Towards Unified Multi-Excitation for Unsupervised Video Prediction

Junyan Wang[1], Likun Qin[23]*, Peng Zhang[4], Yang Long[4], Bingzhang Hu[5], Maurice Pagnucco[1], Shizheng Wang[2], Yang Song[1]

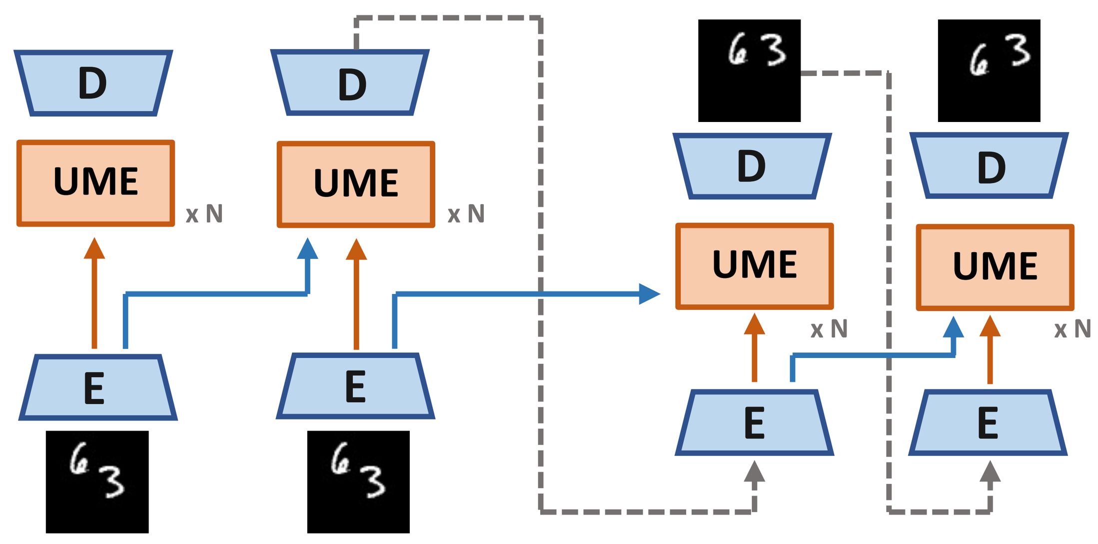
BMVC 2022

Abstract

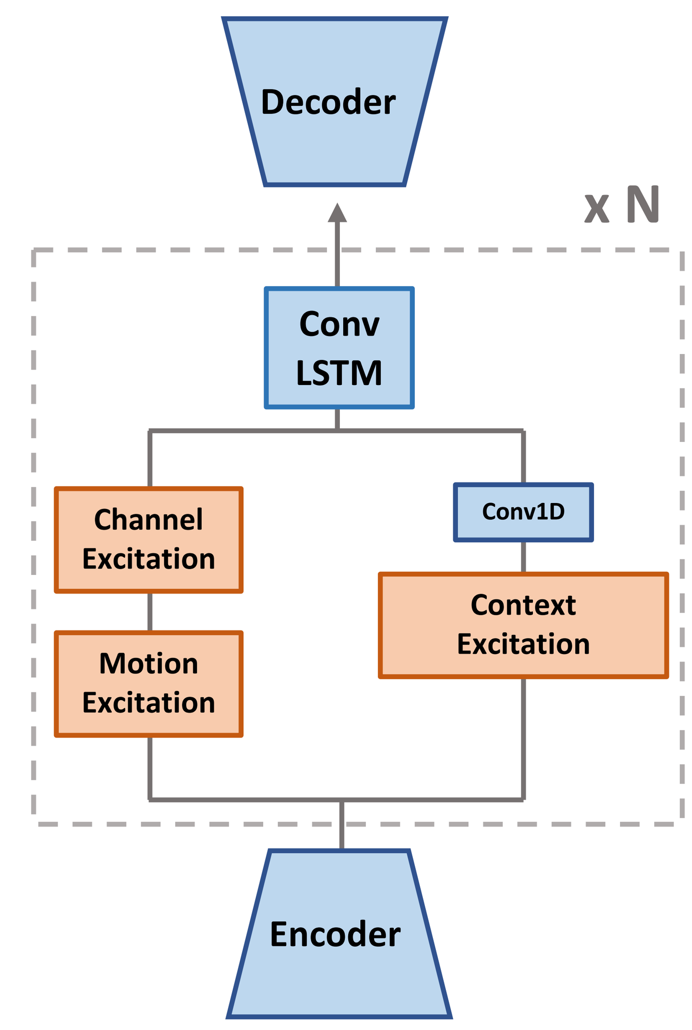
Unsupervised video prediction aims to forecast future frames conditioned on previous frames with the absence of semantic labels. Most existing methods have applied conventional recurrent neural networks, which focus on past memory, while few draw attention to highlight motion and context information. In this work, we propose a Unified Multi-Excitation (UME) block to enhance long-short-term memory, specifically applying an excitation mechanism to learn both channel-wise inter-dependencies and context correlations. Our contributions include: 1) introducing motion and channel excitation to enhance motion-sensitive channels of the features in the short term; and, 2) proposing an adaptive modeling scheme as context excitation inserted between (2+1)D convolution cells. The overall framework employs a multi-excitation block inserted into each ConvLSTM layer to aggregate the motion, channel, and context excitations. The framework achieves state-of-the-art performance on a variety of spatio-temporal predictive datasets including the Moving MNIST, Sea Surface Temperature, Traffic BJ and Human 3.6 datasets. Extensive ablation studies demonstrate the effectiveness of each component of the method.

Contribution

- To the best of our knowledge, we present the first multi-excitation mechanism for unsupervised video prediction, which can effectively enhance both motion and semantic information from previous frames.
- The motion and channel excitation is proposed to highlight motion representations in the channel dimension.
- Context excitation as an adaptive module is inserted between (2+1)D convolutional layers, integrating semantic information from previous frames.
- Quantitative and qualitative experiments on four datasets, Moving MNIST, Traffic BJ, Sea Surface Temperature and Human 3.6 , demonstrate that our proposed UME-Net achieves superior performance over the state-of-the-art methods.



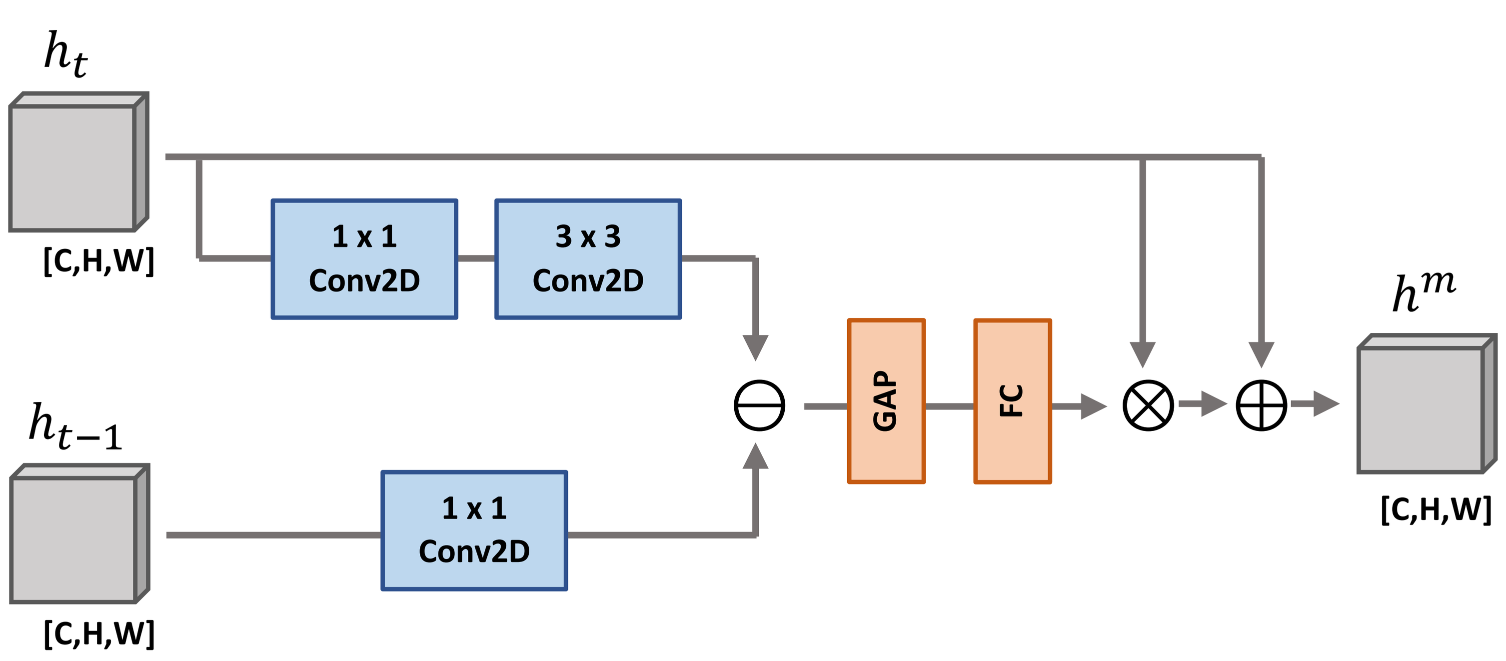
Overall UME-Net Framework



Unified Multi Excitation Block

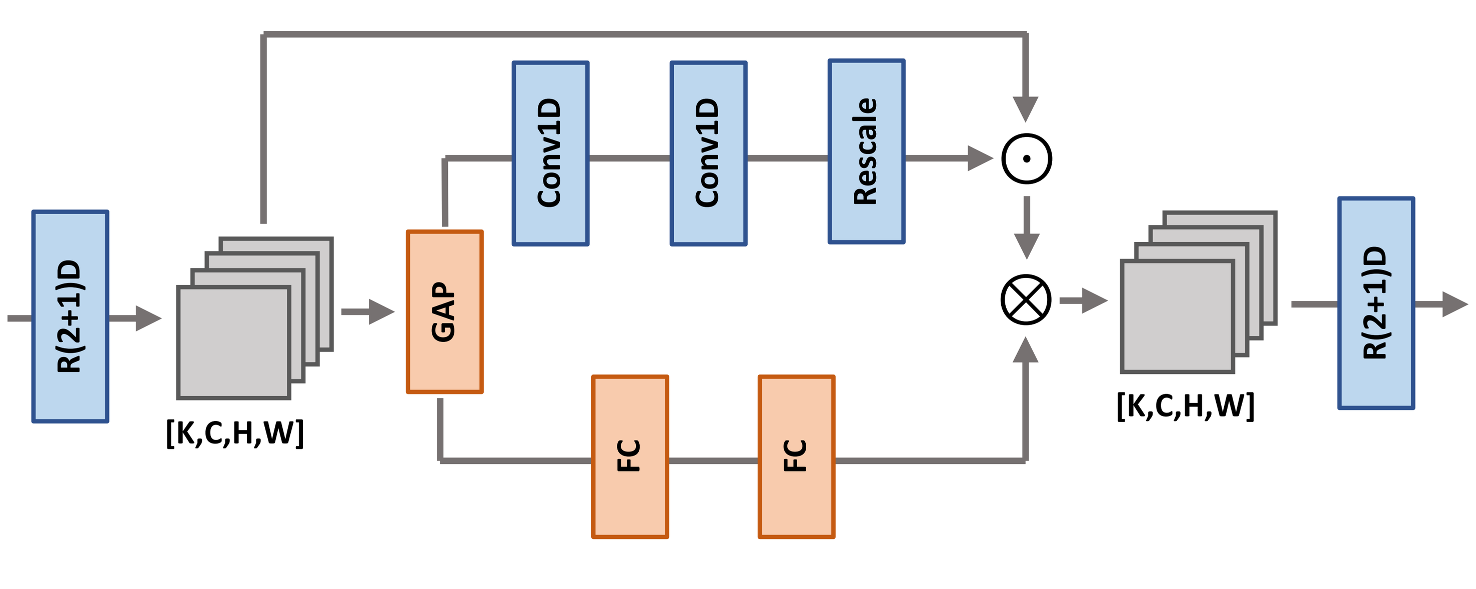
Method

We consider differences between consecutive frames to be the motion representation, and transform it to a series of weights for each channel. Then these motion generated weights are multiplied with the original feature map to enhance some motion-related channels.



Motion and Channel Excitation

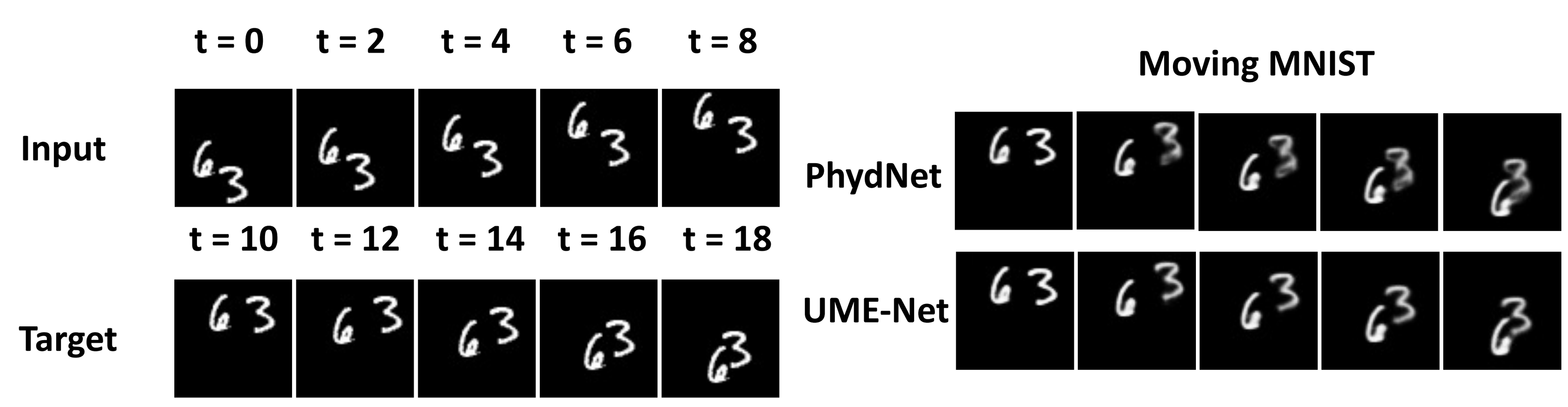
We consider that 3D CNNs can jointly capture the spatio-temporal features in a unified framework, whereas (2+1)D factorizing the 3D convolutional filters into separate spatial and temporal components can provide more efficient performance.



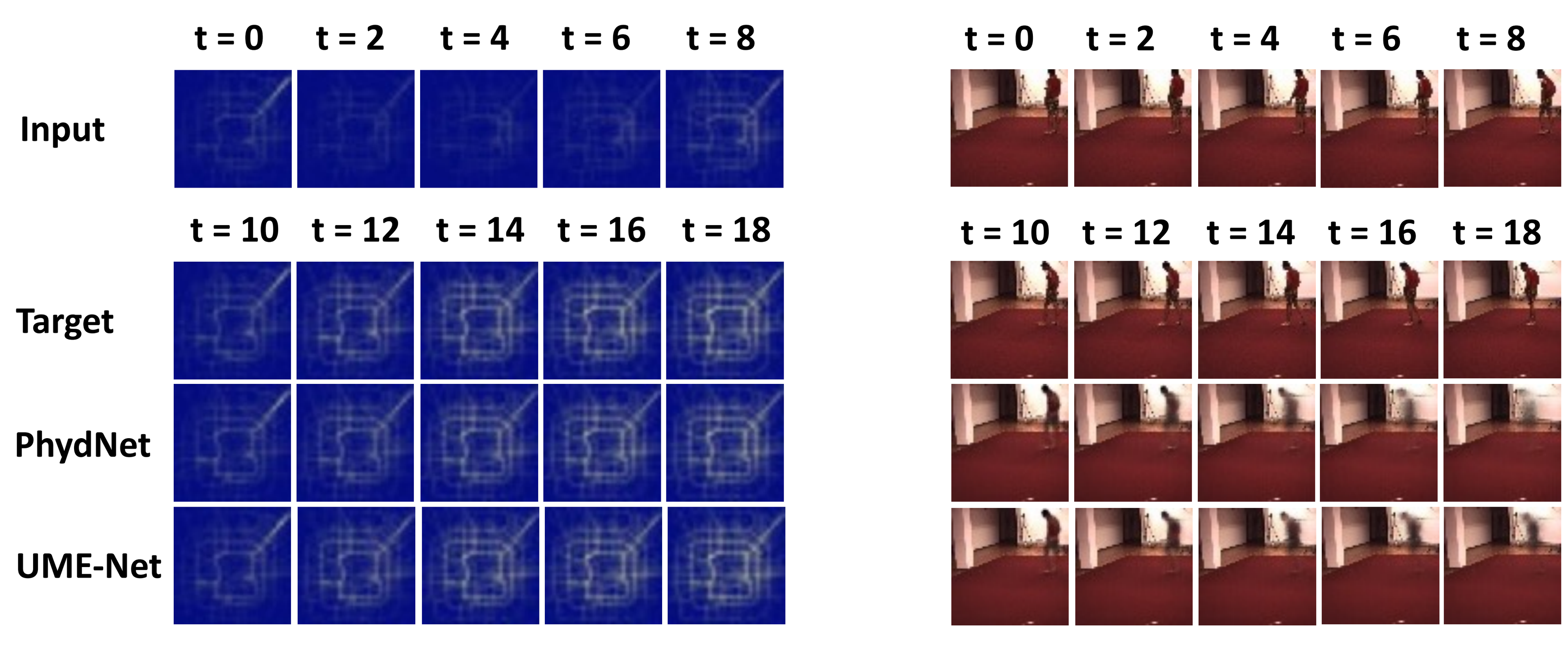
Context Excitation

Results

Method	Moving MNIST			Traffic BJ			SST			Human 3.6		
	MSE	PSNR	SSIM	MSE×10^2	PSNR	SSIM	MSE×10	PSNR	SSIM	MSE×10	PSNR	SSIM
ConvLSTM [17]	103.3	-	0.707	48.5	-	0.978	45.6	-	0.949	50.4	-	0.776
PredRNN [22]	56.8		- 0.867	46.4	-	0.971	41.9	-	0.955	48.4	-	-
Causal LSTM [23]	46.5	-	0.898	44.8	-	0.977	39.1	-	0.929	45.8	-	-
MIM [25]	44.2	-	0.910	42.9	-	0.971	42.1	-	0.955	42.9	-	-
E3D-LSTM [24]	41.3	-	0.920	43.2	-	**0.979**	34.7	-	0.969	46.4	-	-
PhydNet [3] *	24.19	23.36	0.9471	34.2	38.35	0.9761	31.9	**35.02**	0.9718	34.3	21.49	0.8321
UME-Net	**22.61**	**23.81**	**0.9523**	**32.22**	**38.7**	**0.9788**	**31.4**	34.79	**0.9742**	**33.5**	**21.71**	**0.8417**



Moving MNIST



Traffic BJ **Human 3.6**

1 School of Computer Science and Engineering, University of New South Wales, Sydney, Australia
2 Institute of Microelectronics, Chinese Academy of Science, Beijing, China
3 School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
4 Department of Computer Science, Durham University, Durham, UK
5 Institute of Physical Science, Chinese Academy of Science, Hefei, China