Towards Unified Multi-Excitation for Unsupervised Video Prediction

Junyan Wang[1*]
junyan.wang@unsw.edu.au
Likun Qin[2,3*†]
qinlk@mail.sysu.edu.cn
Peng Zhang[4]
peng.zhang@durham.ac.uk
Yang Long[4]
yang.long@ieee.org
Bingzhang Hu[5]
hubzh@aiofm.ac.cn
Maurice Pagnucco[1]
morri@cse.unsw.edu.au
Shizheng Wang[2]
shizheng.wang@foxmail.com
Yang Song[1]
yang.song1@unsw.edu.au

[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

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. Code and datasets are available at https://github.com/captaincj/UMENet.git.

* equal contribution
† Work done in Institute of Microelectronics, Chinese Academy of Science.

1 Introduction

With the tremendous growth of video materials uploaded to various online video platforms like YouTube, video prediction has received increasing attention in recent years. Future sequences are predicted based on the given previous frames, and can be used in various contexts, such as weather forecasting [17] and autonomous driving [9]. Unsupervised video prediction is different from conventional supervised video understanding tasks, like video recognition, which usually require a large amount of human labeling. By using the future frames as ground truth, it is possible to perform video prediction without additional manual labels and hence the task can be considered unsupervised.

The essential problem of this task is to explore better representations for videos [16] in an unsupervised way. A line of research focuses on improvements on recurrent neural network (RNN), especially Long Short Term Memory (LSTM). Starting from the initial patch-based method utilizing primitive RNN [16], some applied the LSTM Encoder-Decoder structure to directly predict pixel values [18]. Later, some methods focused on enhancing LSTM with 2D convolution [17] or 3D convolution [24] for better modeling spatio-temporal relations. Others focused on RNN itself, incorporating adversarial learning with LSTM Encoder-Decoder to solve the problem of blurry prediction [14] [15]. These enhanced LSTM layers have been adopted by subsequent studies as basic building blocks for their networks. However, these methods lack attention on motion information between frames. Recently, another line of studies is dedicated to disentangle motion and content information, such as using frame difference [19] and optical flow [12] to model motions, while visual content is mostly extracted by conventional 2D convolutions.

There are thus still many challenges in unsupervised video prediction. Firstly, although previous methods involving explicit modeling of motions have been proved successful, the best way to combine motion into a RNN framework remains an open issue: 1) using only difference of adjacent frames inevitably leads to loss of static information which would be useful when reconstructing future frames; and 2) using optical flow as supplementary motion representation brings significant computational cost and the quality of the extracted flow will influence the prediction. Secondly, simple RNN cells lack specific designs to highlight key semantic information, as conventional RNNs treat both long-term memory and short-term context in a unified way, where all information is integrated into the cell state through the same gating mechanism. This might cause the RNN cells to focus more on some information from distant frames, as the work of [23] points out, whereas the nearby frames should play a much more important role in future prediction especially when sudden changes happen.

On the other hand, the excitation mechanism [5] gains significant improvement on image-level tasks. It is designed to force the network to pay more attention on relevant features and suppress the irrelevant ones. To this end, the above two key challenges motivate us to design a **Unified Multi-Excitation** (UME) block consisting of motion, channel, and context excitation. Concretely, 1) motion and channel excitation transforms the difference between adjacent frames to a series of weights for each channel, and multiplies it with the original feature map to enhance motion-related channels; and, 2) context excitation aggregates the information from previous frames into a dynamic adaptive kernel in a convolutional manner to enhance semantic features. The overall **UME-Net** includes both motion and context excitation operations in parallel before each ConvLSTM layer to form our unified multi-excitation block. In summary, our contributions include:

- To the best of our knowledge, we present the first multi-excitation mechanism for un-

supervised 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 [18], Traffic BJ [27], Sea Surface Temperature [2] and Human 3.6 [6], demonstrate that our proposed UME-Net achieves superior performance over the state-of-the-art methods.

2 Related Work

Video Prediction. Deep learning methods have recently achieved excellent performance for video prediction [3, 7, 16, 18, 21, 23]. Given a series of previous frames, early works defined the video prediction task as generating the next frame. Ranzato *et al*. [16] firstly used RNNs to tackle this task, and Srivastava *et al*. [18] used a LSTM Encoder-Decoder framework to extend the one-frame prediction task to the prediction of a long sequence of frames. Meanwhile, some studies paid more attention to spatial and temporal modeling. Shi *et al*. [17] improved the original fully-connected LSTM layer by incorporating 2D convolution in their ConvLSTM layer, and Wang *et al*. [24] incorporated 3D convolution into LSTM. Recently, the unsupervised video prediction task has drawn increasing attention. For example, Wang *et al*. [23] proposed Causal LSTM with cascaded dual memories along with a Gradient Highway Unit, and that of Kalchbrenner *et al*. [7] adopts the Video Pixel Network to encode the time, space and color structure of video tensors as a four-dimensional dependency chain. Moreover, [3] proposed PhyCell in parallel with ConvLSTM cell to enhance its ability to capture motion information. However, there still lacks semantic information, which is typically considered as important in generative tasks. Our work focuses on unsupervised learning video prediction by extending the excitation mechanism.

Excitation Scheme in Video Understanding. The concept of excitation is recently put forward as a method for providing attention at the channel level [1, 5, 11, 13, 20, 26]. Excitation is designed to force the network to pay more attention on relevant features and suppress irrelevant ones, making the network more effective. Before the idea of excitation formally came out, Wang *et al*. [20] adopted a kind of mixed attention mechanism to focus on specific locations and channels, which can be viewed as the prototype of excitation. However, their attention module is relatively complicated and computationally demanding. Later on, Hu *et al*. [5] firstly introduced excitation in their Squeeze-and-Excitation (SE) network. The SE method recalibrates channel-wise features by modeling inter-dependency of channels in a simple yet effective way. Following SE, Woo *et al*. [26] further proposed spatial attention in addition to channel excitation in their Convolutional Block Attention Module (CBAM). Recently, Li *et al*. [11] extended the idea of excitation from the image classification to action recognition domain in their Motion Excitation (ME) block. Nevertheless, in the field of video prediction, excitation is rarely used to help generate future frames. The work of [13] proposed a temporal adaptive module (TAM) to generate video-specific temporal kernels based on its own feature map. In this work, we focus on exploring excitation mechanism in both motion and context aspects for generating future frames.

3 The Proposed Approach

Although there has been significant research investigating ConvLSTM-based methods in video prediction, learning spatio-temporal features still remains a challenging issue. As an extension of RNN, ConvLSTM learns spatial information by the convolutional mechanism and temporal information by the LSTM mechanism. However, RNN cells, such as LSTM, focus more on memorizing past states rather than exploring the pattern of changes. Compared to RNN-based approaches, temporal excitation can provide a new solution to learn spatio-temporal information in video prediction. To achieve this, we propose a unified multi-excitation block to enhance the spatio-temporal features and an end-to-end UME-Net for unsupervised video prediction.

3.1 Revisiting the Excitation Scheme

The excitation scheme was introduced in Squeeze-and-Excitation (SE) blocks [5], which is designed to add squeeze and excitation operations into each residual-like block to enhance informative channels and suppress non-relevant ones. Concretely, given input feature maps $H \in \mathbb{R}^{C\times H\times W}$, SE blocks first squeeze global spatial information into a channel *descriptor* $z \in \mathbb{R}^C$, such that the cth element of z is calculated by:

$$z_c = \mathcal{F}_{squeeze}(h_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} h_c(i,j)\ , \tag{1}$$

where $h_c(i,j)$ denotes the value at position (i,j) in channel c. To make use of the information aggregated in the squeeze operation, SE blocks employ a simple gating mechanism with a sigmoid activation σ to make the excitation function flexible and learn a non-mutually-exclusive relationship, as:

$$s = \mathcal{F}_{excitation}(z,W) = \sigma(W_2\delta(W_1 z))\ , \tag{2}$$

where $s \in \mathbb{R}^{c\times 1\times 1}$ denotes the acquired weights for each channel of H and δ refers to the ReLU function. Two dimensionality-reduction layers with reduction ratio γ are used to generate weights for each channel, $W_1 \in \mathbb{R}^{\frac{C}{\gamma}\times C}$ and $W_2 \in \mathbb{R}^{C\times\frac{C}{\gamma}}$. The final output of the block is obtained by rescaling as:

$$\widetilde{h_c} = \mathcal{F}_{scale}(h_c, s_c) = s_c h_c + h_c\ , \tag{3}$$

where $\mathcal{F}_{scale}$ denotes channel-wise multiplication. In this regard, SE blocks intrinsically introduce dynamics conditioned on the input, which can be regarded as a self-attention function on channels whose relationships are not confined to the local receptive field the convolutional filters are responsive to. Existing excitation-related models have performed well on image-level tasks, demonstrating that the excitation scheme is beneficial in learning image-level features. However in video-level tasks, especially in video prediction, data is complex and requires modeling of spatio-temporal information. To this end, we design a multi-excitation block to apply the excitation scheme to effectively predict future frames.

3.2 Unified Multi-excitation Block

The essential part of a UME block contains motion, channel and context excitations. The details of each excitation cell are shown in Figure 1.

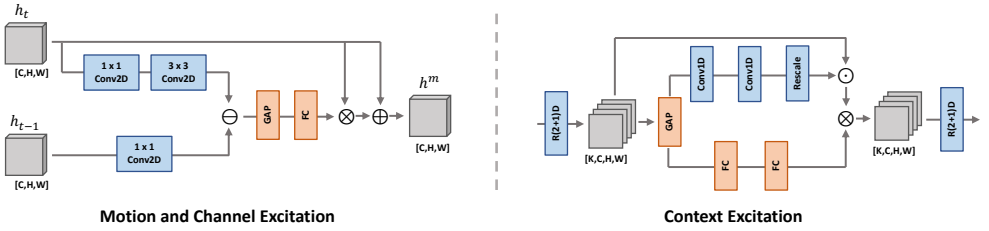

Figure 1: Illustration of motion and channel excitation, and context excitation.

3.2.1 Motion and Channel Excitation

Previous methods typically used difference of adjacent frames [19] or optical flow [12] as motion representations. However, simply computing differences will lose the static scene information and calculating optical flow is computationally expensive. Recent video understanding studies [11, 13] show that the excitation mechanism can be effective in enhancing temporal channel-wise information. Hence, we propose to combine motion excitation and channel excitation in an efficient way. Concretely, 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.

Specifically, our motion and channel excitation block receives feature maps of current frame and the previous frame as input. Given an input feature $h_t \in \mathbb{R}^{C\times H\times W}$, we first feed it into a Conv1D layer to reduce the number of channels instead of average pooling in Eq. 1, which is defined as:

$$m_t = Conv_{squeeze}(h_t),\ m_t \in \mathbb{R}^{\frac{C}{r}\times H\times W} \tag{4}$$

where m_t represents the channel-reduced feature and r denotes the reduction ratio. Instead of directly subtracting the feature map of the previous frame from that of the current frame, we apply a Conv2D operation to offset minor displacement between them before subtraction, as:

$$\hat{m}_t = Conv_{trans} * m_t - m_{t-1}, \tag{5}$$

where $\hat{m}_t \in \mathbb{R}^{\frac{C}{r}\times H\times W}$ denotes the desired motion feature. After that, we adopt global average pooling to summarize the spatial information and another Conv1D to expand the channel number back to c. Then, the attention weights A^m and final output h_t^m are defined as

$$\begin{aligned} A^m &= \sigma(Conv_{exp} * Pool_{avg}(\hat{m}_t)) - 0.5 \\ h_t^m &= h_t + h_t \odot A^m,\ h_t^m \in \mathbb{R}^{C\times H\times W} \end{aligned} \tag{6}$$

where σ denotes the sigmoid function and $\odot$ represents the channel-wise multiplication. Compared to the SE block, our proposed motion and channel excitation focuses on enhancing motion-sensitive spatio-temporal information, instead of static background information.

3.2.2 Context Excitation

As a generative task, the video prediction model also requires semantic information for generating future frames. Although ConvLSTM itself can model progression in both time and spatial dimensions, it lacks a bottom-up component which explicitly combines low-level

details to form high-level semantics. Besides, RNNs are generally too shallow to extract complex yet abstract representations. We argue that such deficiency in ConvLSTM hinders it from utilizing high-level semantics to guide future prediction. To this end, we propose context excitation to highlight meaningful representations from context features efficiently. We consider that 3D CNNs [4] 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. To deal with complex temporal variations in videos, we thus adopt an adaptive mechanism between (2+1)D cells, inspired by the work of [13].

Specifically, given an input feature $h_n \in \mathbb{R}^{N\times C\times H\times W}$ from a previous (2+1)D cell, where $N < T$ denotes the selected previous frames, we first employ a global spatial average pooling to squeeze the feature map as $Pool(h_t) \in \mathbb{R}^{N\times C}$. As shown in Figure 1, the local branch first reduces the number of channels from C to $\frac{C}{\beta}$ by a Conv1D. Then, the second Conv1D yields the importance weights A^c by a sigmoid function. Finally, the excitation is formulated as:

$$Z = Rescale(A^c) \odot h_t,\ A^c \in \mathbb{R}^{N\times C} \tag{7}$$

where $Z \in \mathbb{R}^{N\times C\times H\times W}$ and the $Rescale$ function rescale A^c to $\mathbb{R}^{N\times C\times H\times W}$ by replicating the spatial dimension. On the other hand, the global branch generates the dynamic kernel and aggregates semantic information in a convolutional manner as shown in Figure 1. Formally, for the c^{th} kernel, the adaptive kernel θ is learned as:

$$\theta = Softmax(\mathcal{F}(W_2, \sigma(\mathcal{F}(W_1, Pool(h_t)_c)))), \tag{8}$$

where $\theta \in \mathbb{R}^K$ denotes the adaptive kernel for the c^{th} channel, and $\mathcal{F}$ represents the fully-connected layer. The adaptive kernel is learned based on the squeezed feature map, and fully-connected layers can leverage long-term information. Thus, this kernel can aggregate temporal features guided by the global context. Combined with Z, the overall process can be defined as:

$$h^c_{n,c,j,i} = \theta \otimes Z = \sum_k \theta_{c,k} \cdot Z_{n+k,c,j,i}, \tag{9}$$

where $h^c \in \mathbb{R}^{N\times C\times H\times W}$ represents the output context feature map. By inserting context excitation between (2+1)D convolutional layers, it integrates information from previous frames to learn a more general representation of a specific moment. Compared to the SE block, the context excitation introduces an adaptive module with information aggregation, where the local information and global semantics are derived from given features. We can thus get high-level semantics gradually.

3.3 Overall Architecture Details

Based on the designed UME block, we propose a UME-Net for unsupervised video prediction, as shown in Figure 2, which contains encoder (E), multiple ConvLSTM layers and then a decoder (D). Before each ConvLSTM layer, we insert our two-branch UME block which has a motion and excitation block and a context excitation block in parallel. In this way, more critical information for prediction such as motion and context can be fused into the ConvLSTM memory to boost its performance.

Fusion Component. With motion and channel excitation and context excitation as two independent parallel branches, we introduce a simple yet efficient way to aggregate them, as

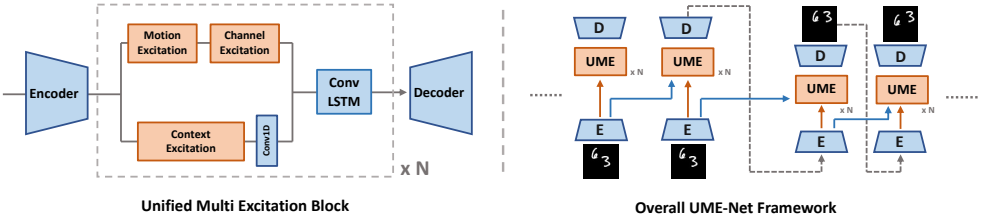

Figure 2: Illustration of the UME block and overall framework.

shown in Figure 2. As motion and channel excitation produces 3D tensors h^m in the shape of $C \times H \times W$, while context excitation outputs 4D tensors h^c in the shape of $N \times C \times H \times W$, an intuitive way is to reduce the dimension of the outputs of context excitation. Firstly, we divide $h^c \in \mathbb{R}^{N \times C \times H \times W}$ at time dimension to N smaller tensors and then we concatenate all of them at channel dimension to $\hat{h^c} \in \mathbb{R}^{NC \times H \times W}$. Then we use an additional Conv1D layer to reduce the channel number of $\hat{h^c}$. Then, we concatenate outputs from motion and channel excitation and the dimension-reduced context representations in the channel dimension to obtain the final output of our UME block.

4 Experiment

4.1 Experiment Setup

Datasets. There are four datasets on which we carried out our experiments. 1) **Moving MNIST** [18] contains 10000 frame sequences of length 20 each. The frames are of size 64*64 pixels and consist of two moving digits. Such sequences of frames are generated each time for training by choosing two digits randomly in the MNIST dataset [10] and placing these digits in a random location in the first frame with a random velocity. These digits bounce back each time they hit the wall. 2) **Traffic BJ** [27] dataset is composed of taxi GPS data and meteorology data in Beijing from five periods: Jul. 1st, 2003 to Oct. 30th, 2003, 1st Mar. 2014 to 30-th Jun. 2014, 1st Mar. 2015 to 30th Jun. 2015, 1st Nov. 2015 to 10th Apr. 2016. Data from the last four weeks serves as testing data while the rest are used for training. 3) **Sea Surface Temperature** (SST) [2] dataset consists of sea surface temperature data from 2006-12-28 to 2017-04-05 in the Atlantic ocean generated by the NEMO (Nucleus for European Modeling of the Ocean) engine. Patches of 64*64 pixels at the same location were extracted to form the sequences. 4) **Human 3.6** [6] is an action and human pose dataset with 3.6 million human poses categorized into 15 classes. There are altogether 11 professional actors on the videos, including 5 females and 6 males.

Evaluation Metrics. For fair comparison, we apply commonly used metrics used in video prediction methods: Peak Signal Noise Ratio (PSNR), Mean Squared Error (MSE) and Structure Similarity Index Measure (SSIM).

Implementation Details. In our network, we set the number l of UME blocks between the encoder and decoder to 3. In each UME block, both reduce radios γ and β of motion excitation and context excitation is set to 4, which is considered a good balance between model size and performance. The number of frames N considered in multi-frame aggregation block is 6, and the number of a Res(2+1)D layers is set to 4 in context excitaion. Another important hyper-parameter in each ConvLSTM layer is the number of channels, which we set

to 128 following [3]. Following the work of [3] and fair comparision, the encoder contains 5 convolutional layers and the decoder contains 5 deconvolutional layers. For training, we adopt mean square error as the loss function, ADAM optimizer [8] to train the model, and the learning rate is set to 0.001 for Moving MNIST and Human 3.6, and 0.0001 for SST and Traffic BJ. All experiments are performed on 4 Nvidia RTX 2080Ti GPUs with a batch size of 32.

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 []	103.3	-	0.707	48.5	-	0.978	45.6	-	0.949	50.4	-	0.776
PredRNN []	56.8		- 0.867	46.4	-	0.971	41.9	-	0.955	48.4	-	-
Causal LSTM []	46.5	-	0.898	44.8	-	0.977	39.1	-	0.929	45.8	-	-
MIM []	44.2	-	0.910	42.9	-	0.971	42.1	-	0.955	42.9	-	-
E3D-LSTM []	41.3	-	0.920	43.2	-	**0.979**	34.7	-	0.969	46.4	-	-
PhydNet [8] *	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**

Table 1: Quantitative results of UME-Net compared to baselines using various datasets. * indicates results obtained by reproducing their work, while other results are obtained from their papers. "-" indicates the results are not available for us.

4.2 Quantitative Result

We evaluate our method in comparison with current state-of-the-art unsupervised video prediction methods. As shown in Table 1, our method achieves state-of-the-art performance on all datesets. On Moving MNIST and Human 3.6 datasets, our UME-Net outperforms others with a large margin on all 4 metrics. Compared to other RNN-based methods, our proposed method achieves substantial improvements, which means simply employing an excitation mechanism can be more effective than network improvement. Meanwhile, it indicates that enhancing motion and semantic information from previous frames is important in the generative task. Compared to PhydNet, UME-Net gains better performance, which indicates that excitation can be more suitable than incorporating physical knowledge in unsupervised video prediction. Besides, unlike Moving MNIST, Traffic BJ and Human 3.6, the temperature maps exhibit large variety in shape even for adjacent frames. Therefore, it is challenging for our context excitation to learn more representative high-level semantics. Nevertheless, our UME-Net still achieves overall the best performance on SST by integrating the two-branch excitation design.

4.3 Ablation Study

We construct ablation study models on the Moving MNIST dataset, including: 1) "ConvLSTM" model without any any excitation operations; 2) "M & C" model with only motion and channel excitation; 3) "Context" model with only context excitation; 4) Both "R3D" and "R(2+1)D" models without any excitation operations on the context branch; 5) "Difference" model using only feature difference as motion excitation; 6) "GAP" model using global average pooling as the fusion component; 7) "DA" model using our proposed dimension aggregation fusion component. Results are shown in Tables 2 to 5.

Motion and Channel Excitation. Table 2 shows that, even though both "M & C" and "Context" can outperform the basic "ConvLSTM", "M & C" gains better performance. It indicates that motion information is more important in video prediction. When replacing the

Model	MSE	MAE	PSNR	SSIM
ConvLSTM	103.3	182.9	-	0.707
M & C	23.9	68.83	23.39	0.9479
Context	26.14	72.55	22.97	0.9427
UME-Net	**22.61**	**66.89**	**23.81**	**0.9523**

Table 2: Ablation study on each component.

Model	MSE	MAE	PSNR	SSIM
R3D	29.04	81.58	22.39	0.9342
R(2+1)D	27.3	76.93	22.75	0.9395
Context	**26.14**	**72.55**	**22.97**	**0.9427**

Table 3: Ablation study on motion and channel excitation.

Model	MSE	MAE	PSNR	SSIM
Difference	123.01	256.2	12.97	0.5116
M & C	**23.9**	**68.83**	**23.39**	**0.9479**

Table 4: Ablation study on context excitation.

Model	MSE	MAE	PSNR	SSIM
GAP	28.24	75.83	22.63	0.9399
DA	**22.61**	**66.89**	**23.81**	**0.9523**

Table 5: Ablation study on fusion style.

proposed excitation mechanism by simple feature difference, the performance of "Difference" drops significantly, which means that the excitation mechanism can be more effective when extracting motion information. The simple feature difference operation will not benefit motion information extraction and even results in loss of important information.

Context Excitation. Although the performance of "Context" is lower than "M & C", it still achieves large improvement over no excitation model, which indicates that semantic information plays an indispensable role in generating future frames. When there is no adaptive module, we find that "R(2+1)D" outperforms "R3D", which means factorizing the 3D convolutional filters into separate spatial and temporal components yields improvements. Moreover, the adaptive module can enhance semantic information as shown in Table 3.

Fusion Method. Comparing the results between "GAP" and "DA", we can see that the learned dimension aggregation outperforms average pooling. We consider that the pooling operation might lose information when fusing the various excited features, whereas our proposed fusion component is more effective in motion and context fusion.

4.4 Qualitative Result

We also provide qualitative results compared with PhydNet. As shown in Figure 3, from the input frames in Moving MNIST, we notice that the two digits move synchronously in the same direction with the same speed until one of them hits the boundary and digit '6' firstly bounces back in the final input frame. Such bounce-back is quickly learned by our method and utilized analogously to help predict the motion of the other digit '3'. In contrast, PhydNet had trouble predicting the future digit '3', which might be due to the lack of emphasis on context. As Traffic BJ is a small dataset, both PhydNet and UME-Net achieves good qualitative results, we can still see that the quality of our results are slightly better, demonstrating that the proposed excitation method can benefit future frames prediction. For Human 3.6 dataset, the content of our method is clearer than PhydNet, which indicates that the context excitation can enhance the semantic information. Also, the human motion is more correct that proves that the motion information has been enhanced by M&C excitation.

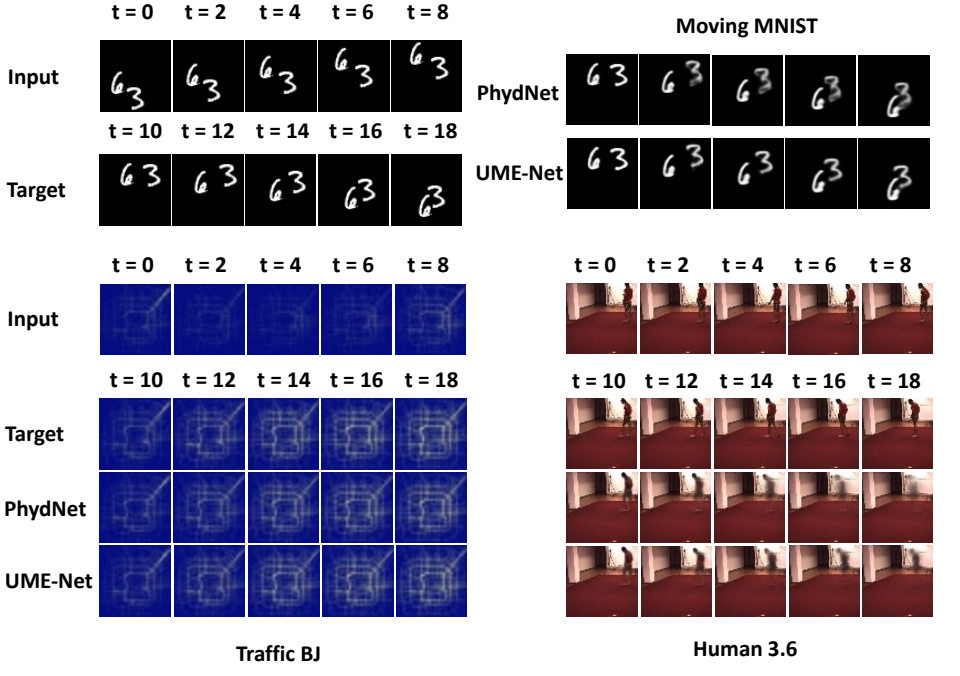

Figure 3: Qualitative results on three datasets: Moving MNIST, Traffic BJ and Human 3.6.

5 Conclusion

In this paper, we proposed a Unified Multi-Excitation block including a motion and channel excitation operation and a context excitation operation. The motion and channel excitation recalibrates channels at a feature map level, to enhance motion representations in the channel dimension. The context excitation inserted an adaptive module between (2+1)D convonlutional layers, to enhance semantic information also in the channel dimension. Moreover, the UME block fuses motion and semantic excited features in a learned dimension aggregation operation. Extensive experiments demonstrate that the proposed UME-Net can achieve state-of-the-art performance on four public datasets: Moving MNIST, Traffic BJ, Sea Surface Temperature and Human 3.6. In the future, we will focus on predicting high-quality and more complicated video frames.

6 Acknowledgments

This work was supported by SunwayAI computing platform (SXHZ202103).

References

[1] Yang Bai, Junyan Wang, Yang Long, Bingzhang Hu, Yang Song, Maurice Pagnucco, and Yu Guan. Discriminative latent semantic graph for video captioning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3556–3564, 2021.

[2] Emmanuel De Bézenac, Arthur Pajot, and Patrick Gallinari. Deep learning for physical processes: Incorporating prior scientific knowledge. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12), 2019.

[3] Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11474–11484, 2020.

[4] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In *CVPR*, pages 6546–6555, 2018.

[5] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[6] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 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, 2014. doi: 10.1109/TPAMI.2013.248.

[7] Nal Kalchbrenner, Aäron van den Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video pixel networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, pages 1771–1779, 06–11 Aug 2017.

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[9] Yong-Hoon Kwon and Min-Gyu Park. Predicting future frames using retrospective cycle GAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1811–1820, 2019.

[10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[11] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. TEA: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 909–918, 2020.

[12] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P. Xing. Dual motion gan for future-flow embedded video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1762–1770, 2017. doi: 10.1109/ICCV.2017.194.

[13] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu. TAM: Temporal adaptive module for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13708–13718, October 2021.

[14] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[15] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *International Conference on Learning Representations*, 2016.

[16] MarcAurelio Ranzato, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*, 2014.

[17] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wangchun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, page 802–810, 2015.

[18] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 843–852, Lille, France, 07–09 Jul 2015.

[19] 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.

[20] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

[21] Junyan Wang, Bingzhang Hu, Yang Long, and Yu Guan. Order matters: Shuffling sequence generation for video prediction. In *Proceedings of British Machine Vision Conference*, pages 275.1–275.14, 2019.

[22] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal LSTMs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 879–888, 2017.

[23] 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.

[24] Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li Fei-Fei. Eidetic 3D LSTM: A model for video prediction and beyond. In *International Conference on Learning Representations*, 2018.

[25] Yunbo Wang, Jianjin Zhang, Hongyu Zhu, Mingsheng Long, Jianmin Wang, and Philip S. Yu. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[26] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[27] Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 1655–1661. AAAI Press, 2017.