

---

# Stochastic Latent Residual Video Prediction

---

Jean-Yves Franceschi<sup>\*1</sup> Edouard Delasalles<sup>\*1</sup> Mickaël Chen<sup>1</sup> Sylvain Lamprier<sup>1</sup> Patrick Gallinari<sup>1 2</sup>

## Abstract

Designing video prediction models that account for the inherent uncertainty of the future is challenging. Most works in the literature are based on stochastic image-autoregressive recurrent networks, which raises several performance and applicability issues. An alternative is to use fully latent temporal models which untie frame synthesis and temporal dynamics. However, no such model for stochastic video prediction has been proposed in the literature yet, due to design and training difficulties. In this paper, we overcome these difficulties by introducing a novel stochastic temporal model whose dynamics are governed in a latent space by a residual update rule. This first-order scheme is motivated by discretization schemes of differential equations. It naturally models video dynamics as it allows our simpler, more interpretable, latent model to outperform prior state-of-the-art methods on challenging datasets.

## 1. Introduction

Being able to predict the future of a video from a few conditioning frames in a self-supervised manner has many applications in fields such as reinforcement learning (Gregor et al., 2019) or robotics (Babaeizadeh et al., 2018). More generally, it challenges the ability of a model to capture visual and dynamic representations of the world. Video prediction has received a lot of attention from the computer vision community. However, most proposed methods are deterministic, reducing their ability to capture video dynamics, which are intrinsically stochastic (Denton & Fergus, 2018).

Stochastic video prediction is a challenging task which has been tackled by recent works. Most state-of-the-art approaches are based on image-autoregressive models (Denton & Fergus, 2018; Babaeizadeh et al., 2018), built around

Recurrent Neural Networks (RNNs), where each generated frame is fed back to the model to produce the next frame. However, performances of their temporal models innately depend on the capacity of their encoder and decoder, as each generated frame has to be re-encoded in a latent space. Such autoregressive processes induce a high computational cost, and strongly tie the frame synthesis and temporal models, which may hurt the performance of the generation process and limit its applicability (Gregor et al., 2019; Rubanova et al., 2019).

An alternative approach consists in separating the dynamic of the state representations from the generated frames, which are independently decoded from the latent space. In addition to removing the aforementioned link between frame synthesis and temporal dynamics, this is computationally appealing when coupled with a low-dimensional latent space. Moreover, such models can be used to shape a complete representation of the state of a system, e.g. for reinforcement learning applications (Gregor et al., 2019), and are more interpretable than autoregressive models (Rubanova et al., 2019). Yet, these State-Space Models (SSMs) are more difficult to train as they require non-trivial inference schemes (Krishnan et al., 2017) and a careful design of the dynamic model (Karl et al., 2017). This leads most successful SSMs to only be evaluated on small or artificial toy tasks.

In this work, we introduce a novel stochastic dynamic model for the task of video prediction which successfully leverages structural and computational advantages of SSMs that operate on low-dimensional latent spaces. Its dynamic component determines the temporal evolution of the system through residual updates of the latent state, conditioned on learned stochastic variables. This formulation allows us to implement an efficient training strategy and process in an interpretable manner complex high-dimensional data such as videos. This residual principle can be linked to recent advances relating residual networks and Ordinary Differential Equations (ODEs) (Chen et al., 2018). This interpretation opens new perspectives such as generating videos at different frame rates, as demonstrated in our experiments. The proposed approach outperforms current state-of-the-art models on the task of stochastic video prediction, as demonstrated by comparisons with competitive baselines on representative benchmarks.

---

<sup>\*</sup>Equal contribution. <sup>1</sup>Sorbonne Université, CNRS, LIP6, F-75005 Paris, France <sup>2</sup>Criteo AI Lab, Paris, France. Correspondence to: Jean-Yves Franceschi <jean-yves.franceschi@lip6.fr>, Edouard Delasalles <edouard.delasalles@lip6.fr>.

## 2. Related Work

Video synthesis covers a range of different tasks, such as video-to-video translation (Wang et al., 2018), super-resolution (Caballero et al., 2017), interpolation between distant frames (Jiang et al., 2018), generation (Tulyakov et al., 2018), and video prediction, which is the focus of this paper.

**Deterministic models.** Inspired by prior sequence generation models using RNNs (Graves, 2013), a number of video prediction methods (Srivastava et al., 2015; Villegas et al., 2017; van Steenkiste et al., 2018; Wichers et al., 2018; Jin et al., 2020) rely on LSTMs (Long Short-Term Memory networks, Hochreiter & Schmidhuber, 1997), or, like Ranzato et al. (2014), Jia et al. (2016) and Xu et al. (2018a), on derived networks such as ConvLSTMs (Shi et al., 2015). Indeed, computer vision approaches are usually tailored to high-dimensional video sequences and propose domain-specific techniques such as pixel-level transformations and optical flow (Shi et al., 2015; Walker et al., 2015; Finn et al., 2016; Jia et al., 2016; Walker et al., 2016; Vondrick & Torralba, 2017; Liang et al., 2017; Liu et al., 2017; Lotter et al., 2017; Lu et al., 2017a; Fan et al., 2019; Gao et al., 2019) that help to produce high-quality predictions. Such predictions are, however, deterministic, thus hurting their performance as they fail to generate sharp long-term video frames (Babaeizadeh et al., 2018; Denton & Fergus, 2018). Following Mathieu et al. (2016), some works proposed to use adversarial losses (Goodfellow et al., 2014) on the model predictions to sharpen the generated frames (Vondrick & Torralba, 2017; Liang et al., 2017; Lu et al., 2017a; Xu et al., 2018b; Wu et al., 2020). Nonetheless, adversarial losses are notoriously hard to train (Goodfellow, 2016), and lead to mode collapse, thereby preventing diversity of generations.

**Stochastic and image-autoregressive models.** Some approaches rely on exact likelihood maximization, using pixel-level autoregressive generation (van den Oord et al., 2016; Kalchbrenner et al., 2017; Weissenborn et al., 2020) or normalizing flows through invertible transformations between the observation space and a latent space (Kingma & Dhariwal, 2018; Kumar et al., 2020). However, they require careful design of complex temporal generation schemes manipulating high-dimensional data, thus inducing a prohibitive temporal generation cost. More efficient continuous models rely on Variational Auto-Encoders (VAEs, Kingma & Welling, 2014; Rezende et al., 2014) for the inference of low-dimensional latent state variables. Except Xue et al. (2016) and Liu et al. (2019) who learn a one-frame-ahead VAE, they model sequence stochasticity by incorporating a random latent variable per frame into a deterministic RNN-based image-autoregressive model. Babaeizadeh et al. (2018) integrate stochastic variables into the ConvLSTM

architecture of Finn et al. (2016). Concurrently with He et al. (2018), Denton & Fergus (2018) use a prior LSTM conditioned on previously generated frames in order to sample random variables that are fed to a predictor LSTM; performance of such methods were improved in follow-up works by increasing networks capacities (Castrejon et al., 2019; Villegas et al., 2019). Finally, Lee et al. (2018) combine the ConvLSTM architecture and this learned prior, adding an adversarial loss on the predicted videos to sharpen them at the cost of a diversity drop. Yet, all these methods are image-autoregressive, as they feed their predictions back into the latent space, thereby tying the frame synthesis and temporal models and increasing their computational cost. Concurrently to our work, Minderer et al. (2019) propose to use the autoregressive VRNN model (Chung et al., 2015) on learned image key-points instead of raw frames. It remains unclear to which extent this change could mitigate the aforementioned problems. We instead tackle these issues by focusing on video dynamics, and propose a model that is state-space and acts on a small latent space. This approach yields better experimental results despite weaker video-specific priors.

**State-space models.** Many latent state-space models have been proposed for sequence modelization (Bayer & Osendorfer, 2014; Fraccaro et al., 2016; 2017; Krishnan et al., 2017; Karl et al., 2017; Hafner et al., 2019), usually trained by deep variational inference. These methods, which use locally linear or RNN-based dynamics, are designed for low-dimensional data, as learning such models on complex data is challenging, or focus on control or planning tasks. In contrast, our fully latent method is the first one to be successfully applied to complex high-dimensional data such as videos, thanks to a temporal model based on residual updates of its latent state. It falls within the scope of a recent trend linking differential equations with neural networks (Lu et al., 2017b; Long et al., 2018), leading to the integration of ODEs, that are seen as continuous residual networks (He et al., 2016), in neural network architectures (Chen et al., 2018). However, the latter work as well as follow-ups and related works (Rubanova et al., 2019; Yıldız et al., 2019; Le Guen & Thome, 2020) are either limited to low-dimensional data, prone to overfitting or unable to handle stochasticity within a sequence. Another line of works considers stochastic differential equations with neural networks (Ryder et al., 2018; De Brouwer et al., 2019), but are limited to continuous Brownian noise, whereas video prediction additionally requires to model punctual stochastic events.

## 3. Model

We consider the task of stochastic video prediction, consisting in approximating, given a number of conditioning video

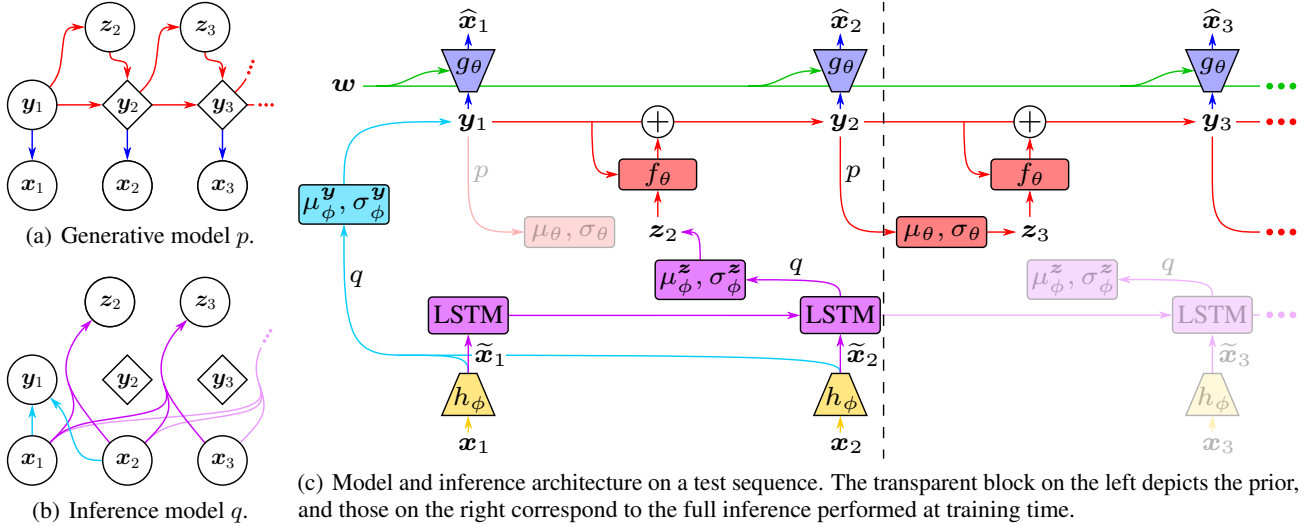


Figure 1. (a), (b) Proposed generative and inference models. Diamonds and circles represent, respectively, deterministic and stochastic states. (c) Corresponding architecture with two parts: inference on conditioning frames on the left, generation for extrapolation on the right.  $h_\phi$  and  $g_\theta$  are deep Convolutional Neural Networks (CNNs), and other named networks are Multilayer Perceptrons (MLPs).

frames, the distribution of possible future frames.

### 3.1. Latent Residual Dynamic Model

Let  $\mathbf{x}_{1:T}$  be a sequence of  $T$  video frames. We model their evolution by introducing latent variables  $\mathbf{y}$  that are driven by a dynamic temporal model. Each frame  $x_t$  is then generated from the corresponding latent state  $\mathbf{y}_t$  only, making the dynamics independent from the previously generated frames.

We propose to model the transition function of the latent dynamic of  $\mathbf{y}$  with a stochastic residual network. State  $\mathbf{y}_{t+1}$  is chosen to deterministically depend on the previous state  $\mathbf{y}_t$ , conditionally to an auxiliary random variable  $z_{t+1}$ . These auxiliary variables encapsulate the randomness of the video dynamics. They have a learned factorized Gaussian prior that depends on the previous state only. The model is depicted in Figure 1(a), and defined as follows:

$$\begin{cases} \mathbf{y}_1 \sim \mathcal{N}(\mathbf{0}, I), \\ z_{t+1} \sim \mathcal{N}(\mu_\theta(\mathbf{y}_t), \sigma_\theta(\mathbf{y}_t)I), \\ \mathbf{y}_{t+1} = \mathbf{y}_t + f_\theta(\mathbf{y}_t, z_{t+1}), \\ \mathbf{x}_t \sim \mathcal{G}(g_\theta(\mathbf{y}_t)), \end{cases} \quad (1)$$

where  $\mu_\theta$ ,  $\sigma_\theta$ ,  $f_\theta$  and  $g_\theta$  are neural networks, and  $\mathcal{G}(g_\theta(\mathbf{y}_t))$  is a probability distribution parameterized by  $g_\theta(\mathbf{y}_t)$ . In our experiments,  $\mathcal{G}$  is a normal distribution with mean  $g_\theta(\mathbf{y}_t)$  and constant diagonal variance. Note that  $\mathbf{y}_1$  is assumed to have a standard Gaussian prior, and, in our VAE setting, will be inferred from conditioning frames for the prediction task, as shown in Section 3.3.

The residual update rule takes inspiration in the Euler dis-

cretization scheme of differential equations. The state of the system  $\mathbf{y}_t$  is updated by its first-order movement, i.e., the residual  $f_\theta(\mathbf{y}_t, z_{t+1})$ . Compared to a regular RNN, this simple principle makes our temporal model lighter and more interpretable. Equation (1), however, differs from a discretized ODE because of the introduction of the stochastic discrete-time variables  $z$ . Nonetheless, we propose to allow the Euler step size  $\Delta t$  to be smaller than 1, as a way to make the temporal model closer to a continuous dynamics. The updated dynamics becomes, with  $\frac{1}{\Delta t} \in \mathbb{N}$  to synchronize the step size with the video frame rate:

$$\mathbf{y}_{t+\Delta t} = \mathbf{y}_t + \Delta t \cdot f_\theta(\mathbf{y}_t, z_{\lfloor t \rfloor + 1}). \quad (2)$$

For this formulation, the auxiliary variable  $z_t$  is kept constant between two integer time steps. Note that a different  $\Delta t$  can be used during training or testing. This allows our model to generate videos at an arbitrary frame rate since each intermediate latent state can be decoded in the observation space. This ability enables us to observe the quality of the learned dynamic as well as challenge its ODE inspiration by testing its generalization to the continuous limit in Section 4. In the following, we consider  $\Delta t$  as a hyperparameter. For the sake of clarity, we consider that  $\Delta t = 1$  in the remaining of this section; generalizing to a smaller  $\Delta t$  is straightforward as Figure 1(a) remains unchanged.

### 3.2. Content Variable

Some components of video sequences can be static, such as the background or shapes of moving objects. They may not impact the dynamics; we therefore model them separately, in the same spirit as Denton & Birodkar (2017) and Yingzhen & Mandt (2018). We compute a content variable  $w$  that

remains constant throughout the whole generation process and is fed together with  $\mathbf{y}_t$  into the frame generator. It enables the dynamical part of the model to focus only on movement, hence being lighter and more stable. Moreover, it allows us to leverage architectural advances in neural networks, such as skip connections (Ronneberger et al., 2015), to produce more realistic frames.

This content variable is a deterministic function  $c_\psi$  of a fixed number  $k < T$  of frames  $\mathbf{x}_c^{(k)} = (\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_k})$ :

$$\begin{cases} \mathbf{w} = c_\psi(\mathbf{x}_c^{(k)}) = c_\psi(\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_k}) \\ \mathbf{x}_t \sim \mathcal{G}(g_\theta(\mathbf{y}_t, \mathbf{w})) \end{cases} \quad (3)$$

During testing,  $\mathbf{x}_c^{(k)}$  are the last  $k$  conditioning frames (usually between 2 and 5).

This content variable is not endowed with any probabilistic prior, contrary to the dynamic variables  $\mathbf{y}$  and  $\mathbf{z}$ . Thus, the information it contains is not constrained in the loss function (see Section 3.3), but only architecturally. To prevent temporal information from leaking in  $\mathbf{w}$ , we propose to uniformly sample these  $k$  frames within  $\mathbf{x}_{1:T}$  during training. We also design  $c_\psi$  as a permutation-invariant function (Zaheer et al., 2017), consisting in an MLP fed with the sum of individual frame representations, following Santoro et al. (2017).

This absence of prior and its architectural constraint allows  $\mathbf{w}$  to contain as much non-temporal information as possible, while preventing it from containing dynamic information. On the other hand, due to their strong standard Gaussian priors,  $\mathbf{y}$  and  $\mathbf{z}$  are encouraged to discard unnecessary information. Therefore,  $\mathbf{y}$  and  $\mathbf{z}$  should only contain temporal information that could not be captured by  $\mathbf{w}$ .

Note that this content variable can be removed from our model, yielding a more classical deep state-space model. An experiment in this setting is presented in Appendix E.

### 3.3. Variational Inference and Architecture

Following the generative process depicted in Figure 1(a), the conditional joint probability of the full model, given a content variable  $\mathbf{w}$ , can be written as:

$$\begin{aligned} p(\mathbf{x}_{1:T}, \mathbf{z}_{2:T}, \mathbf{y}_{1:T} \mid \mathbf{w}) \\ = p(\mathbf{y}_1) \prod_{t=2}^T p(\mathbf{z}_t, \mathbf{y}_t \mid \mathbf{y}_{t-1}) \prod_{t=1}^T p(\mathbf{x}_t \mid \mathbf{y}_t, \mathbf{w}), \end{aligned} \quad (4)$$

with

$$p(\mathbf{z}_t, \mathbf{y}_t \mid \mathbf{y}_{t-1}) = p(\mathbf{z}_t \mid \mathbf{y}_{t-1})p(\mathbf{y}_t \mid \mathbf{y}_{t-1}, \mathbf{z}_t). \quad (5)$$

According to the expression of  $\mathbf{y}_{t+1}$  in Equation (1),  $p(\mathbf{y}_t \mid \mathbf{y}_{t-1}, \mathbf{z}_t) = \delta(\mathbf{y}_t - \mathbf{y}_{t-1} - f_\theta(\mathbf{y}_{t-1}, \mathbf{z}_t))$ , where  $\delta$  is the Dirac delta function centered on  $\mathbf{0}$ . Hence, in order to

optimize the likelihood of the observed videos  $p(\mathbf{x}_{1:T} \mid \mathbf{w})$ , we need to infer latent variables  $\mathbf{y}_1$  and  $\mathbf{z}_{2:T}$ . This is done by deep variational inference using the inference model parameterized by  $\phi$  and shown in Figure 1(b), which comes down to considering a variational distribution  $q_{Z,Y}$  defined and factorized as follows:

$$\begin{aligned} q_{Z,Y} &\triangleq q(\mathbf{z}_{2:T}, \mathbf{y}_{1:T} \mid \mathbf{x}_{1:T}, \mathbf{w}) \\ &= q(\mathbf{y}_1 \mid \mathbf{x}_{1:k}) \prod_{t=2}^T q(\mathbf{z}_t \mid \mathbf{x}_{1:t}) \underbrace{q(\mathbf{y}_t \mid \mathbf{y}_{t-1}, \mathbf{z}_t)}_{=p(\mathbf{y}_t \mid \mathbf{y}_{t-1}, \mathbf{z}_t)}, \end{aligned} \quad (6)$$

with  $q(\mathbf{y}_t \mid \mathbf{y}_{t-1}, \mathbf{z}_t) = p(\mathbf{y}_t \mid \mathbf{y}_{t-1}, \mathbf{z}_t)$  being the aforementioned Dirac delta function. This yields the following evidence lower bound (ELBO), whose full derivation is given in Appendix A:

$$\begin{aligned} \log p(\mathbf{x}_{1:T} \mid \mathbf{w}) &\geq \mathcal{L}(\mathbf{x}_{1:T}; \mathbf{w}, \theta, \phi) \\ &\triangleq -D_{\text{KL}}(q(\mathbf{y}_1 \mid \mathbf{x}_{1:k}) \parallel p(\mathbf{y}_1)) \\ &\quad + \mathbb{E}_{(\tilde{\mathbf{z}}_{2:T}, \tilde{\mathbf{y}}_{1:T}) \sim q_{Z,Y}} \left[ \sum_{t=1}^T \log p(\mathbf{x}_t \mid \tilde{\mathbf{y}}_t, \mathbf{w}) \right. \\ &\quad \left. - \sum_{t=2}^T D_{\text{KL}}(q(\mathbf{z}_t \mid \mathbf{x}_{1:t}) \parallel p(\mathbf{z}_t \mid \tilde{\mathbf{y}}_{t-1})) \right], \end{aligned} \quad (7)$$

where  $D_{\text{KL}}$  denotes the Kullback–Leibler (KL) divergence (Kullback & Leibler, 1951).

The sum of KL divergence expectations implies to consider the full past sequence of inferred states for each time step, due to the dependence on conditionally deterministic variables  $\mathbf{y}_{2:T}$ . However, optimizing  $\mathcal{L}(\mathbf{x}_{1:T}; \mathbf{w}, \theta, \phi)$  with respect to model parameters  $\theta$  and variational parameters  $\phi$  can be done efficiently by sampling a single full sequence of states from  $q_{Z,Y}$  per example, and computing gradients by backpropagation (Rumelhart et al., 1988) through all inferred variables, using the reparameterization trick (Kingma & Welling, 2014; Rezende et al., 2014). We classically choose  $q(\mathbf{y}_1 \mid \mathbf{x}_{1:k})$  and  $q(\mathbf{z}_t \mid \mathbf{x}_{1:t})$  to be factorized Gaussian so that all KL divergences can be computed analytically.

We include an  $\ell_2$  regularization term on residuals  $f_\theta$  applied to  $\mathbf{y}$  which stabilizes the temporal dynamics of the residual network, as noted by Behrmann et al. (2019), de Bézenac et al. (2019) and Rousseau et al. (2019). Given a set of videos  $\mathcal{X}$ , the full optimization problem, where  $\mathcal{L}$  is defined as in Equation (7), is then given as:

$$\begin{aligned} \arg \max_{\theta, \phi, \psi} \sum_{\mathbf{x} \in \mathcal{X}} &\left[ \mathbb{E}_{\mathbf{x}_c^{(k)}} \mathcal{L}(\mathbf{x}_{1:T}; c_\psi(\mathbf{x}_c^{(k)}), \theta, \phi) \right. \\ &\left. - \lambda \cdot \mathbb{E}_{(\mathbf{z}_{2:T}, \mathbf{y}_{1:T}) \sim q_{Z,Y}} \sum_{t=2}^T \|f_\theta(\mathbf{y}_{t-1}, \mathbf{z}_t)\|_2 \right]. \end{aligned} \quad (8)$$

### Stochastic Latent Residual Video Prediction

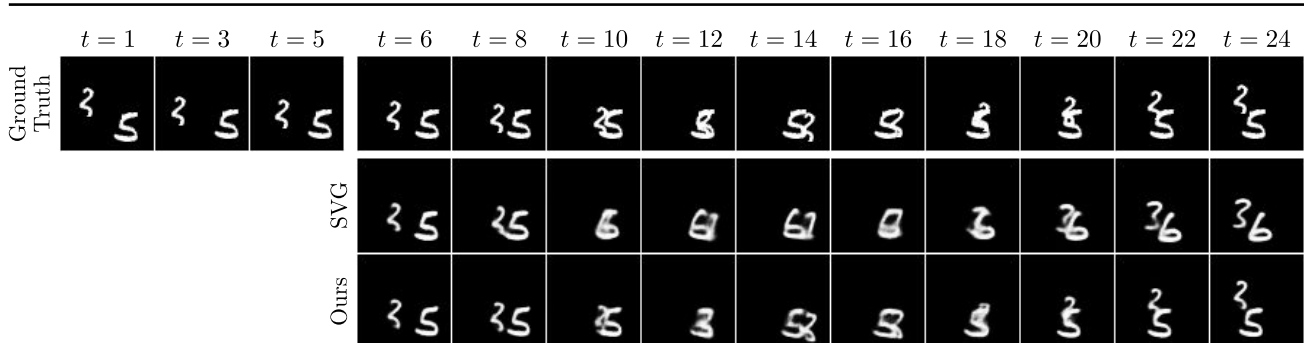


Figure 2. Conditioning frames and corresponding ground truth and best samples with respect to PSNR from SVG and our method for an example of the Stochastic Moving MNIST dataset.

Figure 1(c) depicts the full architecture of our temporal model, showing how the model is applied during testing. The first latent variables are inferred with the conditioning framed and are then predicted with the dynamic model. In contrast, during training, each frame of the input sequence is considered for inference, which is done as follows. Firstly, each frame  $x_t$  is independently encoded into a vector-valued representation  $\tilde{x}_t$ , with  $\tilde{x}_t = h_\phi(x_t)$ .  $y_1$  is then inferred using an MLP on the first  $k$  encoded frames  $\tilde{x}_{1:k}$ . Each  $z_t$  is inferred in a feed-forward fashion with an LSTM on the encoded frames. Inferring  $z$  this way experimentally performs better than, e.g., inferring them from the whole sequence  $x_{1:T}$ ; we hypothesize that this follows from the fact that this filtering scheme is closer to the prediction setting, where the future is not available.

## 4. Experiments

This section exposes the experimental results of our method on four standard stochastic video prediction datasets.<sup>1</sup> We compare our method with state-of-the-art baselines on stochastic video prediction. Furthermore, we qualitatively study the dynamics and latent space learned by our model.<sup>2</sup> Training details are described in Appendix C.

The stochastic nature and novelty of the task of stochastic video prediction make it challenging to evaluate (Lee et al., 2018): since videos and models are stochastic, comparing the ground truth and a predicted video is not adequate. We thus adopt the common approach (Denton & Fergus, 2018; Lee et al., 2018) consisting in, for each test sequence, sampling from the tested model a given number (here, 100) of possible futures and reporting the best performing sample against the true video. We report this discrepancy for three commonly used metrics that are computed frame-wise and averaged over time: Peak Signal-to-Noise Ratio (PSNR,

higher is better), Structured Similarity (SSIM, higher is better), and Learned Perceptual Image Patch Similarity (LPIPS, lower is better, Zhang et al., 2018). PSNR greatly penalizes errors in predicted dynamics, as it is a pixel-level measure derived from the  $\ell_2$  distance, but might also favor blurry predictions. SSIM (only reported in Appendix D for the sake of concision) rather compares local frame patches to circumvent this issue, but loses some dynamics information. LPIPS compares images through a learned distance between activations of deep CNNs trained on image classification tasks, and has been shown to better correlate with human judgment on real images. Finally, the recently proposed Fréchet Video Distance (FVD, lower is better, Unterthiner et al., 2018) aims at directly comparing the distribution of predicted videos with the ground truth distribution through the representations computed by a deep CNN trained on action recognition tasks. It has been shown, independently from LPIPS, to better capture the realism of predicted videos than PSNR and SSIM. We treat all four metrics as complementary, as they capture different scales and modalities.

We present experimental results on a simulated dataset and three real-world datasets, that we briefly present in the following and detail in Appendix B. The corresponding numerical results can be found in Appendix D. For the sake of concision, we only display a handful of qualitative samples in this section, and refer to Appendix H and our website for additional samples. We compare our model against several variational state-of-the-art models: SV2P (Babaeizadeh et al., 2018), SVG (Denton & Fergus, 2018), SAVP (Lee et al., 2018), and StructVRNN (Minderer et al., 2019). Note that SVG has the closest training and architecture to ours among the state of the art. Therefore, we use the same neural architecture as SVG for our encoders and decoders in order to perform fair comparisons with this method.

All baseline results are presented only on the datasets on which they were tested in the original articles. They were obtained with pretrained models released by the authors, except those of SVG on the Moving MNIST dataset and StructVRNN on the Human3.6M dataset, for which we

<sup>1</sup>Code and datasets are available at <https://github.com/edouardelasalles/srvp>. Pretrained models are downloadable at <https://data.lip6.fr/srvp/>.

<sup>2</sup>Animated video samples are available at <https://sites.google.com/view/srvp/>.

## Stochastic Latent Residual Video Prediction

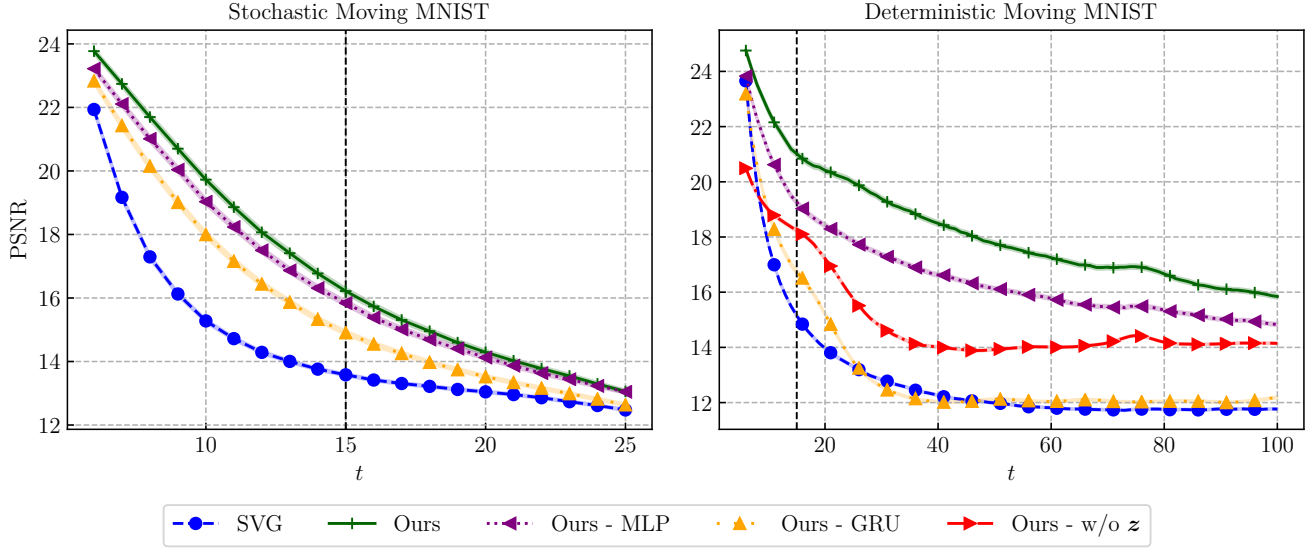


Figure 3. Mean PSNR scores with respect to  $t$  for all tested models on the Moving MNIST dataset, with their 95%-confidence intervals. Vertical bars mark the length of training sequences.

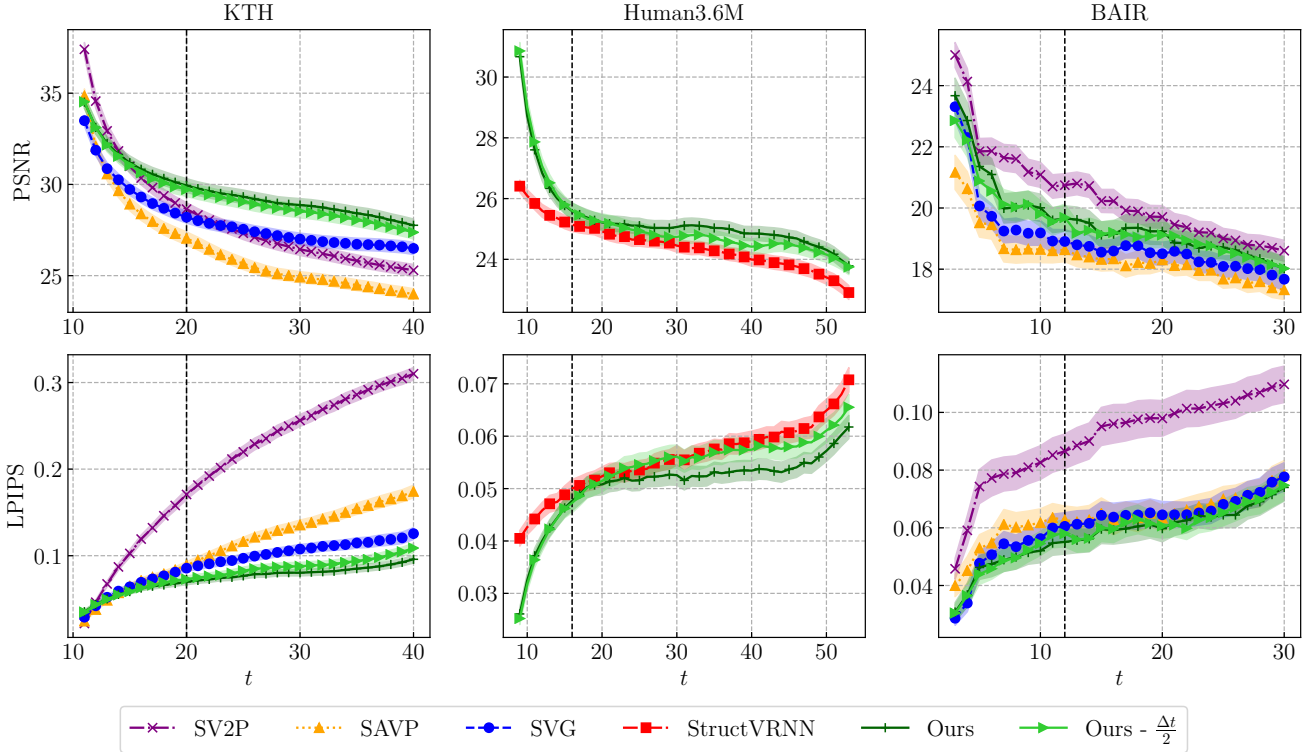


Figure 4. PSNR and LPIPS scores with respect to  $t$  for all tested models on the KTH (left column), Human3.6M (center) and BAIR (right) datasets, with their 95%-confidence intervals. Vertical bars mark the length of training sequences.

Table 1. FVD scores for all tested methods on the KTH, Human3.6M and BAIR datasets with their 95%-confidence intervals over five different samples from the models. Bold scores indicate the best performing method for each dataset.

Dataset	SV2P	SAVP	SVG	StructVRNN	Ours	Ours - $\frac{\Delta t}{2}$	Ours - MLP	Ours - GRU
KTH	636 $\pm$ 1	374 $\pm$ 3	377 $\pm$ 6	—	<b>222 <math>\pm</math> 3</b>	244 $\pm$ 3	255 $\pm$ 4	240 $\pm$ 5
Human3.6M	—	—	—	556 $\pm$ 9	<b>416 <math>\pm</math> 5</b>	<b>415 <math>\pm</math> 3</b>	582 $\pm$ 4	1050 $\pm$ 20
BAIR	965 $\pm$ 17	<b>152 <math>\pm</math> 9</b>	255 $\pm$ 4	—	163 $\pm$ 4	222 $\pm$ 42	162 $\pm$ 4	178 $\pm$ 10

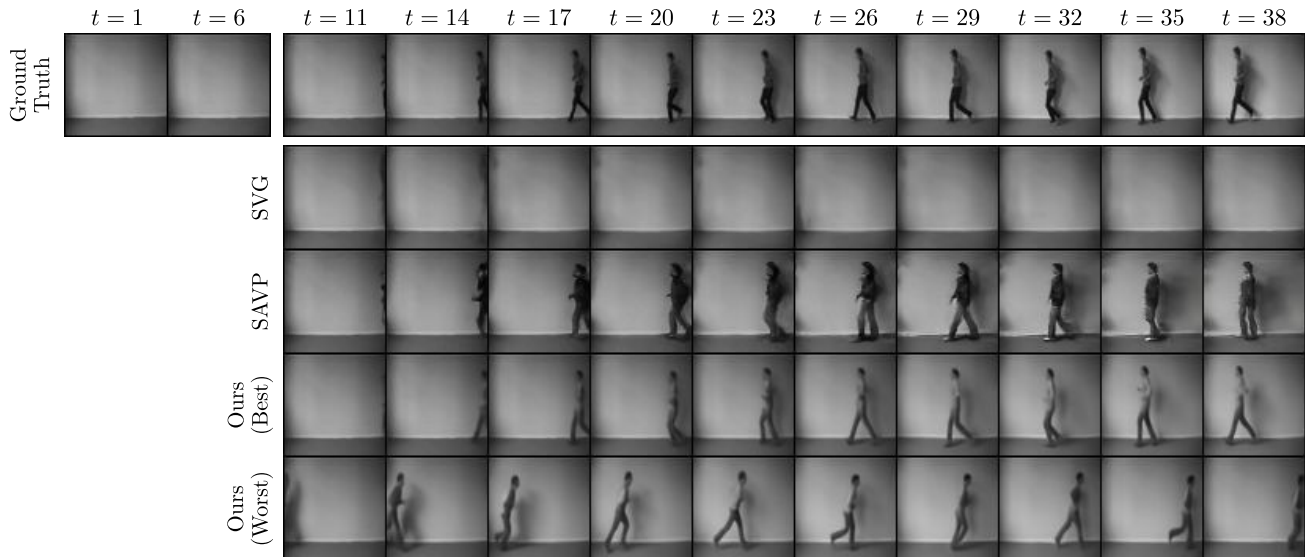


Figure 5. Conditioning frames and corresponding ground truth, best samples from SVG, SAVP and our method, and worst sample from our method, for a video of the KTH dataset. Samples are chosen according to their LPIPS with respect to the ground truth. SVG fails to make a person appear, unlike SAVP and our model. The latter better predicts the subject pose and produces more realistic predictions.

trained models using the code and hyperparameters provided by the authors (see Appendix B). Unless specified otherwise, our model is tested with the same  $\Delta t$  as in training (see Equation (2)).

**Stochastic Moving MNIST.** This dataset consists of one or two MNIST digits (LeCun et al., 1998) moving linearly and randomly bouncing on walls with new direction and velocity sampled randomly at each bounce (Denton & Fergus, 2018).

Figure 3 (left) shows quantitative results with two digits. Our model outperforms SVG on both PSNR and SSIM; LPIPS and FVD are not reported as they are not relevant for this synthetic task. Decoupling dynamics from image synthesis allows our method to maintain temporal consistency despite high-uncertainty frames where crossing digits become indistinguishable. For instance in Figure 2, the digits shapes change after they cross in the SVG prediction, while our model predicts the correct digits. To evaluate the predictive ability on a longer horizon, we perform experiments on the deterministic version of the dataset (Srivastava et al., 2015) with only one prediction per model to compute PSNR and SSIM. We show the results up to  $t + 95$  in Figure 3 (right). We can see that our model better captures the dynamics of the problem compared to SVG as its performance decreases significantly less over time, especially at a long-term horizon.

We also compare to two alternative versions of our model in Figure 3, where the residual dynamic function is replaced by an MLP or a GRU (Gated Recurrent Unit, Cho et al., 2014). Our residual model outperforms both versions on

the stochastic, and especially on the deterministic version of the dataset, showing its intrinsic advantage at modeling long-term dynamics. Finally, on the deterministic version of Moving MNIST, we compare to an alternative where  $z$  is entirely removed, resulting in a temporal model close to the one presented by Chen et al. (2018). The loss of performance of this alternative model is significant, showing that our stochastic residual model offers a substantial advantage even when used in a deterministic environment.

**KTH Action dataset (KTH).** This dataset is composed of real-world videos of people performing a single action per video in front of different backgrounds (Schüldt et al., 2004). Uncertainty lies in the appearance of subjects, the actions they perform, and how they are performed.

We substantially outperform on this dataset every considered baseline for each metric, as shown in Figure 4 and Table 1. In some videos, the subject only appears after the conditioning frames, requiring the model to sample the moment and location of the subject appearance, as well as its action. This critical case is illustrated in Figure 5. There, SVG fails to even generate a moving person; only SAVP and our model manage to do so, and our best sample is closer to the subject’s poses compared to SAVP. Moreover, the worst sample of our model demonstrates that it captures the diversity of the dataset by making a person appear at different time steps and with different speeds. An additional experiment on this dataset in Appendix G studies the influence of the encoder and decoder architecture on SVG and our model.

Finally, Table 1 and appendix Table 3 compare our method to its MLP and GRU alternative versions, leading to two

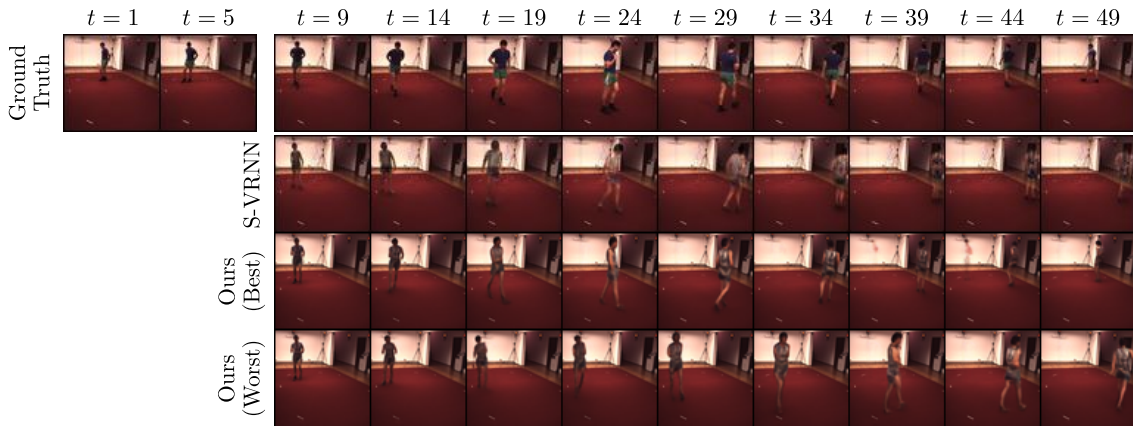


Figure 6. Conditioning frames and corresponding ground truth, best samples from StructVRNN and our method, and worst sample from our method, with respect to LPIPS, for a video of the Human3.6M dataset. Our method better captures the dynamic of the subject and produces less artefacts than StructVRNN.

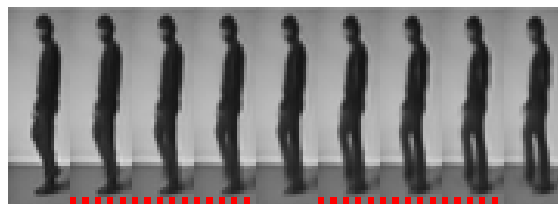
conclusions. Firstly, it confirms the structural advantage of residual dynamics observed on Moving MNIST. Indeed, both MLP and GRU lose on all metrics, and especially in terms of realism according to LPIPS and FVD. Secondly, all three versions of our model (residual, MLP, GRU) outperform prior methods. Therefore, this improvement is due to their common inference method, latent nature and content variable, strengthening our motivation to propose a non-autoregressive model.

**Human3.6M.** This dataset is also made of videos of subjects performing various actions (Ionescu et al., 2011; 2014). While there are more actions and details to capture with less training subjects than in KTH, the video backgrounds are less varied, and subjects always remain within the frames.

As reported in Figure 4 and Table 1, we significantly outperform, with respect to all considered metrics, StructVRNN, which is the state of the art on this dataset and has been shown to surpass both SAVP and SVG by Minderer et al. (2019). Figure 6 shows the dataset challenges; in particular, both methods do not capture well the subject appearance. Nonetheless, our model better captures its movements, and produces more realistic frames.

Comparisons to the MLP and GRU versions demonstrate once again the advantage of using residual dynamics. GRU obtains low scores on all metrics, which is coherent with similar results for SVG reported by Minderer et al. (2019). While the MLP version remains close to the residual model on PSNR, SSIM and LPIPS, it is largely beaten by the latter in terms of FVD.

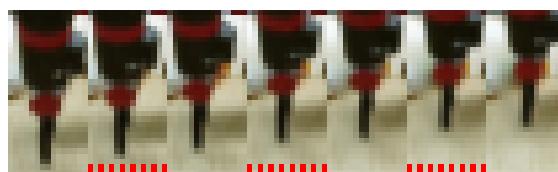
**BAIR robot pushing dataset (BAIR).** This dataset contains videos of a Sawyer robotic arm pushing objects on a tabletop (Ebert et al., 2017). It is highly stochastic as the arm can change its direction at any moment. Our method



(a) Cropped KTH sample.



(b) Cropped Human3.6M sample.



(c) Cropped BAIR sample.

Figure 7. Generation examples at doubled or quadrupled frame rate, using a halved  $\Delta t$  compared to training. Frames including a bottom red dashed bar are intermediate frames.

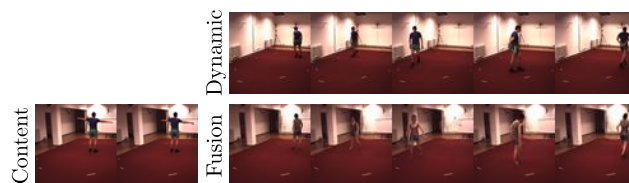


Figure 8. Video (bottom right) generated from the dynamic latent state  $y$  inferred with a video (top) and the content variable  $w$  computed with the conditioning frames of another video (left). The generated video keeps the same background as the bottom left frames, while the subject moves accordingly to the top frames.



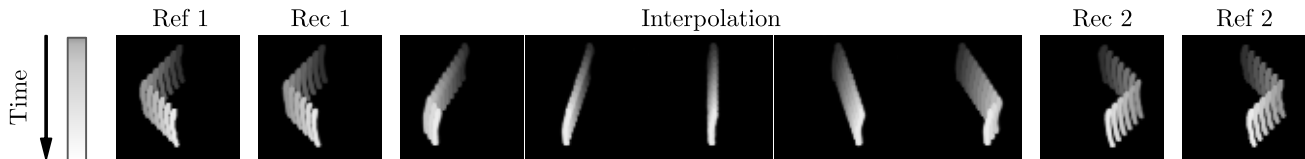


Figure 9. From left to right,  $\mathbf{x}^s$ ,  $\hat{\mathbf{x}}^s$  (reconstruction of  $\mathbf{x}^s$  by the VAE of our model), results of the interpolation in the latent space between  $\mathbf{x}^s$  and  $\mathbf{x}^t$ ,  $\hat{\mathbf{x}}^t$  and  $\mathbf{x}^t$ . Each trajectory is materialized in shades of grey in the frames.

achieves similar or better results compared to state-of-the-art models in terms of PSNR, SSIM and LPIPS, as shown in Figure 4, except for SV2P that produces very blurry samples, as seen in Appendix H, yielding good PSNR but prohibitive LPIPS scores. Our method obtains second-best FVD score, close to SAVP whose adversarial loss enables it to better model small objects, and outperforms SVG, whose variational architecture is closest to ours, demonstrating the advantage of non-autoregressive methods. Recent advances (Villegas et al., 2019) indicate that performance of such variational models can be improved by increasing networks capacities, but this is out of the scope of this paper.

**Varying frame rate in testing.** We challenge here the ODE inspiration of our model. Equation (2) amounts to learning a residual function  $f_{z_{\lfloor t \rfloor + 1}}$  over  $t \in [\lfloor t \rfloor, \lfloor t \rfloor + 1)$ . We aim at testing whether this dynamics is close to its continuous generalization:

$$\frac{d\mathbf{y}}{dt} = f_{z_{\lfloor t \rfloor + 1}}(\mathbf{y}), \quad (9)$$

which is a piecewise ODE. To this end, we refine this Euler approximation during testing by halving  $\Delta t$ ; if this maintains the performance of our model, then the dynamic rule of the latter is close to the piecewise ODE. As shown in Figure 4 and Table 1, prediction performances overall remain stable while generating twice as many frames (cf. Appendix F for further discussion). Therefore, the justification of the proposed update rule is supported by empirical evidence. This property can be used to generate videos at a higher frame rate, with the same model, and without supervision. We show in Figure 7 and Appendix F frames generated at a double and quadruple frame rate on KTH, Human3.6M and BAIR.

**Disentangling dynamics and content.** Let us show that the proposed model actually separates content from dynamics as discussed in Section 3.2. To this end, two sequences  $\mathbf{x}^s$  and  $\mathbf{x}^t$  are drawn from the Human3.6M testing set. While  $\mathbf{x}^s$  is used for extracting our content variable  $\mathbf{w}^s$ , dynamic states  $\mathbf{y}^t$  are inferred with our model from  $\mathbf{x}^t$ . New frame sequences  $\hat{\mathbf{x}}$  are finally generated from the fusion of the content vector and the dynamics. This results in a content corresponding to the first sequence  $\mathbf{x}^s$  and a movement following the dynamics of the second sequence  $\mathbf{x}^t$ , as

observed in Figure 8. More samples for KTH, Human3.6M, and BAIR can be seen in Appendix H.

**Interpolation of dynamics.** Our state-space structure allows us to learn semantic representations in  $\mathbf{y}_t$ . To highlight this feature, we test whether two deterministic Moving MNIST trajectories can be interpolated by linearly interpolating their inferred latent initial conditions. We begin by generating two trajectories  $\mathbf{x}^s$  and  $\mathbf{x}^t$  of a single moving digit. We infer their respective latent initial conditions  $\mathbf{y}_1^s$  and  $\mathbf{y}_1^t$ . We then use our model to generate frame sequences from latent initial conditions linearly interpolated between  $\mathbf{y}_1^s$  and  $\mathbf{y}_1^t$ . If it learned a meaningful latent space, the resulting trajectory should also be a smooth interpolation between the directions of reference trajectories  $\mathbf{x}^s$  and  $\mathbf{x}^t$ , and this is what we observe in Figure 9. Additional examples can be found in Appendix H.

## 5. Conclusion

We introduce a novel dynamic latent model for stochastic video prediction which, unlike prior image-autoregressive models, decouples frame synthesis and dynamics. This temporal model is based on residual updates of a small latent state that is showed to perform better than RNN-based models. This endows our method with several desirable properties, such as temporal efficiency and latent space interpretability. We experimentally demonstrate the performance and advantages of the proposed model, which outperforms prior state-of-the-art methods for stochastic video prediction. This work is, to the best of our knowledge, the first to propose a latent dynamic model that scales for video prediction. The proposed model is also novel with respect to the recent line of work dealing with neural networks and ODEs for temporal modeling; it is the first such residual model to scale to complex stochastic data such as videos.

We believe that the general principles of our model (state-space, residual dynamic, static content variable) can be generally applied to other models as well. Interesting future works include replacing the VRNN of Minderer et al. (2019) with our residual dynamics in order to model the evolution of key-points, supplementing our model with more video-specific priors, or leveraging its state-space nature in model-based reinforcement learning.

## Acknowledgements

We would like to thank all members of the MLIA team from the LIP6 laboratory of Sorbonne Université for helpful discussions and comments, as well as Matthias Minderer for his help to process the Human3.6M dataset and reproduce StructVRNN results.

We acknowledge financial support from the LOCUST ANR project (ANR-15-CE23-0027). This work was granted access to the HPC resources of IDRIS under the allocation 2020-AD011011360 made by GENCI (Grand Equipement National de Calcul Intensif).

## References

- Babaeizadeh, M., Finn, C., Erhan, D., Campbell, R., and Levine, S. Stochastic variational video prediction. In *International Conference on Learning Representations*, 2018.
- Bayer, J. and Osendorfer, C. Learning stochastic recurrent networks. *arXiv preprint arXiv:1411.7610*, 2014.
- Behrmann, J., Grathwohl, W., Chen, R. T. Q., Duvenaud, D., and Jacobsen, J.-H. Invertible residual networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 573–582, Long Beach, California, USA, June 2019. PMLR.
- Caballero, J., Ledig, C., Aitken, A., Acosta, A., Totz, J., Wang, Z., and Shi, W. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2848–2857, July 2017.
- Castrejon, L., Ballas, N., and Courville, A. Improved conditional VRNNs for video prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, pp. 7608–7617, October 2019.
- Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. Neural ordinary differential equations. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 6571–6583. Curran Associates, Inc., 2018.
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.
- Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A., and Bengio, Y. A recurrent latent variable model for sequential data. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 2980–2988. Curran Associates, Inc., 2015.
- De Brouwer, E., Simm, J., Arany, A., and Moreau, Y. GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series. In Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 7379–7390. Curran Associates, Inc., 2019.
- de Bézenac, E., Ayed, I., and Gallinari, P. Optimal unsupervised domain translation. 2019.
- Denton, E. and Birodkar, V. Unsupervised learning of disentangled representations from video. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4414–4423. Curran Associates, Inc., 2017.
- Denton, E. and Fergus, R. Stochastic video generation with a learned prior. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning Research*, pp. 1174–1183, Stockholmsmässan, Stockholm, Sweden, July 2018. PMLR.
- Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., and Garcia, R. Incorporating second-order functional knowledge for better option pricing. In Leen, T. K., Dietterich, T. G., and Tresp, V. (eds.), *Advances in Neural Information Processing Systems 13*, pp. 472–478. MIT Press, 2001.
- Ebert, F., Finn, C., Lee, A. X., and Levine, S. Self-supervised visual planning with temporal skip connections. In Levine, S., Vanhoucke, V., and Goldberg, K. (eds.), *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pp. 344–356. PMLR, November 2017.
- Fan, H., Zhu, L., and Yang, Y. Cubic LSTMs for video prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8263–8270, 2019.
- Finn, C., Goodfellow, I., and Levine, S. Unsupervised learning for physical interaction through video prediction. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 64–72. Curran Associates, Inc., 2016.

- Fraccaro, M., Sønderby, S. K., Paquet, U., and Winther, O. Sequential neural models with stochastic layers. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2199–2207. Curran Associates, Inc., 2016.
- Fraccaro, M., Kamronn, S., Paquet, U., and Winther, O. A disentangled recognition and nonlinear dynamics model for unsupervised learning. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 3601–3610. Curran Associates, Inc., 2017.
- Gao, H., Xu, H., Cai, Q.-Z., Wang, R., Yu, F., and Darrell, T. Disentangling propagation and generation for video prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, pp. 9006–9015, October 2019.
- Goodfellow, I. NIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014.
- Graves, A. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.
- Gregor, K., Papamakarios, G., Besse, F., Buesing, L., and Weber, T. Temporal difference variational auto-encoder. In *International Conference on Learning Representations*, 2019.
- Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., and Davidson, J. Learning latent dynamics for planning from pixels. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565, Long Beach, California, USA, June 2019. PMLR.
- He, J., Lehrmann, A., Marino, J., Mori, G., and Sigal, L. Probabilistic video generation using holistic attribute control. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y. (eds.), *The European Conference on Computer Vision (ECCV)*, pp. 466–483. Springer International Publishing, September 2018.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016.
- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A.  $\beta$ -VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.
- Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- Ionescu, C., Li, F., and Sminchisescu, C. Latent structured models for human pose estimation. In *2011 International Conference on Computer Vision*, pp. 2220–2227, November 2011.
- Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, July 2014.
- Jia, X., De Brabandere, B., Tuytelaars, T., and Van Gool, L. Dynamic filter networks. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 667–675. Curran Associates, Inc., 2016.
- Jiang, H., Sun, D., Jampani, V., Yang, M.-H., Learned-Miller, E., and Kautz, J. Super SloMo: High quality estimation of multiple intermediate frames for video interpolation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9000–9008, June 2018.
- Jin, B., Hu, Y., Tang, Q., Niu, J., Shi, Z., Han, Y., and Li, X. Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4554–4563, June 2020.
- Kalchbrenner, N., van den Oord, A., Simonyan, K., Danihelka, I., Vinyals, O., Graves, A., and Kavukcuoglu, K. Video pixel networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1771–1779, International Convention Centre, Sydney, Australia, August 2017. PMLR.
- Karl, M., Soelch, M., Bayer, J., and van der Smagt, P. Deep variational Bayes filters: Unsupervised learning of state space models from raw data. In *International Conference on Learning Representations*, 2017.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

- Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 10215–10224. Curran Associates, Inc., 2018.
- Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.
- Krishnan, R. G., Shalit, U., and Sontag, D. Structured inference networks for nonlinear state space models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pp. 2101–2109, 2017.
- Kullback, S. and Leibler, R. A. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1): 79–86, 1951.
- Kumar, M., Babaeizadeh, M., Erhan, D., Finn, C., Levine, S., Dinh, L., and Kingma, D. VideoFlow: A conditional flow-based model for stochastic video generation. In *International Conference on Learning Representations*, 2020.
- Le Guen, V. and Thome, N. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11474–11484, June 2020.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.
- Lee, A. X., Zhang, R., Ebert, F., Abbeel, P., Finn, C., and Levine, S. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018.
- Liang, X., Lee, L., Dai, W., and Xing, E. P. Dual motion GAN for future-flow embedded video prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, pp. 1762–1770, October 2017.
- Liu, Z., Yeh, R. A., Tang, X., Liu, Y., and Agarwala, A. Video frame synthesis using deep voxel flow. In *The IEEE International Conference on Computer Vision (ICCV)*, pp. 4473–4481, October 2017.
- Liu, Z., Wu, J., Xu, Z., Sun, C., Murphy, K., Freeman, W. T., and Tenenbaum, J. B. Modeling parts, structure, and system dynamics via predictive learning. In *International Conference on Learning Representations*, 2019.
- Long, Z., Lu, Y., Ma, X., and Dong, B. PDE-net: Learning PDEs from data. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3208–3216, Stockholmsmässan, Stockholm Sweden, July 2018. PMLR.
- Lotter, W., Kreiman, G., and Cox, D. Deep predictive coding networks for video prediction and unsupervised learning. In *International Conference on Learning Representations*, 2017.
- Lu, C., Hirsch, M., and Schölkopf, B. Flexible spatio-temporal networks for video prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2137–2145, July 2017a.
- Lu, Y., Zhong, A., Li, Q., and Dong, B. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. *arXiv preprint arXiv:1710.10121*, 2017b.
- Mathieu, M., Couprie, C., and LeCun, Y. Deep multi-scale video prediction beyond mean square error. In *International Conference on Learning Representations*, 2016.
- Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., and Wu, H. Mixed precision training. In *International Conference on Learning Representations*, 2018.
- Minderer, M., Sun, C., Villegas, R., Cole, F., Murphy, K., and Lee, H. Unsupervised learning of object structure and dynamics from videos. In Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 92–102. Curran Associates, Inc., 2019.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8026–8037. Curran Associates, Inc., 2019.
- Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016.
- Ranzato, M., Szeliski, A., Bruna, J., Mathieu, M., Collobert, R., and Chopra, S. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*, 2014.

- Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1278–1286, Beijing, China, June 2014. PMLR.
- Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F. (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing.
- Rousseau, F., Drumetz, L., and Fablet, R. Residual networks as flows of diffeomorphisms. *Journal of Mathematical Imaging and Vision*, May 2019.
- Rubanova, Y., Chen, R. T. Q., and Duvenaud, D. Latent ordinary differential equations for irregularly-sampled time series. In Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 5320–5330. Curran Associates, Inc., 2019.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. *Neurocomputing: Foundations of Research*, chapter Learning Representations by Back-Propagating Errors, pp. 696–699. MIT Press, Cambridge, MA, USA, 1988.
- Ryder, T., Golightly, A., McGough, A. S., and Prangle, D. Black-box variational inference for stochastic differential equations. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4423–4432, Stockholm, Sweden, July 2018. PMLR.
- Santoro, A., Raposo, D., Barrett, D. G. T., Malinowski, M., Pascanu, R., Battaglia, P., and Lillicrap, T. A simple neural network module for relational reasoning. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4967–4976. Curran Associates, Inc., 2017.
- Schüldt, C., Laptev, I., and Caputo, B. Recognizing human actions: A local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pp. 32–36, August 2004.
- Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., and Woo, W.-c. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 802–810. Curran Associates, Inc., 2015.
- Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.
- Srivastava, N., Mansimov, E., and Salakhudinov, R. Unsupervised learning of video representations using LSTMs. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 843–852, Lille, France, July 2015. PMLR.
- Tulyakov, S., Liu, M.-Y., Yang, X., and Kautz, J. MoCoGAN: Decomposing motion and content for video generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1526–1535, June 2018.
- Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., and Gelly, S. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.
- van den Oord, A., Kalchbrenner, N., Espeholt, L., Kavukcuoglu, K., Vinyals, O., and Graves, A. Conditional image generation with PixelCNN decoders. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 4790–4798. Curran Associates, Inc., 2016.
- van Steenkiste, S., Chang, M., Greff, K., and Schmidhuber, J. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. In *International Conference on Learning Representations*, 2018.
- Villegas, R., Yang, J., Hong, S., Lin, X., and Lee, H. Decomposing motion and content for natural video sequence prediction. In *International Conference on Learning Representations*, 2017.
- Villegas, R., Pathak, A., Kannan, H., Erhan, D., Le, Q. V., and Lee, H. High fidelity video prediction with large stochastic recurrent neural networks. In Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 81–91. Curran Associates, Inc., 2019.
- Vondrick, C. and Torralba, A. Generating the future with adversarial transformers. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2992–3000, July 2017.

- Walker, J., Gupta, A., and Hebert, M. Dense optical flow prediction from a static image. In *The IEEE International Conference on Computer Vision (ICCV)*, pp. 2443–2451, December 2015.
- Walker, J., Doersch, C., Gupta, A., and Hebert, M. An uncertain future: Forecasting from static images using variational autoencoders. In Leibe, B., Matas, J., Sebe, N., and Welling, M. (eds.), *The European Conference on Computer Vision (ECCV)*, pp. 835–851. Springer International Publishing, October 2016.
- Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., and Catanzaro, B. Video-to-video synthesis. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 1144–1156. Curran Associates, Inc., 2018.
- Weissenborn, D., Täckström, O., and Uszkoreit, J. Scaling autoregressive video models. In *International Conference on Learning Representations*, 2020.
- Wichers, N., Villegas, R., Erhan, D., and Lee, H. Hierarchical long-term video prediction without supervision. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 6038–6046, Stockholm, Sweden, July 2018. PMLR.
- Wu, Y., Gao, R., Park, J., and Chen, Q. Future video synthesis with object motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5539–5548, June 2020.
- Xu, J., Ni, B., Li, Z., Cheng, S., and Yang, X. Structure preserving video prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1460–1469, June 2018a.
- Xu, J., Ni, B., and Yang, X. Video prediction via selective sampling. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 1705–1715. Curran Associates, Inc., 2018b.
- Xue, T., Wu, J., Bouman, K. L., and Freeman, W. T. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 91–99. Curran Associates, Inc., 2016.
- Yingzhen, L. and Mandt, S. Disentangled sequential autoencoder. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5670–5679, Stockholm, Sweden, July 2018. PMLR.
- Yıldız, C., Heinonen, M., and Lahdesmaki, H. ODE<sup>2</sup>VAE: Deep generative second order odes with Bayesian neural networks. In Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 13412–13421. Curran Associates, Inc., 2019.
- Zaheer, M., Kottur, S., Ravanbakhsh, S., Póczos, B., Salakhutdinov, R., and Smola, A. J. Deep sets. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 3391–3401. Curran Associates, Inc., 2017.
- Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, June 2018.