

Supplementary Material

Wide and Narrow: Video Prediction from Context and Motion

Jaehoon Cho¹
rehoon@yonsei.ac.kr

Jiyoung Lee¹
easy00@yonsei.ac.kr

Changjae Oh²
c.oh@qmul.ac.uk

Wonil Song¹
swonil92@yonsei.ac.kr

Kwanghoon Sohn¹
khsohn@yonsei.ac.kr

¹ School of Electrical and
Electronic Engineering
Yonsei University
Seoul, South Korea

² Centre for Intelligent Sensing
Queen Mary University of London
London, United Kingdom

1 Experiment details

1.1 Network Architecture

We present a detailed description of the architecture of our overall model in Table 1. We describe each layer with its input/output features' names, channel numbers, batch normalization, activation function, and down/upscale. For all convolutional layers, kernel size is set to 3×3 , except for filter generating networks. All max-pooling layers have both kernel size and stride set to 2×2 , resulting in the output features being downscaled with a factor of 2. Similarly, all Upconv. layers, implemented with transposed convolutional layers, have an upscaling factor of 2. $\mathbf{X} = (X_{t-(\delta-1)}, \dots, X_t)$, $\hat{\mathbf{Z}}$, and \hat{Y}_{t+1} denotes input video sequences, output features of GCPN convolved with the filter kernels from LFMN, and estimated future frame, respectively.

1.2 Comparison with state-of-the-art methods

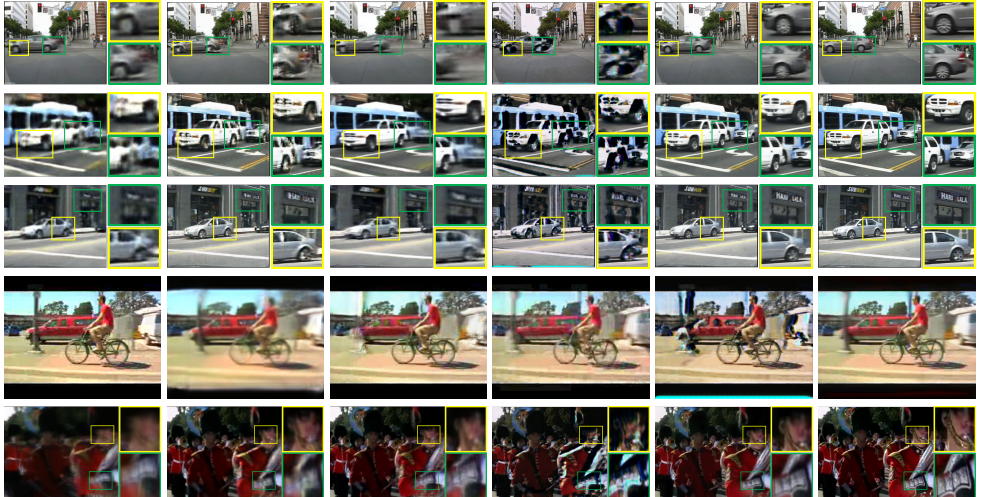
This section shows more visual examples of next frame and multi-frame prediction results on Caltech pedestrian [2] and UCF101 [3] datasets. All our models are trained for the next frame and multi-frame prediction given 4 input frames.

1.2.1 Next frame prediction.

Fig. 1 shows the next frame prediction with state-of-the-art methods.

Table 1: The architecture of our overall network. C_in, C_out denotes the number of channels of the input and output features, respectively. BNorm and Act. means batch normalization and activation function. $\{.,.\}$ denotes concatenation operator.

Encoder						
Layer/Output	C_in	C_out	BNorm	Act.	DownScale	Input
conv_E1a	12	64	o	ReLU	-	X
conv_E1b	64	64	o	ReLU	-	conv_E1a
pool_E1	64	64	-	-	o	conv_E1b
conv_E2a	64	64	o	ReLU	-	pool_E1
conv_E2b	64	64	o	ReLU	-	conv_E2a
pool_E2	64	64	-	-	o	conv_E2b
conv_E3a	64	64	o	ReLU	-	pool_E2
conv_E3b	64	64	o	ReLU	-	conv_E3a
pool_E3	64	64	-	-	o	conv_E3b
conv_E4a	64	64	o	ReLU	-	pool_E3
conv_E4b	64	64	-	-	-	conv_E4a
Decoder						
Layer/Output	C_in	C_out	BNorm	Act.	Upscale	Input
conv_D4a	128	64	o	ReLU	-	{conv_E4b, \hat{Z} }
conv_D4b	64	64	o	ReLU	-	conv_D4a
upconv_D3	64	64	o	ReLU	o	conv_D4b
conv_D3a	128	64	o	ReLU	-	{upconv_D3, conv_E3b}
conv_D3b	64	64	o	ReLU	-	Conv_D3a
upconv_D2	64	64	o	ReLU	o	Conv_D3b
conv_D2a	128	64	o	ReLU	-	{upconv_D2, conv_E2b}
conv_D2b	64	64	o	ReLU	-	conv_D2a
upconv_D1	64	64	o	ReLU	o	conv_D2b
conv_D1a	128	64	o	ReLU	-	{upconv_D1, conv_E1b}
conv_D1b	64	64	o	ReLU	-	conv_D1a
\hat{Y}_{t+1}	64	3	-	-	-	conv_D1b



(a) PredNet [1] (b) MCnet [1] (c) ContVP [1] (d) Liu [1] (e) Ours (f) Ground-truth
Figure 1: Qualitative comparisons of the predicted next frame on the Caltech Pedestrian and UCF101 dataset.

1.2.2 Multi-frame prediction.

Fig. 2, 3, and 4 compare multi-frame predictions of our best model (i.e our model trained with GCPN and LFMN) with state-of-the-art methods.

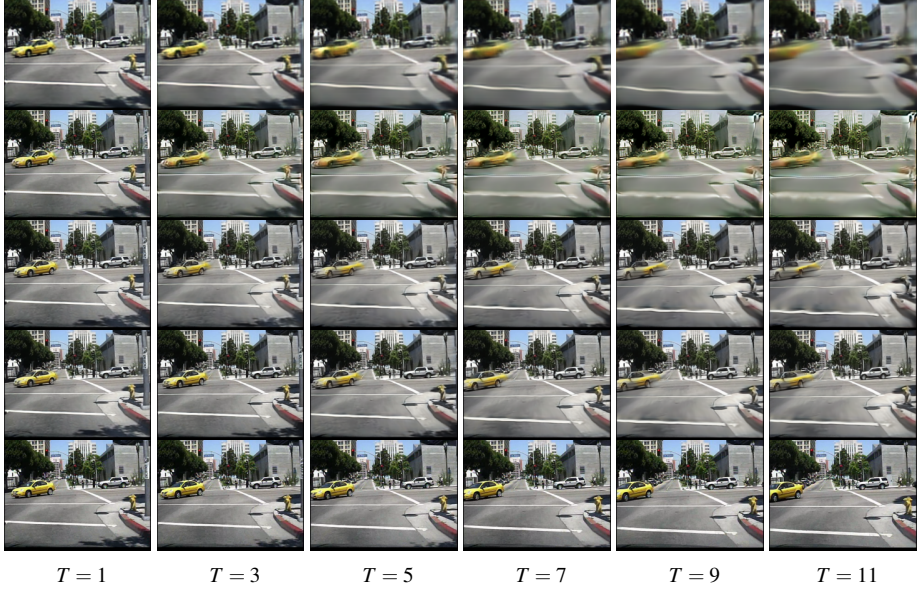


Figure 2: Multi-frame prediction on the Caltech pedestrian dataset. From top to bottom: results of PredNet [4], MCnet [6], ours with LFMN and GCPN, and ground-truth frame.

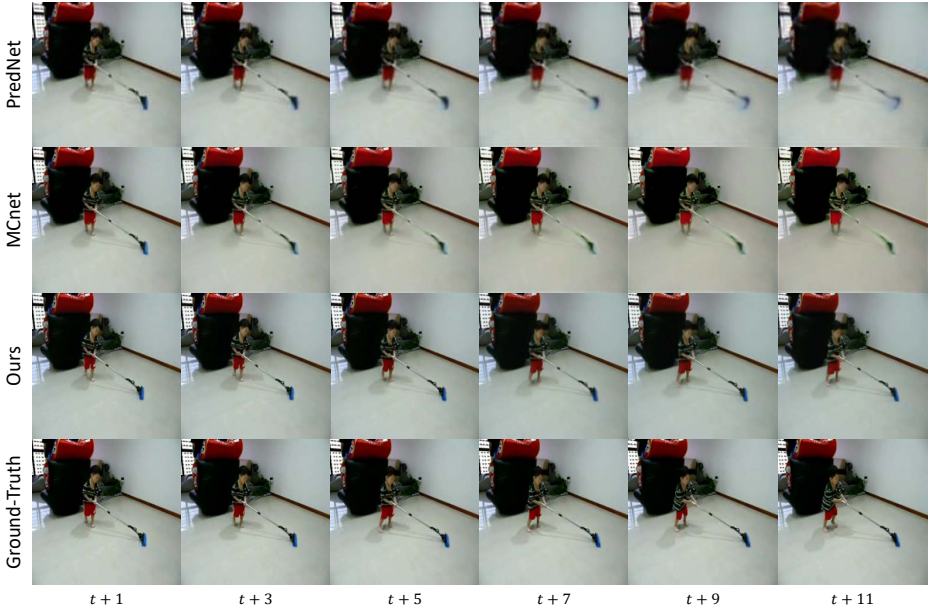


Figure 3: Multi-frame prediction on the UCF101 dataset. From top to bottom: results of PredNet [4], MCnet [6], ours with LFMN and GCPN, and ground-truth frame.

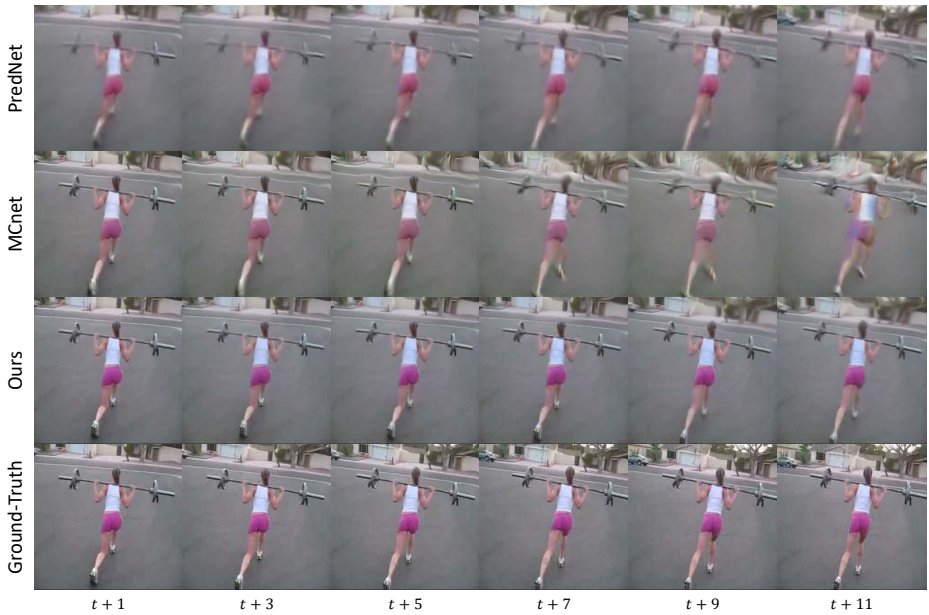


Figure 4: Multi-frame prediction on the UCF101 dataset. From top to bottom: results of PredNet [9], MCnet [6], ours with LFMN and GCPN, and ground-truth frame.

1.3 Analysis

Qualitative evaluation. Fig. 5 presents an example of qualitative prediction results using our method. Similar to previous works [6], we show the predicted frame and optical flow [2] extracted from the predicted frames. The predicted frame and extracted flow map show that our method is able to capture the global and motion information effectively.



Figure 5: Qualitative evaluation of the proposed method on Caltech pedestrian dataset.

References

- [1] W. Byeon, Q. Wang, R. K. Srivastava, and K. Petros. Contextvp: Fully context-aware video prediction. In *Proc. ECCV*, pages 753–769, 2018.
- [2] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Proc. CVPR*, 2009.
- [3] W. Liu, W. Luo, D. Lian, and S. Gao. Future frame prediction for anomaly detection—a new baseline. In *Proc. CVPR*, pages 6536–6545, 2018.
- [4] W. Lotter, G. Kreiman, and D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. In *ICLR*, 2017.
- [5] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. In *arXiv preprint arXiv:1212.0402*, 2012.
- [6] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee. Decomposing motion and content for natural video sequence prediction. In *ICLR*, 2017.
- [7] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In *ICCV*, 2013.