Stochastic Video Prediction with Perceptual Loss

Donghun Lee, Ingook Jang, Seonghyun Kim, Chanwon Park, Junhee Park
Electronics and Telecommunications Research Institute (ETRI)
Daejeon, South Korea
{donghun, ingook, kim-sh, cwp, juni}@etri.re.kr

Abstract

Predicting future states is a challenging process in the decision-making system because of its inherently uncertain nature. Most works in this literature are based on deep generative networks such as variational autoencoder which uses pixel-wise reconstruction in their loss functions. Predicting the future with pixel-wise reconstruction could fail to capture the full distribution of high-level representations and result in inaccurate and blurred predictions. In this paper, we propose stochastic video generation with perceptual loss (SVG-PL) to improve uncertainty and blurred area in future prediction. The proposed model combines perceptual loss function and pixel-wise loss function for image reconstruction and future state predictions. The model is built on a variational autoencoder to reduce high dimensionality to latent variable to capture both spatial information and temporal dynamics of future prediction. We show that utilization of perceptual loss on video prediction improves reconstruction ability and result in clear predictions. Improvements in video prediction could further help the decision-making process in multiple downstream applications.

1 Introduction

Future prediction is an important and challenging process in a decision-making process. Learning full distribution of spatial dynamics and temporal dynamics could result accurate prediction of future states which could be utilized in multiple downstream applications such as pedestrian prediction [1], autonomous driving [2], sport games [3], networks [4], anomaly detection [5] and model-based reinforcement learning [6]. One of the challenging issues in video prediction research field is inherently uncertain model dynamics. For example, when a bouncing random ball hits a wall, it is uncertain that which direction and velocity the ball would bounce in future states. Video prediction in this scenario is to predict features of the target object in future states such as direction, position, and velocity of the bouncing ball in continuous images where past states are given as input. Recent researches in pixel-based video prediction shows noticeable advances in stochastic state prediction [6]. The application of deep generative model outperformed research approaches with traditional methods because of their improved ability to infer temporal dynamics of past and future and learn spatial dynamics of high dimensional state environments [7].

Video is a collection of consecutive static images which contains temporal information and motion patterns. Prediction on video consists with two components which are reconstruction of spatial information and inference of temporal dynamics. Disentangle these two and find spatio-temporal correlation is an important research area [8]. Prediction in future state could be affected by external interventions such as occlusion, change in camera positions, direction of lights, background objects, and many other factors. The future state is multimodal and contains uncertainty by nature. In deterministic approach, each possible future state is learned to have same averaged probability and

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

visually shown as a blurred area especially if time range is in long horizon. Their deterministic model has limited application in real-world downstream applications which contains inherent uncertainty. In stochastic approach, recent studies applied variational autoencoder [9] to reduce high dimensionality to latent variables and to better incorporate uncertainty.

Variational autoencoder (VAE) [9] has been employed for stochastic prediction problems. Walker et al. [10] proposed VAE conditioned on images to predict dense trajectories from pixels. Xue et al. [11] proposed cross convolution network with VAE to learn future frames with single input image. Babaeizadeh et al. [12] proposed stochastic variational video prediction based on variational inference with real-world videos. Denton and Fergus [6] proposed stochastic video generation with learned prior. Villegas et al. [13] extended SVG while minimizing inductive bias and maximizing network capacities. Previous approaches highlighted pixelwise reconstruction of future states in stochastic model. Our approach differs from these in that per-pixel reconstruction is extended with perceptual loss function to extract high-level features.

Perceptual loss [14] has been introduced to extract and differentiate high-level feature representations from pretrained perceptual networks. Ledig et al. [15] and Sajjadi et al. [16] utilized perceptual loss and adversarial loss to generate super-resolution single image. Zhu et al. [17] learns natural image manifold directly from data. Reda et al. [18] proposed spatially-displaced convolution to learn a motion vector and a kernel to synthesize images. Hou et al. [19] proposed perceptual loss for VAE training to extract natural visual appearance and perceptual quality of a image. Pihlgren et al. [20] applied perceptual loss to improve image embeddings in static settings. These approaches improved reconstruction quality of images by utilizing high-level features from perceptual loss, however, limited research is done on video prediction with uncertainty. Our approach utilizes both pixel-wise loss function and perceptual loss function to predict future states in continuous time frame.

In this paper, stochastic video generation with perceptual loss (SVG-PL) is proposed. A pixel-wise loss and a perceptual loss are both used to predict future states in stochastic approach to improve blurriness and uncertainty of the video prediction. Especially high-level representations are extracted from pretrained perceptual network from natural image dataset. This approach could improve reconstruction ability and result in clear predictions. Also state prediction with high-level representations could help learning control policy in model-based reinforcement learning because high-level representations such as positions of a target and an acting agent are more important than pixel-wise reconstruction.

2 Method

The formal description of a pixel-based future state prediction is addressed. Let $x_t \in R^{c*w*h)}$ is a state for given timestep $t \in T$ where T, c, w, h, stands for total timesteps, channel, weight, and height. Let $X = \{x_{(t-n)}, x_{(t-n+1)}, \ldots, x_{(t-2)}, x_{(t-1)}\}$ is a set of states from n past states to current state $t-1$. Main goal is to predict the next frame x_t based on previous sequence set X and latent variable z_t.

Proposed approach consists of three module which are the Inference Module, the Prior Learning Module and the Prediction Module. Combination of MSE, Perceptual loss and KL loss [21] is applied to form Perceptual Reconstruction Loss function. The Inference Module learns posterior $q_\phi(z_t \mid x_{(1:t)})$ which is a distribution of a latent variable z where states from timestep 1 to t are given. The latent variable z contains temporal information of states starting from timestep 1 to t. The Prior Learning Module learns the prior $p_\psi(z_t \mid x_{(1:t-1)})$ at each timestep which is a distribution of a latent variable z where state from timestep 1 to $t-1$ are given. KL loss is calculated with the posterior $q_\phi(z_t \mid x_{(1:t)})$ to have close distribution to the prior $p_\psi(z_t \mid x_{(1:t-1)})$. The Prediction Module infers a future state image $\widehat{x_t}$ from a past state x_{t-1} and a latent variable z_t learned from the Inference Module and the Prior Learning Module. Perceptual loss and MSE are both used as perceptual reconstruction loss to evaluate predicted image $\widehat{x_t}$ with original image x_t. The KL loss and the perceptual reconstruction loss are combined for learning processes.

2.1 Inference Module

The inference module is adapted with formalism of beta variational auto-encoder [22]. It learns posterior $q_\phi(z_t \mid x_{(1:t)})$ which is a distribution of a latent variable z where states from timestep 1 to t

are given. The latent variable z has reduced dimensionality and contains temporal information about past t states. Reduced dimensionality makes learning processes fast and easy to learn the distribution of representation. The module is learned with maximization of the variational lower bound:

$$L_{\theta,\phi}(x_{1:T}) = \sum_{t=1}^{T}[L_{perceptual\_recon} - \beta D_{KL}(q_{\phi}(z_t|x_{1:t})||p(z))]$$

2.2 Prior Learning Module

The Prior Learning module is built on Stochastic Video Generation algorithm with learned prior (SVG-LP)[6]. Gaussian distribution with fixed variance $\mathcal{N}$(0,1) is not used here. Instead this approach learns a prior which varies across time. Prior $p_{\psi}(z_t \mid x_{(1:t-1)})$ is calculated at every timestep with all past frames up to timestep $t-1$ and conditional gaussian distribution $\mathcal{N}(\mu(x_{(1:t)}), \phi(x_{(1:t)}))$ is used. The module is learned with maximization of the variational lower bound:

$$L_{\theta,\phi,\psi}(x_{1:T}) = \sum_{t=1}^{T}[L_{perceptual\_recon} - \beta D_{KL}(q_{\phi}(z_t|x_{1:t})||p_{\psi}(z_t|x_{1:t-1}))]$$

2.3 Prediction Module

The Prediction Module infers a future state image $\widehat{x_t}$ from a past state x_{t-1} and a latent variable z_t learned from the Inference Module and the Prior Learning Module. Temporal information is conditioned and high dimensionality is reduced to the latent variable z. The inference process could infer temporal dynamics and how the future state would change based on the past states. The prediction module predicts the future states for b period of timestep where a period of timestep of past states are given. The prior is calculated in the prior learning module at every timestep.

2.4 Perceptual Reconstruction Loss

Loss function is built on beta variational autoencoder. A perceptual loss function is applied to improve limited ability of pixel-wise loss function and extract high level representations. A per-pixel loss function such as mean squared error (MSE) is vulnerable to slight deformation on predicted images which could damage the learning process. Using perceptual loss could improve this limitations. In perceptual reconstruction loss, pixel-wise reconstruction loss and perceptual reconstruction loss are combined. Pre-trained VGG16 network [23] with ImageNet dataset [24] is used as perceptual network to extract high level features and MSE is used as pixel-wise loss. The loss function of our model is described as:

$$L_{\theta,\phi,\psi}(x_{1:T}) = \sum_{t=1}^{T}[L_{t,p} + \beta D_{KL}(q_{\phi}(z_t|x_{1:t})||p_{\psi}(z_t|x_{1:t-1}))] \tag{1}$$

$$L_{t,p} = \sum_{t}\sum_{i} w_i L_p^i \tag{2}$$

$$L_p^i = \left\|\Phi(x)^i - \Phi(\widehat{x})^i\right\| \tag{3}$$

Equation 1 shows complete loss function. KL loss is calculated with the posterior $q_{\phi}(z_t \mid x_{(1:t)})$ to have close distribution to the prior $p_{\psi}(z_t \mid x_{(1:t-1)})$. Equation 2 shows perceptual reconstruction loss that perceptual loss of each layer i as L_p^i in pretrained perceptual network are aggregated for a given timestep t. Equation 3 shows detail of perceptual reconstruction loss. The difference between original image x and reconstructed image $\widehat{x}$ is calculated at each layer.

3 Experiments

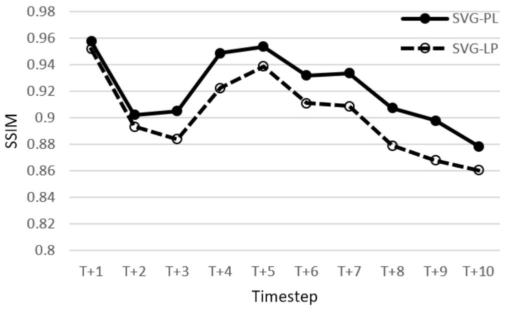

Figure 1: Average SSIM score for 10 frame prediction on BAIR Push Dataset. The SSIM score is high on SGV-PL compare to SVG-LP. Compounding error is shown as timestep increases.

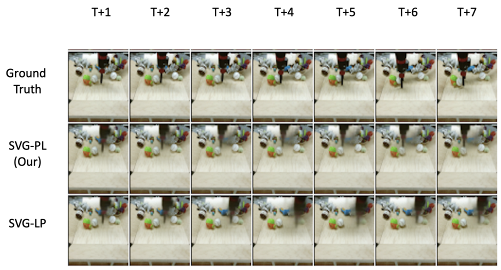

Figure 2: Qualitative comparion on BAIR Push Data. Seven consecutive future frames are predicted. Compared to SVG-LP, SVG-PL better captures temporal dynamics.

We evaluated our SVG-PL and baseline SVG-LP model on natural dataset BAIR Push [25]. Structural similarity index measure (SSIM) [26] is used for quantitative comparisons. Past 5 frames are used to train the model to predict 10 future frames. For each reconstructed 10 future frames, SSIM is measured compared to original future frames. For each prediction sequences, 100 sample sequences are generated and one with best SSIM score is selected. Since this metric provides limited comparison about perceptual qualities, qualitative comparison is also conducted.

Figure 1 shows average of SSIM results on stochastic moving BAIR Push data with SVG-PL and SVG-LP models. Results shows our model shows better SSIM results on predicting 10 future frames. This shows perceptual reconstruction loss could improve accuracy of future frame prediction by extracting high-level representations than pixel-wise loss function for reconstruction.

Figure 2 shows qualitative results of future generations of BAIR Push data for seven future frames where five past frames are given. The figure shows results of ground truth, proposed method SVG-PL, and baseline method SVG respectively. Result shows SVG-PL captures temporal dynamics of moving robot arm accurately compare to SVG-LP which shows limited representations. As we predict continuous future states, the result shows compounding error that prediction error gets accumulated.

4 Conclusion and Discussion

Our model proposes perceptual reconstruction loss to predict future states in stochastic approach to improve blurriness and uncertainty of the video prediction. Our model shows better results on future state prediction in both qualitative and quantitative metrics. The perceptual reconstruction loss uses both pixel-wise loss and perceptual loss from pretrained perceptual network using VGG16 with ImageNet dataset. Utilization of perceptual loss on video prediction improves reconstruction ability and result in clear predictions. Improvements in video prediction can help learning control policy in model-based reinforcement learning.

There are few limitations in this work and needs future works. For the quantitative comparison only Structural similarity index measure (SSIM) is used here. There exists metrics which are used to compare perceptual similarities such as DeepSim [27], Learned Perceptual Image Patch Similarity (LPIPS) [28] and Berkeley Adobe Perceptual Patch Similarity (BAPPS) [28]. Using these metrics could show clear improvements in perceptual feature extraction in video prediction.

The perceptual network is pretrained with ImageNet dataset. If the perceptual network is pretrained on the target dataset such as BAIR Push data, the model could extract better high-level representations on the target dataset. Other algorithms than VGG16 could be used to extract perceptual representations. Pixel-wise loss and perceptual loss are both used in this model. Instead of giving equal weight on both losses, different weight could improve prediction results.

Acknowledgments and Disclosure of Funding

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [21ZR1100, A Study of Hyper-Connected Thinking Internet Technology by autonomous connecting, controlling and evolving ways].

References

[1] Christoph G Keller and Dariu M Gavrila. Will the pedestrian cross? a study on pedestrian path prediction. *IEEE Transactions on Intelligent Transportation Systems*, 15(2):494–506, 2013.

[2] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, et al. Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE intelligent vehicles symposium (IV)*, pages 163–168. IEEE, 2011.

[3] Sungroh Yoon, Seil Lee, and Seong-Jun Oh. Stochastic modeling and concurrent simulation of the game of golf. *ETRI journal*, 31(6):809–811, 2009.

[4] Hoc Thai Nguyen and Nguyen Huu Thai. Temporal and spatial outlier detection in wireless sensor networks. *ETRI Journal*, 41(4):437–451, 2019.

[5] YeongHyeon Park, Won Seok Park, and Yeong Beom Kim. Anomaly detection in particulate matter sensor using hypothesis pruning generative adversarial network. *ETRI Journal*, 2020.

[6] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *International Conference on Machine Learning*, pages 1174–1183. PMLR, 2018.

[7] Sergiu Oprea, Pablo Martinez-Gonzalez, Alberto Garcia-Garcia, John Alejandro Castro-Vargas, Sergio Orts-Escolano, Jose Garcia-Rodriguez, and Antonis Argyros. A review on deep learning techniques for video prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[8] 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. PMLR, 2018.

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

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

[11] Tianfan Xue, Jiajun Wu, Katherine L Bouman, and William T Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. *arXiv preprint arXiv:1607.02586*, 2016.

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

[13] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033*, 2017.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[15] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[16] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017.

[17] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, pages 597–613. Springer, 2016.

[18] Fitsum A Reda, Guilin Liu, Kevin J Shih, Robert Kirby, Jon Barker, David Tarjan, Andrew Tao, and Bryan Catanzaro. Sdc-net: Video prediction using spatially-displaced convolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 718–733, 2018.

[19] Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1133–1141. IEEE, 2017.

[20] Gustav Grund Pihlgren, Fredrik Sandin, and Marcus Liwicki. Improving image autoencoder embeddings with perceptual loss. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[21] John R Hershey and Peder A Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–317. IEEE, 2007.

[22] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

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

[24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[25] Frederik Ebert, Chelsea Finn, Alex X Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. In *CoRL*, pages 344–356, 2017.

[26] Zhou Wang and Alan C Bovik. A universal image quality index. *IEEE signal processing letters*, 9(3):81–84, 2002.

[27] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29:658–666, 2016.

[28] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

A Appendix

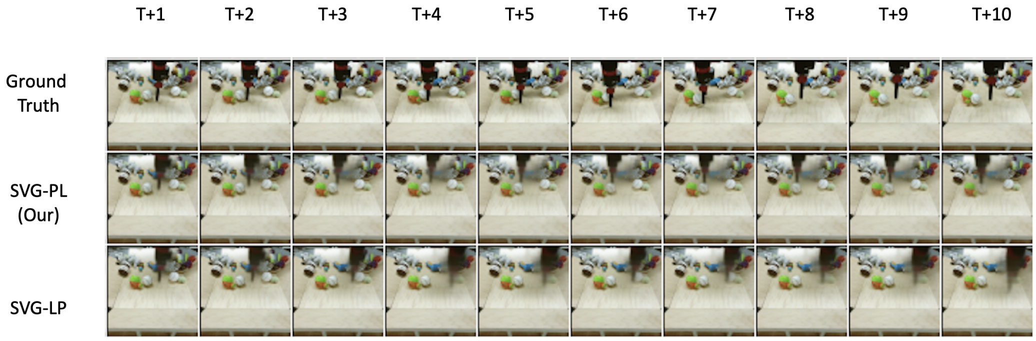

Figure 3: Qualitative comparison on BAIR Push Dataset. Ten consecutive future frames are predicted.

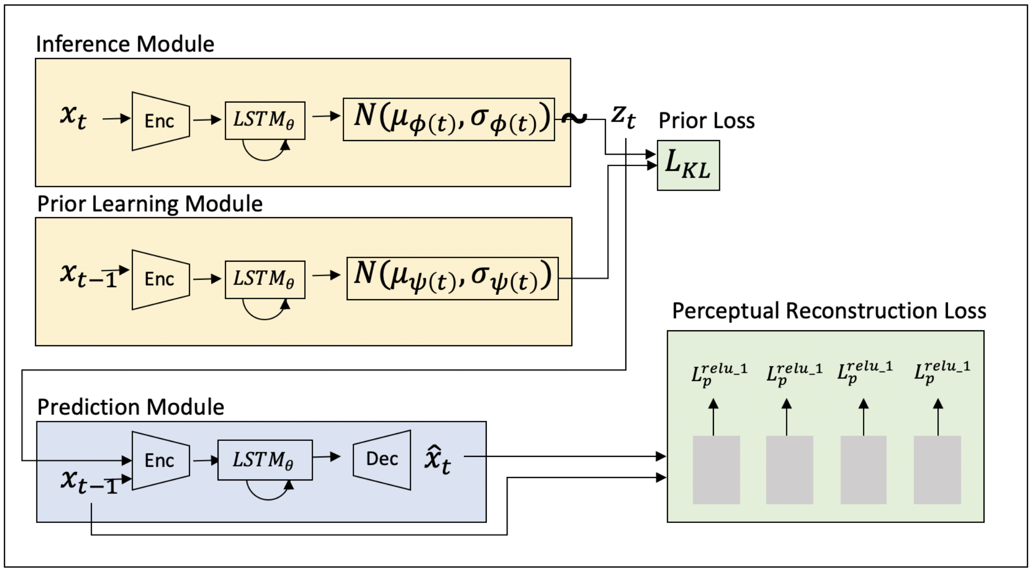

Figure 4: System architecture of the proposed algorithm, Stochastic Video Generation with Perceptual Loss (SVG-PL)

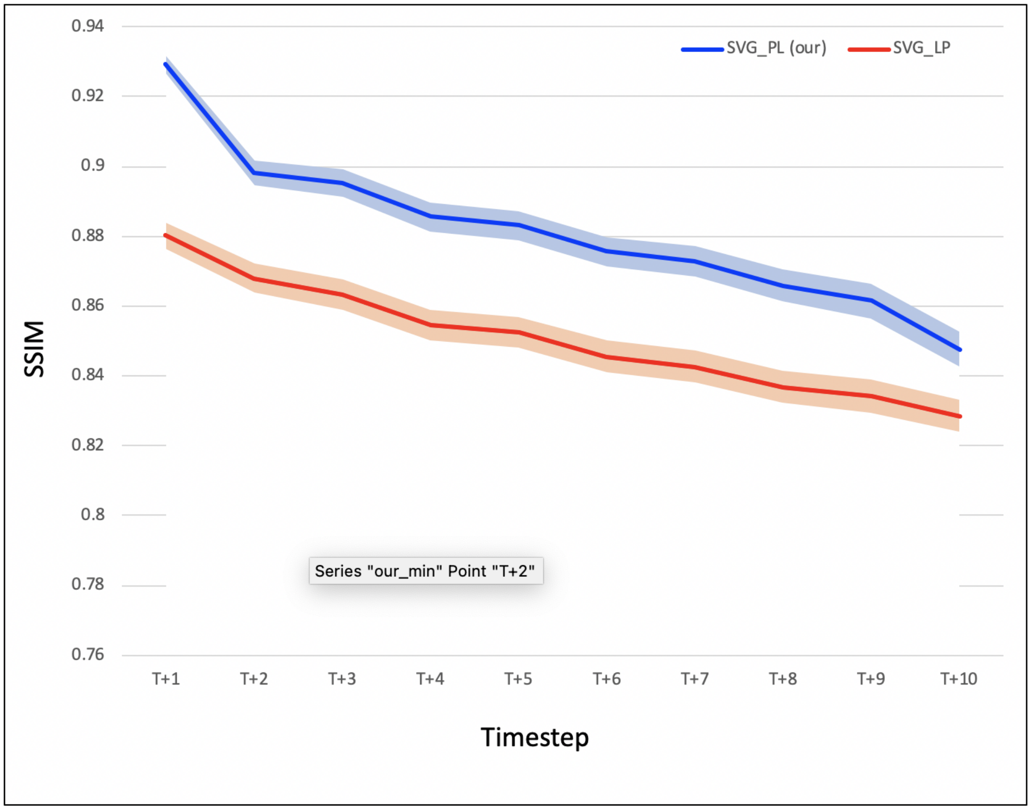

Figure 5: Average SSIM of 256 Test samples for each timestep of 10 consecutive future frames. Blue line shows result of our model SVG-PL and red line shows baseline model SVG-LP.

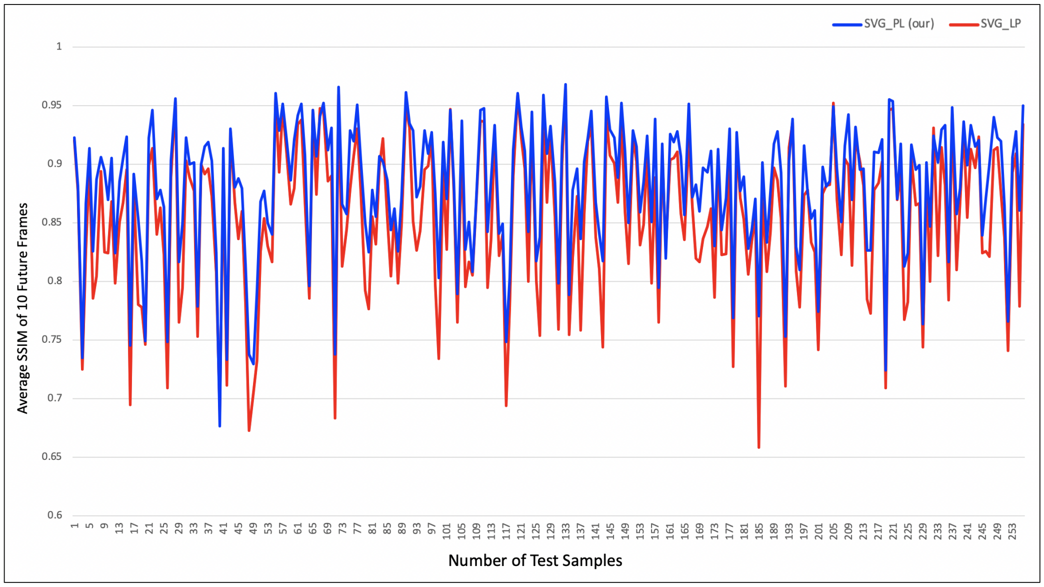

Figure 6: Agerage SSIM of 10 future frame predictions for each 256 test samples. Blue line shows result of our model SVG-PL and red line shows baseline model SVG-LP.