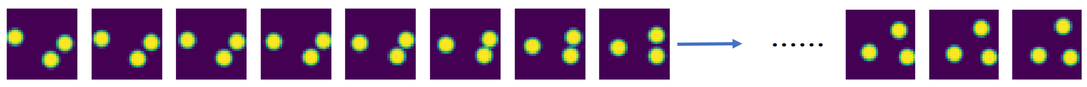
A CNN-based Video Prediction Method

Simon Fraser University

Problem

Given some frames of a video, predict the movement



Motivation

- Traditional CNN model excels in feature extraction, but lacks the ability of capturing movement information
- Recurrent based models (RNN, LSTM, etc.) naturally generate sequences, but are slow to train
- Attention mechanism performs well on selecting effective information in input sequential data

Contribution

We build two specific types of layers to help CNN better learn from sequence pictures

Movement tendency layer

- Tendency is computed by the difference of the latter and former layer
- Feed tendency layer into CNN, extracting the general direction of movement

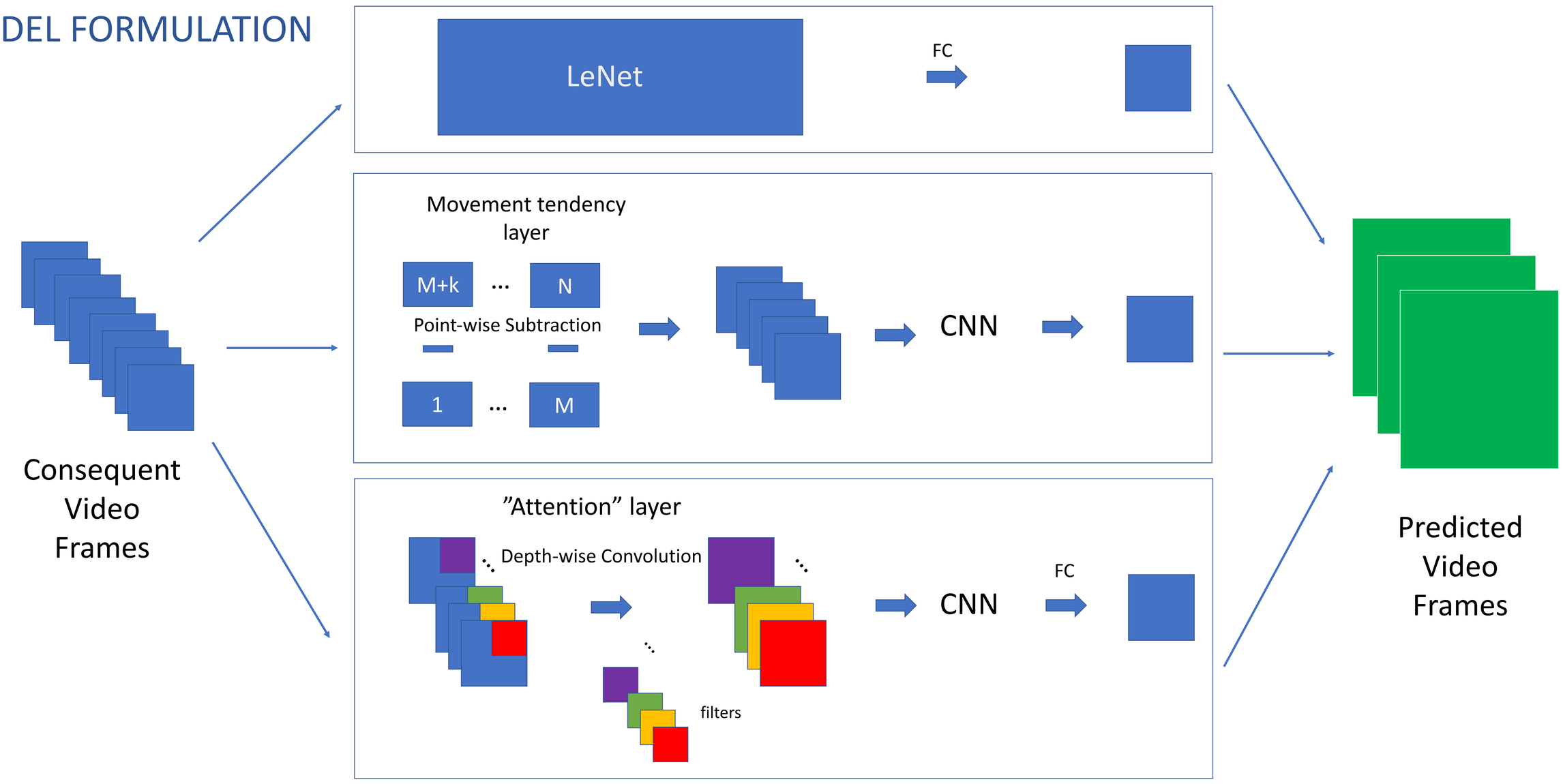
"Attention" layers

- Adopt n different filters for n input frames, and implement depth-wise convolution
- Each filter only works on specific frame
- Filters are trained to be equivalent to the "attention" weight for each frame

High quality MSE:

- Predicted image is meaningless when MSE error is above certain level
- Some extremely wrong input shouldn't be considered

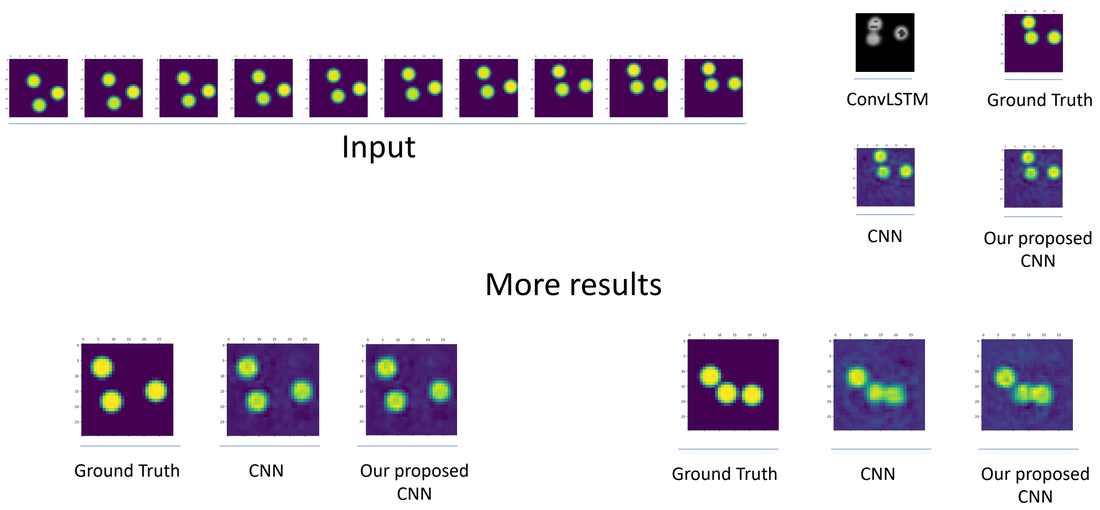
MODEL FORMULATION



EXPERIMENTS

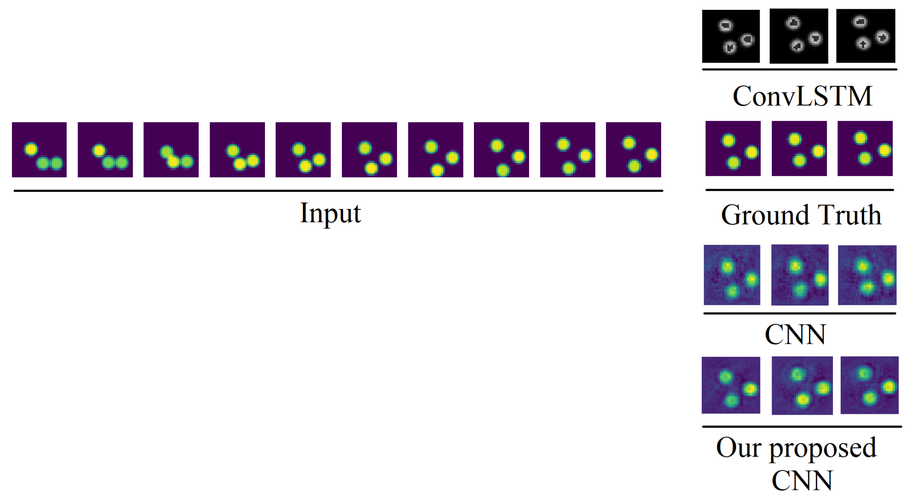
Predict 1 frame, given 10 frames as input

	CNN-only	Attention-Error Enhanced CNN	convLSTM[1]	Last Frame As Input
MSE loss	0.0061	**0.0059**	0.0078	0.0147

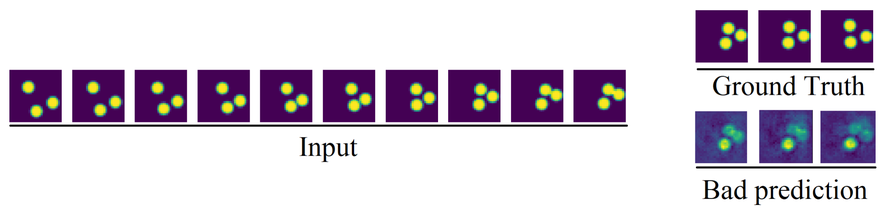


Predict 3 frames, given 10 frames as input

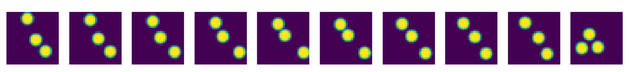
	CNN-only	Attention-Error Enhanced CNN	convLSTM
MSE loss	0.0247	0.0209	**0.0137**
High quality MSE*	0.0123	**0.0112**	/



Bad prediction happens after collision



*Why High Quality MSE?

Input:  MSE loss = 0.040

The input data is clearly wrong. We should not include such example in the computation of MSE loss.

[1]Xingjian, S. H. I., et al. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting." *Advances in neural information processing systems*. 2015.