling latent variables z_t from the approximated posterior at time step t. At test time, by drawing samples from the prior and combining them with a deterministic predictor, SVG can generate varied frames into future. Nevertheless, the existing approaches struggle to generate realistic and high-quality video sequences. For example, when the SVG method is adopted to perform video prediction on the Moving-MNIST dataset [Srivastava *et al.*, 2015], the numbers are always blurry.

We argue that there may be two reasons. Firstly, they follow the assumption in deep VAEs, which advocates all dimensions in variables of approximated posterior are independent of each other. That is, they samples all dimensions of a variable at one time, which hinders the inference process from modeling a more accurate future uncertainty [Zhao *et al.*, 2017; Blei *et al.*, 2017; Hoffman *et al.*, 2013]. Secondly, they use skip connections from the encoder of the last ground-truth frame to the decoder at the current time step t. Although it enables the decoder to copy from the last observed frame directly, it would bring feature misalignment and blurry generation. Here comes to a question that how to build a stochastic generation model which can model future uncertainty more accurately and generate realistic future frames.

In this paper, we investigate how to overcome the above deficiency in existing stochastic video generation methods with a modified recurrent VAE. On the one hand, we relax the independence assumption of approximated posterior in SVG by splitting all dimensions of variable z_t into G groups. The G sub-variables can be represented as z_t^j, where $j = 1, 2, ...G$. In this way, when we sample j-th sub-variable from its corresponding dimensions, we can use the previous $j-1$ sub-variables as the known information. This hierarchical structure leads to more accurate modeling for future uncertainty. On the other hand, we replace the direct skip connections between encoder and decoder at different time steps with a novel feature aligner to deal with the feature misalignment.

First, similarity scores are computed by the point-wise inner product between encoder and decoder features. Then, we normalize these similarity scores via a softmax layer to obtain attention weights. Finally, the representation of each position is computed by a weighted sum of the features at all positions. Thus, the output feature of the feature aligner not only contain aligned spatial context from the observed frames which are encoded by the encoder, but also the decoder feature at the current time step t. These two improvements on SVG motivates the name of our method: **H**ierarchical **S**tochastic **V**ideo **G**eneration network with **A**ligned **F**eatures: HAF-SVG.

Although there are quite a few approaches have been introduced to perform stochastic video generation [Shu *et al.*, 2016; Babaeizadeh *et al.*, 2018; Denton and Fergus, 2018], it is still a challenge to model a more accurate future uncertainty and produce realistic video clips. The aim of this paper is to disclose this feasibility. To demonstrate effectiveness of the proposed HAF-SVG on stochastic video generation, we provide abundant experimental results on Moving-MNSIT [Srivastava *et al.*, 2015], KTH [Schuldt *et al.*, 2004] and BAIR [Ebert *et al.*, 2017] datasets. HAF-SVG outperforms SVG on all datasets. The contribution of this paper is:

- We relax the independence assumption of approximate posterior in SVG and construct hierarchical inference modules by separating latent variable z_t into G inter-dependent groups, which is beneficial to model a more accurate future uncertainty.
- We introduce a novel feature aligner into our modified recurrent VAE to overcome the misalignment between encoder and decoder features at different time steps, which is helpful for generating realistic frames.
- We carry out extensive experiments, and it turns out that HAF-SVG far exceeds the current state-of-the-art method SVG in terms of prediction accuracy and the quality of generated frames with a large margin.

2 Related Work

Due to the latest progress in deep learning, a number of approaches have been proposed to perform video prediction with deterministic models [Denton and others, 2017; Srivastava *et al.*, 2015; Finn *et al.*, 2016; Villegas *et al.*, 2017a; Villegas *et al.*, 2017b; Wang *et al.*, 2017b; Wang *et al.*, 2018a; Wang *et al.*, 2018b]. However, these deterministic models struggle to deal with future uncertainty and always lead to averaging of future states.

Stochastic video Generation. Stochastic models built upon recurrent-VAE are proposed to deal with the inherent uncertainty of future states in videos. SV2P [Babaeizadeh *et al.*, 2018] predicts different possible future for each sample of its latent variables. SVG [Denton and Fergus, 2018] learns a prior distribution at each time step rather than using a standard Gaussian directly. From this learned prior, we can sample the diverse and plausible future sequence at test time. However, we argue that the existing stochastic video generation methods cannot generate realistic and high-quality frames because of their limited inference performance caused by the independence assumption of approximated posterior and the feature misalignment brought by skip connections between the encoder and decoder at different time steps. In contrast, the proposed HAF-SVG is the first attempt to deal with these two problems on the stochastic video prediction task.

Hierarchical Variational Auto-encoders. The original VAEs [Kingma and Welling, 2014] use an approximated posterior $q_\phi(z|x)$ to approach the true posterior, where z is the only latent variable, and its dimension is independent of each other. In contrast, hierarchical VAE (HVAEs) is a series of VAEs stacked on top of each other [Zhao *et al.*, 2017]. It has the following hierarchy latent variables $z = \{z_1, z_2, , , z_G\}$, besides to the observed variables x. By splitting latent variable z as G groups, HVAEs is able to deal with the inactivate stochastic latent variable problem [Tomczak and Welling, 2018] and improve the inference performance.

3 Framework and Formulation

3.1 Preliminaries

Let $x_{1:T}$ represents a video clip with T consecutive frames. Stochastic video generation methods observe first k frames and predict the diverse future sequences with $T-k$ frames recurrently by sampling a latent variable and decoding it at each time step. Here, we briefly review the SVG model [Denton and Fergus, 2018], which achieves the best stochastic video generation quality so far. Analogous to using VAEs to generate images, SVG produce video sequences by optimizing the following Evidence Lower Bound (ELBO):

$$\mathcal{L}^{SVG}(x_{k+1:T}|x_{1:k}) = \sum_{t=k+1}^{T} [\mathbb{E}_{q_\phi(z_t|x_{1:t})} \log p_\theta(x_t|z_t, x_{1:t-1}) - \beta D_{KL}(q_\phi(z_t|x_{1:t})||p_\psi(z_t|x_{1:t-1})], \tag{1}$$

where the first term in the right-hand side (RHS) is the negative prediction loss and the second term is the Kullback-Leibler (KL) divergence between the approximated posterior q_ϕ and the learned prior p_ψ. β is a hyper-parameter that controls the trade off between these two terms. Specifically, $p_\theta(x_t|z_t, x_{1:t-1})$ represents a frame predictor (generator) that constructs frame x_t conditioned on the estimated features of $x_{1:t-1}$ and the sampled latent variable z_t at time step t, as well as the information of $x_{1:t-2}$ stemming from the recurrent nature of the model. $q_\phi(z_t|x_{1:t})$ indicates an inference model (encoder) that is forced to be close to a learned prior distribution p_ψ via the KL term during training, while this encoder is ignored at test time. Instead of adopting $\mathcal{N}(0, \mathbf{I})$ as a prior, SVG learns a prior distribution $p_\psi(z_t|x_{1:t-1})$ that is specified by a conditional Gaussian distribution $\mathcal{N}(\mu_\psi(x_{1:t-1}), \sigma_\psi(x_{1:t-1}))$. At test time, we can draws samples from this learned prior to generate video clips.

Following the original VAEs, SVG treats all dimensions in variable z_t of the approximated posterior $q_\phi(z_t|x_{1:t})$ as independent. Nevertheless, this would hinder inference performance during training. Besides, to obtain dense background information, the decoder in SVG receives multi-level features from the last observed frames x_k with skip connections at the current time step t, which would lead to feature misalignment and blurry generation.

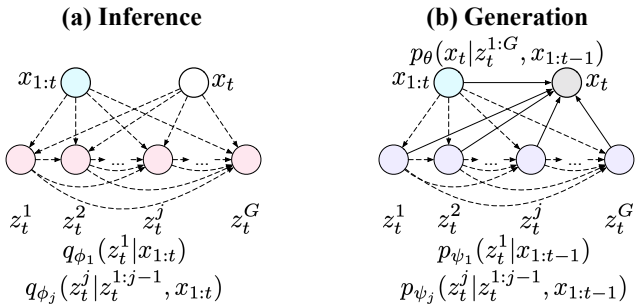

Figure 1: A graphical model of the encoder and the generator in the H-SVG, which are denoted by dashed and solid lines respectively. Left: The inference model. Right: The generative model.

3.2 Hierarchical Stochastic Video Generation

Considering that the dimensions in z_t of the approximated posterior q_ϕ are not necessarily independent one from each other, we relax the independence assumption by splitting all dimensions of variable z_t into G sub-groups as:

$$q_\phi(z_t|x_{1:t}) = q_{\phi_1}(z_t^1|x_{1:t}) \prod_{j=2}^{G} q_{\phi_j}(z_t^j|z_t^{1:j-1}, x_{1:t}). \quad (2)$$

Similarly, the learned prior can be factorized as follows:

$$p_\psi(z_t|x_{1:t-1}) = p_{\psi_1}(z_t^1|x_{1:t-1}) \prod_{j=2}^{G} p_{\psi_j}(z_t^j|z_t^{1:j-1}, x_{1:t-1}), \quad (3)$$

where both $q_{\phi_j}(z_t^j|z_t^{1:j-1}, x_{1:t})$ and $p_{\psi_j}(z_t^j|z_t^{1:j-1}, x_{1:t-1})$ are specified by conditional Gaussian distributions. Each sub-variables z_t^j are treated as independent within groups, but opposite between groups. Besed on this analogy, we propose **H**ierarchical **S**tochastic **V**ideo **G**eneration (H-SVG) which merely relax the independence assumption of approximated posterior and the corresponding learned prior in SVG. As seen in Figure 1, the inference and the generative model of the proposed H-SVG is visualized. It is naturally organized as a hierarchical structure. Each sub-variable z_t^j in H-SVG directly depends on all previous sub-variables $z_t^{1:j-1}$ and $x_{1:t}$. From this perspective, SVG can be regarded as the 1-group H-SVG or H-SVG-1 in short.

Inference model. We design an inference model to approximate the true posterior distribution on G latent sub-variables $z_t^{1:G}$ at time step t. Each approximated sub-posterior distribution $q_{\phi_j}(z_t^j|z_t^{1:j}, x_{1:t})$ on the sub-variable z_t^j is specified by a conditional Gaussian distribution $\mathcal{N}(\mu_{\phi_j}(z_t^{1:j-1}, x_{1:t}), \sigma_{\phi_j}(z_t^{1:j-1}, x_{1:t}))$. During training, we force $q_{\phi_j}(z_t^j|z_t^{1:j-1}, x_{1:t})$ to approach the learned sub-prior $p_{\psi_j}(z_t^j|z_t^{1:j-1}, x_{1:t-1})$ for $j = 1, 2, ..., G$, separately.

Generative model. At time step t, the generative process $p_\theta(x_t|z_t^{1:G}, x_{1:t-1})$ involves $x_{1:t-1}$ and variables $z_t^{1:G}$ where the latter is sampled from the learned prior $p_\psi(z_t|x_{1:t-1})$ via the re-parameterization trick [Kingma and Welling, 2014] at testing time. Note that $p_{\psi_j}(z_t^j|z_t^{1:j-1}, x_{1:t-1})$ only relays on $x_{1:t-1}$ and sampled sub-variables $z_t^{1:j-1}$. The frame predictor p_θ receives $z_t^{1:G}$ and x_{t-1} as input. The dependencies on all previous $x_{1:t-2}$ and $z_{t-1}^{1:G}$ derive from the recurrent nature of our model.

4 Model Architecture

4.1 Pipeline of HAF-SVG

Figure 2 delineates the pipeline of the proposed HAF-SVG. The encoder E is constructed with a deep convolutional network which embeds the frame x_t into a hidden space as h_t. The decoder D consists of an LSTM model and an asymmetrical structured CNN model with the encoder. Here, the decoder D receives previous embedded features $h_{1:t-1}$ and sampled latent variables $z_t^{1:G}$ as input. During training, $z_t^{1:G}$ are sampled from approximated posterior q_ϕ in a hierarchical fashion. Note that HAF-SVG optimizes the variational lower bound in the same form as SVG, however, the dimensions of each variable are no longer independent. That is, HAF-SVG also optimizes reconstruction loss between x_t and $\hat{x}_t$ and KL loss between estimated posterior q_ϕ and learned prior p_ψ.

4.2 Hierarchical Inference Module

We illustrate the hierarchical inference module$_\phi$ in the middle of the Figure 2. The hierarchical inference module$_\psi$ works in a similar fashion. Here, we take the former as an example. Each sub-variable z_t^j is sampled from $\mathcal{N}(\mu_{\phi_j}(z_t^{1:j-1}, x_{1:t}), \sigma_{\phi_j}(z_t^{1:j-1}, x_{1:t}))$ which is estimated by LSTM$_{\phi_j}$. Note that LSTM$_{\phi_j}$ receives the hidden feature h_t and previous sampled sub-variables $z_t^{1:j-1}$, and output the parameters of each conditional Gaussian distribution μ_{ϕ_j} and σ_{ϕ_j}. The whole process can be formulated as follows:

$$\begin{aligned} [\mu_{\phi_1}, \sigma_{\phi_1}] &= LSTM_{\phi_1}(h_t), z_t^1 \sim \mathcal{N}(\mu_{\phi_1}, \sigma_{\phi_1}), \\ [\mu_{\phi_j}, \sigma_{\phi_j}] &= LSTM_{\phi_j}(h_t, z_t^{1:j-1}), \\ z_t^j &\sim \mathcal{N}(\mu_{\phi_j}, \sigma_{\phi_j}),\ j = 2, 3, ..., G. \end{aligned} \quad (4)$$

4.3 Feature Aligner

The top of Figure 3 illustrates the phenomenon of feature misalignment. Green boxes represent the needed receptive field to construct moving objects at the time step t. Red boxes are the corresponding location in the frame x_k which is used to extract multi-level features. However, target objects are missed or partially missed in red boxes, which means the decoder may not be able to utilize the appearance features of target objects in frame x_k to construct realistic frames.

To make full use of multi-level features from the last observed frame x_k without feature misalignment, we propose a feature aligner between the encoder E and the decoder D. As seen in Figure 3, the $\mathcal{F}_t^d \in \mathbb{R}^{C\times N}$ and the $\mathcal{F}_k^e \in \mathbb{R}^{C\times N}$ denote flattened intermediate features extracted by decoder and encoder at different time steps separately, where C is the number of channels, and $N = H \times W$ represents the number of "pixels" of feature maps. First, $\mathcal{F}_t^d$ and $\mathcal{F}_k^e$ are embedded into the same embedding space by the linear transformation as "query" $\mathbf{Q}$ and "key" $\mathbf{K}$, where $\mathbf{Q} = \mathbf{W}_q\mathcal{F}_t^d$

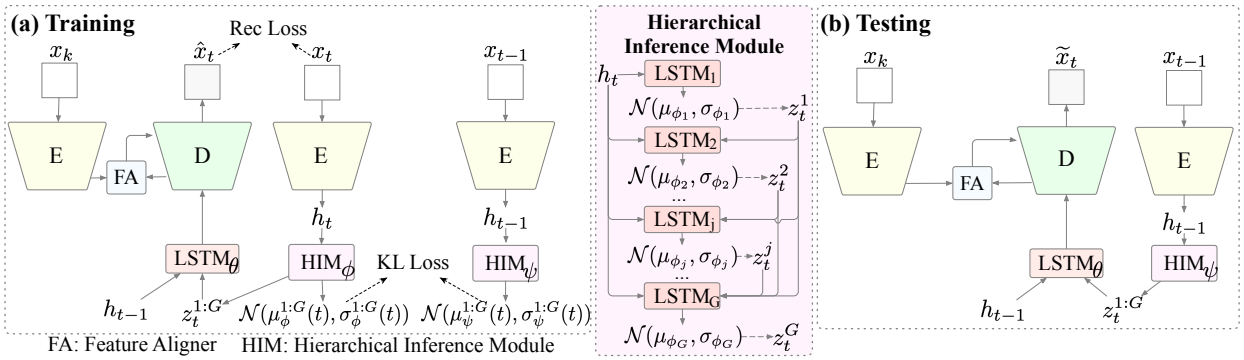

Figure 2: Pipeline of HAF-SVG. During training, the encoder E estimates a latent embedding of image x_t as h_t for each time step. At each time step, the approximated sub-posterior $q_{\phi_j}(z_t^j|z_t^{1:j-1}, x_{1:t})$ and the learned sub-prior $p_{\psi_j}(z_t^j|z_t^{1:j-1}, x_{1:t-1})$ are obtained by the hierarchical inference model with embedded features h_t and h_{t-1} as the input separately, where $j = 1, 2, ..., G$. The information in $h_{1:t}$ and $h_{1:t-2}$ derive from the model recurrence. The decoder D receives h_{t-1} and sampled variables $z_t^{1:G}$ and output reconstruction $\hat{x}_t$ at the training time or the prediction $\widetilde{x}_t$ during the testing phase recurrently. Besides, we introde a feature aligner between the encoder E and decoder D to help decoder obtain aligned context information from the last observed frame x_k.

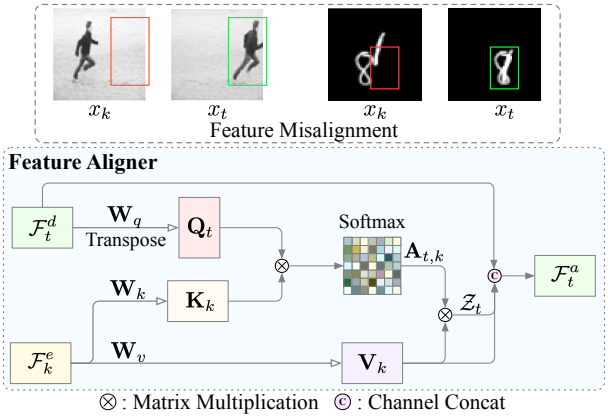

Figure 3: Top: the phenomenon of feature misalignment between features from different time steps (best view in color). Bottom: the architecture of the proposed feature aligner. Features of encoder $\mathcal{F}_k^e$ are aligned to decoder features $\mathcal{F}_t^d$ with an attention approach.

and $\mathbf{K} = \mathbf{W}_k\mathcal{F}_k^e$. The "value" $\mathbf{V}$ can also be obtained by $\mathbf{V} = \mathbf{W}_v\mathcal{F}_t^d$. Then, we compute the similarity score matrix $\mathbf{S} = \mathbf{Q}^T\mathbf{K} \in \mathbb{R}^{N\times N}$ by the point-wise inner product between $\mathbf{Q}$ and $\mathbf{K}$. Next, the attention map $\mathbf{A}$ is calculated via adopting a softmax layer:

$$a_{i,j} = \frac{\exp(\mathbf{s}_{i,j})}{\sum_{l=1}^{N}\exp(\mathbf{s}_{i,l})}. \tag{5}$$

Finally, the aligned feature $\mathcal{Z}_t$ is calculated by the matrix multiplication between $\mathbf{V}$ and $\mathbf{A}$: $\mathcal{Z}_t = \mathbf{V}\mathbf{A}^T \in \mathbb{R}^{C\times N}$. The feature aligner outputs the concatenation of all features as $\mathcal{F}_t^a = [\mathcal{F}_t^d; \mathcal{F}_k^e; \mathcal{Z}_t] \in \mathbb{R}^{3C\times N}$.

The feature aligner is inspired by the self-attention (SA) [Wang *et al.*, 2017a], which has shown its advantages on aggregating context by measuring similarities. Different from the SA which aims to bring global information for each position in a feature map, our feature aligner focus on obtaining aligned spatial context from the last observed frame.

5 Experiments

We perform experiments on synthetic sequences (Moving-MNIST [Srivastava *et al.*, 2015]), as well as real-world videos (KTH action [Schuldt *et al.*, 2004] and BAIR robot [Ebert *et al.*, 2017]). Here, we provide the different settings of HAF-SVG according to different values of G, indicated as HAF-SVG-1, HAF-SVG-2, HAF-SVG-4, where all sub-variables have the same dimensionalities. First, we provide a qualitative comparison between HAF-SVG-1 and the baseline model SVG [Denton and Fergus, 2018], which demonstrates the proposed feature aligner can predict realistic frames with better object content. Besides, we calculate the average Peak Signal to Noise Ratio (PSNR) [Huynh-Thu and Ghanbari, 2008], Structural Similarity Index Measure (SSIM) [Wang *et al.*, 2004] according to the ground-truth sequences to further evaluate SVG and HAF-SVG quantitatively.

5.1 Datasets

Moving-MNIST. This dataset depicts two or three potentially overlapping digits moving with constant velocity and bouncing off the image edges, denoted as Moving-MNIST-2 and Moving-MNIST-3, respectively[Srivastava *et al.*, 2015]. Each training sequence consists of 15 consecutive frames, 5 for the input and 10 for the prediction.

BAIR robot pushing. BAIR robot pushing dataset contains sequences of frames where Sawyer robotic arm pushes various objects on the table [Ebert *et al.*, 2017]. We train both SVG and HAF-SVG on the BAIR dataset by conditioning on 2 frames and predicting the next 10 frames.

KTH action. This dataset includes six types of human actions (walking, jogging, running, boxing, hand waving, and hand clamping) performed by 25 people in 4 different scenes [Schuldt *et al.*, 2004]. During training, we generate the subsequent 10 frames by observing 10 frames.

5.2 Training Details

We adopt the experiment setup in SVG [Denton and Fergus, 2018], where frames are all resized into 64×64. LSTM$_\theta$ is

Figure 4: Qualitative comparison of the two methods on the Moving-MNIST-2. To make a fair comparison and prove the effectiveness of the proposed feature aligner, a 1-group HAF-SVG is used. Given 5 frames, both HAF-SVG-1 and SVG generate the next 25 frames on the test sequences. At each time step, latent variables z_t have been sampled 100 times from the prior $p_\psi(z_t|x_{1:t-1})$, separately. We provide different samples from both methods to reflect the diversity and variability of the future states. All samples show that SVG struggles to maintain the content of digits when overlapping occurs. In contrast, HAF-SVG can reconstruct clear digits "5" and "7" under the overlapping situation.

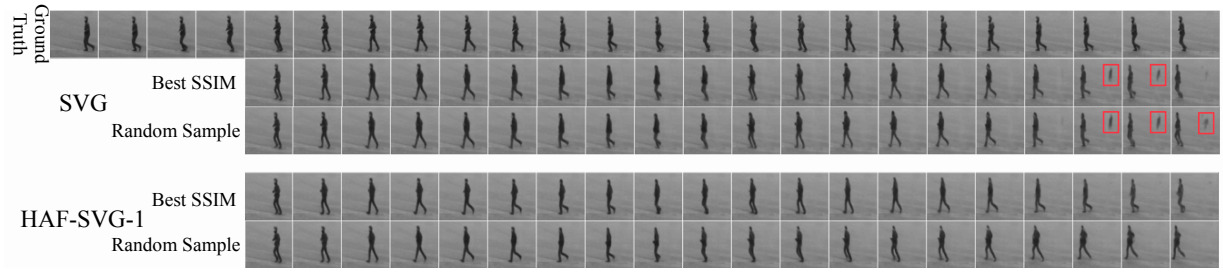

Figure 5: Qualitative comparison between SVG and HAF-SVG-1 on the KTH. Given 10 frames (4 frames are showed in the figure), both HAF-SVG-1 and SVG generate the next 20 frames on the test sequences. We provide different samples from both methods to reflect the diversity and variability of the future states. Compared with SVG, our HAF-SVG-1 generates more clear and crisp frames.

implemented by a two-layer LSTMs with 256 cells in each layer while LSTM_{ϕ_j} and LSTM_{ψ_j} are single-layer LSTMs with 256 cells. The output dimensionalities of the LSTM networks are 128 and $|h_t| = 128$ for all three datasets. For KTH and BAIR, the encoder E adopts the VGG16 [Simonyan and Zisserman, 2015] architecture, and the frame decoder D is the mirrored version of the encoder. $|\mu_{\phi_j}| = |\mu_{\psi_j}|$ are set to 24 on KTH, while 64 on BAIR. For Moving-MNIST, we adopt the DCGAN discriminator architecture [Radford *et al.*, 2016] as our E, the DCGAN generator architecture as D, and $|\mu_{\phi_j}| = |\mu_{\psi_j}| = 16$. Besides, we use β=1e-4 for Moving-MNIST and BAIR and β=1e-6 for KTH.

5.3 Qualitative Comparison

For every test sequence, we use HAF-SVG-1 and SVG to predict 100 different videos separately by drawing 100 samples z_t from the prior at each time step and decoding them into pixel space. Then, we pick the sequences with the best SSIM and the best PSNR with respect to the ground-truth sequence. Figure 4 shows the qualitative results of two models on the Moving-MNIST, from where we can find that SVG model tends to produce chaos content when the two numbers overlap to some extent in all samples. In contrast, HAF-SVG-1 can produce correct numbers even if there is a certain degree of overlap between the two numbers. Besides, we shows the qualitative comparison on the KTH test set in the Figure 5. SVG tends to produce pseudo shadows without the proposed feature aligner, which is marked by red rectangles in the Figure 5. In contrast, HAF-SVG-1 always generate clear and crisp frames and objects during the prediction.

5.4 Quantitative Comparison

We also provide the average SSIM and PSNR as quantitative metrics to evaluate the proposed HAF-SVG. For each test sequence, we draw 100 samples from their corresponding prior at each time step t. Then, we pick those sequences with the best SSIM and the best PSNR with respect to the ground-truth sequence [Denton and Fergus, 2018]. Figure 6 depicts the average SSIM and PSNR scores over the test sequences of Moving-MNSIT-2, Moving-MNSIT-3, KTH, and BAIR using SVG, HAF-SVG-1, HAF-SVG-2, and HAF-SVG-4. Compared with SVG, our HAF-SVG can achieve a higher average PSNR and SSIM on all datasets. The HAF-SVG-4 achieves the best results mostly. However, the performance margins among different models is limited since the number of samples at each time step is big enough to weaken the influence of our hierarchical sampling ways. Next, we would provide an ablation study on the number of samples.

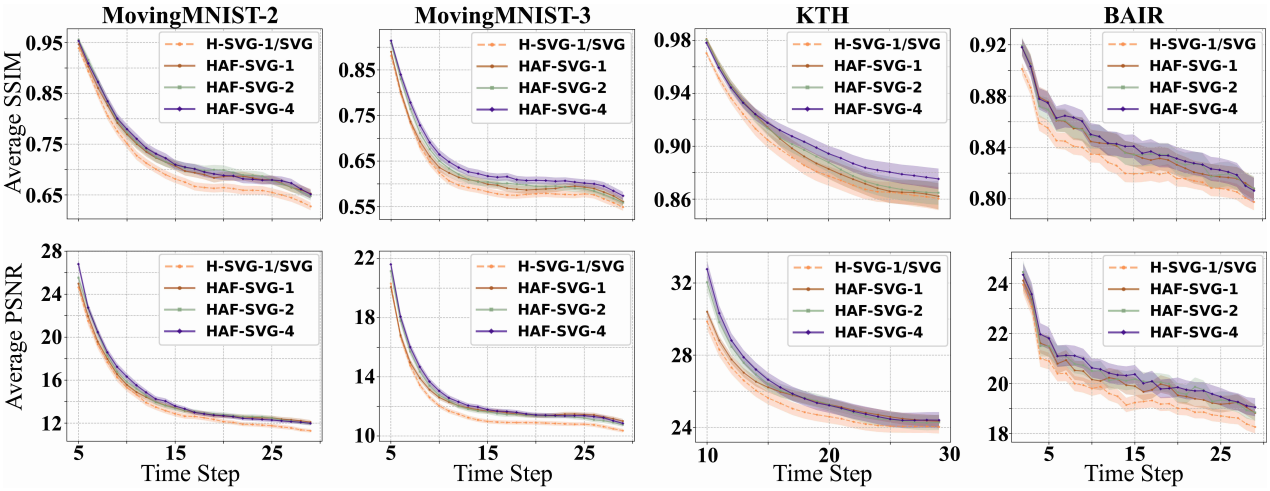

Figure 6: Quantitative comparison between HAF-SVG and SVG on the Moving-MNIST-2, Moving-MNIST-3, KTH, and BAIR datasets with average SSIM and PSNR. Both models are trained to generate consecutive 25 frames based 5 past frames on the Moving-MNIST, 20 frames based on 10 known frames on the KTH, 28 frames with 2 known frames on the BAIR. For each sequence, 100 predictions are sampled and one with the best score with respect to the ground-truth. The plots indicate the average SSIM and PSNR over the unseen test sequences and shadow is the 95% confidence interval. For both average SSIM and PSNR, the higher, the better.

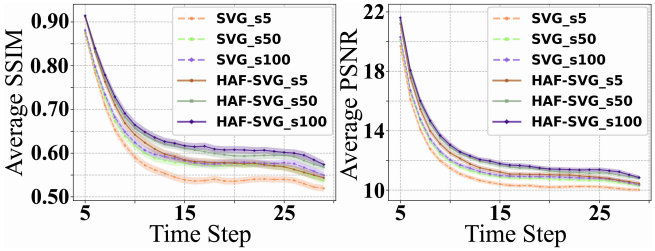

Figure 7: Ablation study on MovingMNIST-3. We use SVG and HAF-SVG-4 to draw 5, 50, and 100 samples separately.

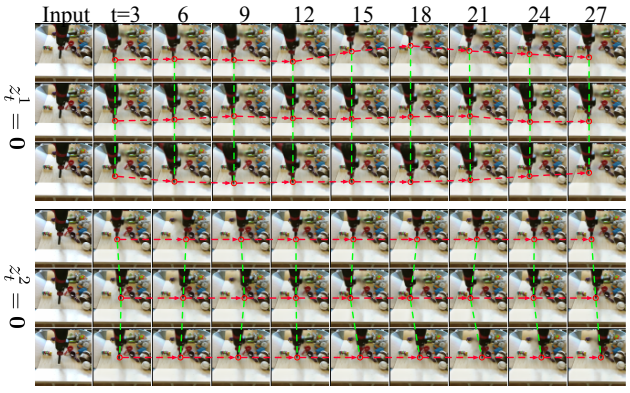

Figure 8: Ablation study on the BAIR with HAF-SVG-2. z_t^1 and z_t^2 are fixed to zero-vectors separately. The visualization clearly shows that z_t^1 and z_t^2 controls the vertical and horizontal movement of the robotic arm separately, which is helpful for uncertainty modeling.

5.5 Ablation Study

The influence of the number of samples. We provide experiments on MovingMNIST-3 using HAF-SVG-4 and SVG with 5, 50, and 100 sampled variables z_t at each step to investigate the influence of the number of samples on the average SSIM and PSNR of models. Figure 7 shows the quantitative results of different models with different sampling numbers. The fewer the sampling number is taken, the more obvious the advantages of our HAF-SVG are.

Is the hierarchy helpful for uncertainty modeling? we further make an ablation study on the BAIR with HAF-SVG-2 to explore whether our methods can learn effective representations via the hierarchical structure. During sampling, we fix the z_t^1 and z_t^2 as zero-vectors separately. The generation results in Figure 8 demonstrates that z_t^1 controls the vertical movement of the robotic arm and z_t^2 controls the movement on the horizontal direction. The hierarchy learns disentangled representations and is beneficial to uncertainty modeling.

6 Conclusion

In this paper, we presented the HAF-SVG model, which adopted a modified recurrent VAE architecture to predict diverse and plausible future frames by sampling variables from a learned prior at each time step. The HAF-SVG relaxed the independence assumption of approximated posterior to improve the inference performance in SVG. Besides, it proposed a novel feature aligner to deal with feature misalignment between encoder and decoder at different time steps. Moreover, we provide extensive experiments on both the synthetic data and the real-world sequences to demonstrate the superiority of the proposed HAF-SVG.

Acknowledgments

This work was supported by NSFC project Grant No. U1833101, Shenzhen Science, Technology and Innovation Commission under Grant No. JCYJ20190809172201639 and the Joint Research Center of Tencent and Tsinghua.

References

[Babaeizadeh *et al.*, 2018] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, RH Campbell, and Sergey Levine. Stochastic variational video prediction. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[Blei *et al.*, 2017] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[Denton and Fergus, 2018] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *International Conference on Machine Learning*, pages 1182–1191, 2018.

[Denton and others, 2017] Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *NIPS*, pages 4414–4423, 2017.

[Ebert *et al.*, 2017] Frederik Ebert, Chelsea Finn, Alex X Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. In *Conference on Robot Learning*, pages 344–356, 2017.

[Finn *et al.*, 2016] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *NIPS*, pages 64–72, 2016.

[Hoffman *et al.*, 2013] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[Huynh-Thu and Ghanbari, 2008] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

[Kingma and Welling, 2014] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *stat*, 1050:10, 2014.

[Radford *et al.*, 2016] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2016.

[Schuldt *et al.*, 2004] 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*, volume 3, pages 32–36. IEEE, 2004.

[Shu *et al.*, 2016] 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*, volume 2, 2016.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3th International Conference on Learning Representations, ICLR 2015*, 2015.

[Srivastava *et al.*, 2015] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *ICML*, pages 843–852, 2015.

[Tomczak and Welling, 2018] Jakub M Tomczak and Max Welling. Vae with a vampprior. In *21st International Conference on Artificial Intelligence and Statistics, AISTATS 2018*, 2018.

[Villegas *et al.*, 2017a] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[Villegas *et al.*, 2017b] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3560–3569. JMLR. org, 2017.

[Wang *et al.*, 2004] 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*, 13(4):600–612, 2004.

[Wang *et al.*, 2017a] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2017.

[Wang *et al.*, 2017b] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S. Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. In *NIPS*, 2017.

[Wang *et al.*, 2018a] Yunbo Wang, Zhifeng Gao, Mingsheng Long, Jianmin Wang, and S Yu Philip. Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In *International Conference on Machine Learning*, pages 5123–5132, 2018.

[Wang *et al.*, 2018b] Yunbo Wang, Jianjin Zhang, Hongyu Zhu, Mingsheng Long, Jianmin Wang, and Philip S. Yu. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9146–9154, 2018.

[Zhao *et al.*, 2017] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Learning hierarchical features from deep generative models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4091–4099. JMLR. org, 2017.