2021 ICCV OCTOBER 11-17 VIRTUAL

SLAMP: Stochastic Latent Appearance and Motion Prediction

Adil Kaan Akan[1], Erkut Erdem[2], Aykut Erdem[1], Fatma Güney[1]

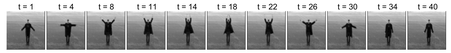
[1] KUIS AI Center, Koç University, [2] Hacettepe University Computer Vision Lab

Stochastic Video Prediction

Goal: Given a few frames in a video, predict next frames.

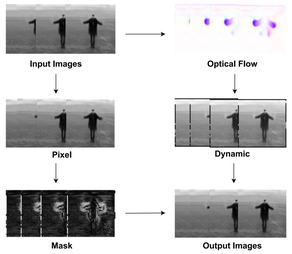
Challenges:

- High dimensionality of the problem
- Uncertainty of the future
- Requires an understanding of the scene structure, motion in the scene, and object relations, etc.



Given few frames of a video, our goal is to predict future frames.

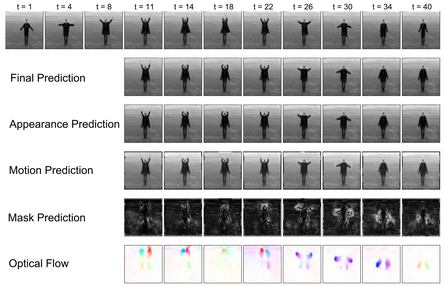
SLAMP



SLAMP Pipeline

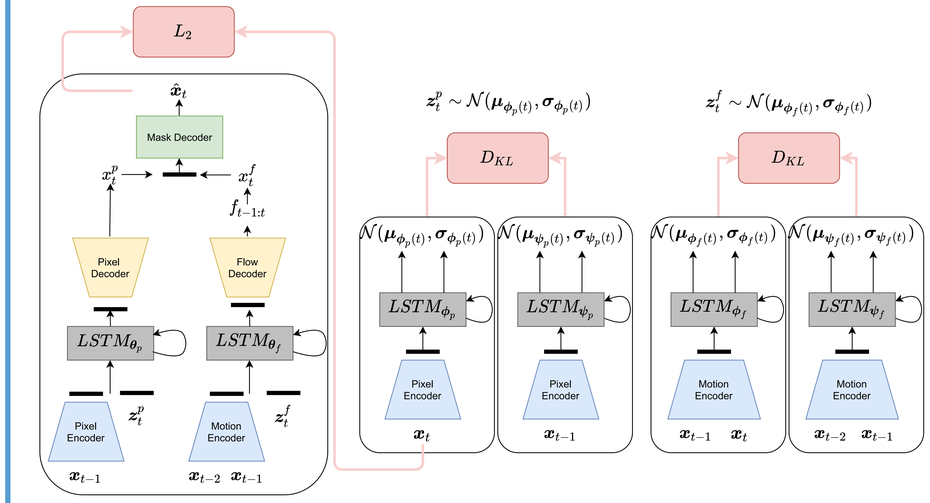
Steps of SLAMP Inference

- Encode past frames to create rich image features
- Sample latent variables from prior distributions
- Predict next frame's features and optical flow features by conditioning on latent variables
- Decode predicted features to next frame and optical flow for warping current frame to the next one
- Predict a mask to decide which prediction to choose
- Combine both predictions, pixel and dynamic, to generate final prediction



Detailed predictions

Model block



SLAMP has three different decoders, pixel, flow and mask. It decodes two candidate images from pixel and flow decoders and combine them via predicted mask. It also has two separate prior and posterior distributions, one models pixel space and the other models motion history.

Methodology

Future Information

We learn two different posterior distributions, pixel and motion, from the future frames.

$$\mathbf{h}_t^p = \text{PixelEnc}(\mathbf{x}_t) \qquad \mathbf{h}_t^f = \text{MotionEnc}(\mathbf{x}_{t-1}, \mathbf{x}_t)$$
$$\boldsymbol{\mu}_{\phi_p(t)}, \boldsymbol{\sigma}_{\phi_p(t)} = \text{LSTM}_{\phi_p}(\mathbf{h}_t^p) \qquad \boldsymbol{\mu}_{\phi_f(t)}, \boldsymbol{\sigma}_{\phi_f(t)} = \text{LSTM}_{\phi_f}(\mathbf{h}_t^f)$$

Past Information

We learn two different prior distributions, beause it is better than using fixed prior as shown in [1], pixel and motion, from the past frames.

$$\mathbf{h}_{t-1}^p = \text{PixelEnc}(\mathbf{x}_{t-1}) \qquad \mathbf{h}_{t-1}^f = \text{MotionEnc}(\mathbf{x}_{t-2}, \mathbf{x}_{t-1})$$
$$\boldsymbol{\mu}_{\psi_p(t-1)}, \boldsymbol{\sigma}_{\psi_p(t-1)} = \text{LSTM}_{\psi_p}(\mathbf{h}_{t-1}^p) \qquad \boldsymbol{\mu}_{\psi_f(t)}, \boldsymbol{\sigma}_{\psi_f(t)} = \text{LSTM}_{\psi_f}(\mathbf{h}_{t-1}^f)$$

Frame prediction

We predict pixel and motion in high-leval feature space.

$$\mathbf{g}_t^p = \text{LSTM}_{\theta_p}(\mathbf{h}_{t-1}^p, \mathbf{z}_t^p) \qquad \mathbf{g}_t^f = \text{LSTM}_{\theta_f}(\mathbf{h}_{t-1}^f, \mathbf{z}_t^f)$$

Decoding

We decode predicted features and combine them with a predicted mask.

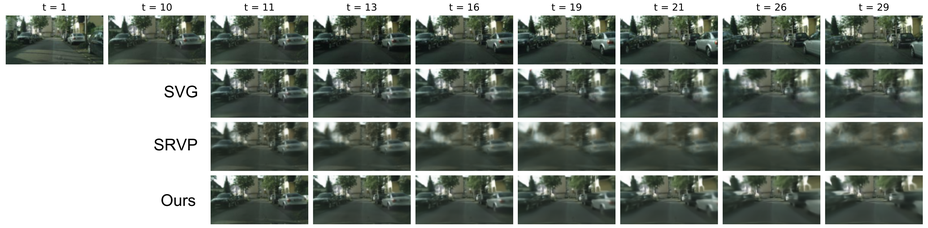
$$\boldsymbol{\mu}_{\theta_p} = \text{PixelDec}(\mathbf{g}_t^p) \qquad \boldsymbol{\mu}_{\theta_f} = \text{FlowDec}(\mathbf{g}_t^f)$$
$$\mathbf{x}_t^p = \boldsymbol{\mu}_{\theta_p} \qquad \mathbf{x}_t^f = Warp(\mathbf{x}_{t-1}, \boldsymbol{\mu}_{\theta_f})$$
$$\hat{\mathbf{x}}_t = \mathbf{m}(\mathbf{x}_t^p, \mathbf{x}_t^f) \odot \mathbf{x}_t^p + (\mathbf{1} - \mathbf{m}(\mathbf{x}_t^p, \mathbf{x}_t^f)) \odot \mathbf{x}_t^f$$

Evidence Lower Bound (ELBO)

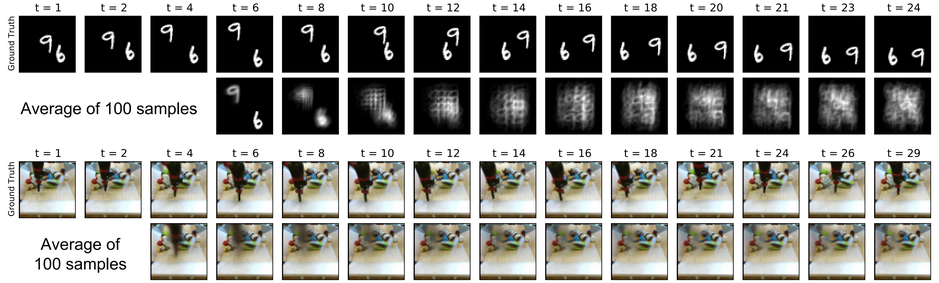
We maximize the following Evidence Lower Bound to optimize our model.

$$\log\, p_{\boldsymbol{\theta}}(\mathbf{x}) \geq \mathcal{L}_{\boldsymbol{\theta},\phi_p,\phi_f,\psi_p,\psi_f}(\mathbf{x}_{1:T}) = \sum_t E_{\substack{\mathbf{z}_{1:t}^p \sim q_{\phi_p} \\ \mathbf{z}_{1:t}^f \sim q_{\phi_f}}} \log\, p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t}^p, \mathbf{z}_{1:t}^f) - \beta\Big[q(\mathbf{z}_t^p|\mathbf{x}_{1:t})p(\mathbf{z}_t^p|\mathbf{x}_{1:t-1}) + q(\mathbf{z}_t^f|\mathbf{x}_{1:t})p(\mathbf{z}_t^f|\mathbf{x}_{1:t-1})\Big]$$

Qualitative Results



Comparisons with baseline and state-of-the-art methods on Cityscapes.



Diversity Experiments

Quantitative Results

	Models	PSNR (↑)	SSIM (↑)	LPIPS (↓)
KTH	SV2P [2]	28.19 ± 0.31	0.8141 ± 0.0050	0.2049 ± 0.0053
	SAVP [3]	26.51 ± 0.29	0.7564 ± 0.0062	0.1120 ± 0.0039
	SVG [1]	28.06 ± 0.29	0.8438 ± 0.0054	0.0923 ± 0.0038
	SRVP [4]	**29.69 ± 0.32**	**0.8697 ± 0.0046**	**0.0736 ± 0.0029**
	SLAMP	29.39 ± 0.30	0.8646 ± 0.0050	0.0795 ± 0.0034
BAIR	SV2P [2]	**20.39 ± 0.27**	0.8169 ± 0.0086	0.0912 ± 0.0053
	SAVP [3]	18.44 ± 0.25	0.7887 ± 0.0092	0.0634 ± 0.0026
	SVG [1]	18.95 ± 0.26	0.8058 ± 0.0088	0.0609 ± 0.0034
	SRVP [4]	19.59 ± 0.27	**0.8196 ± 0.0084**	**0.0574 ± 0.0032**
	SLAMP	19.67 ± 0.26	0.8161 ± 0.0086	0.0639 ± 0.0037

Results on generic video prediction datasets

Our model performs comparable to state-of-the-art, SRVP [4], on simple datasets, Moving MNIST, KTH, BAIR. However, it outperforms state-of-the-art SRVP on real-world datasets, KITTI and Cityscapes with challenging background motion.

	Models	PSNR (↑)	SSIM (↑)	LPIPS (↓)
KITTI	SVG [1]	12.70 ± 0.70	0.329 ± 0.030	0.594 ± 0.034
	SRVP [4]	13.41 ± 0.42	0.336 ± 0.034	0.635 ± 0.021
	SLAMP	**13.46 ± 0.74**	**0.337 ± 0.034**	**0.537 ± 0.042**
CITY	SVG [1]	20.42 ± 0.63	0.606 ± 0.023	0.340 ± 0.022
	SRVP [4]	20.97 ± 0.43	0.603 ± 0.016	0.447 ± 0.014
	SLAMP	**21.73 ± 0.76**	**0.649 ± 0.025**	**0.2941 ± 0.022**

Results on Real-world datasets with moving background

Concluding Remarks

- Explicit motion modelling via optical flow
- Implicit learning of moving parts of the scene via mask
- Comparable results on generic video prediction datasets
- State-of-the-art results on challenging real-world datasets with moving background

References

[1] E. Denton and R. Fergus, "Stochastic video generation with a learned prior," 2018.

[2] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine, "Stochastic variational video prediction," 2018.

[3] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine, "Stochastic adversarial video prediction," 2018.

[4] J.-Y. Franceschi, E. Delasalles, M. Chen, S. Lamprier, and P. Gallinari, "Stochastic latent residual video prediction," 2020.

Contact Information

Our results are seen best in video, please check our website: https://kuis-ai.github.io/slamp.

[Email] {kakan20, aerdem, fguney}@ku.edu.tr, erkut@cs.hacettepe.edu.tr