Object Detection and Tracking via Deep Neural Networks in Complicated Environments Using Dynamic Template Module

Yiming Lin Yilin Liu Haotian Fang Chengrui Zhang Supervisor: Kaizhu Huang

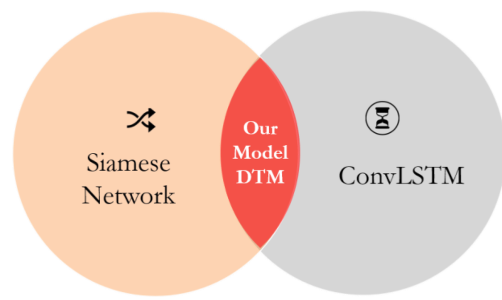
SURF Summer Undergraduate Research Fellowship

ABSTRACT

Among recent well-performed Siamese Network-based tracking algorithms, SiamMask reaches a precise object appearance identification by introducing the Mask Segmentation with tandem inputs. However, most existing SiamMask algorithms did not exploit sufficiently spatial-temporal information. In this research, we focus on building a dynamic modular based on ConvLSTM where a dynamic template is constructed following the exemplar frame by frame. Our model improves the robustness by establishing a dynamic template, while SiamMask merely provides an unalterable one. Also, the temporal features are efficiently extracted which are being neglected in SiamMask Algorithm.

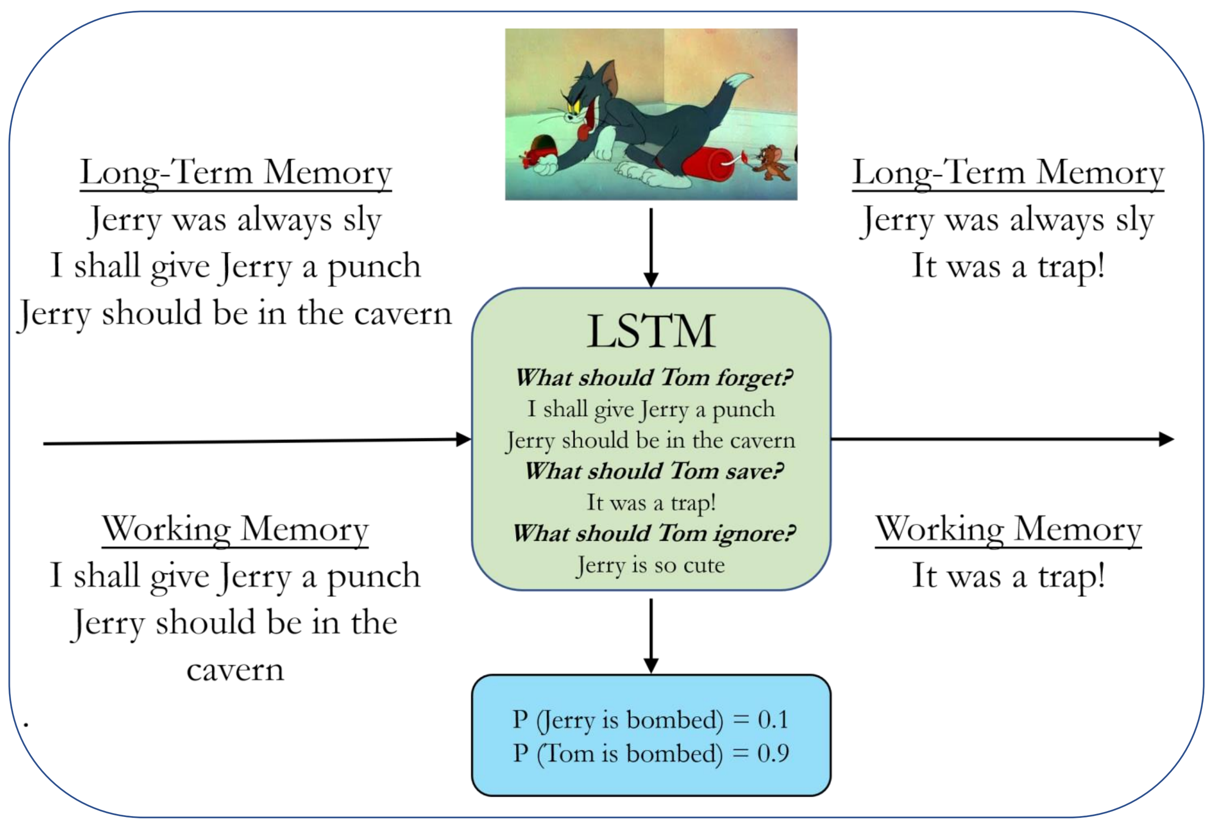
INTRODUCTION

Siamese Network architecture has exceptional speed and accuracy by putting the object tracking problem transformed into a similarity comparison. Siamese Mask algorithm uses the Video Object Segmentation (VOS) method to determine object appearance more precisely, to obtain better bounding box and classification results.

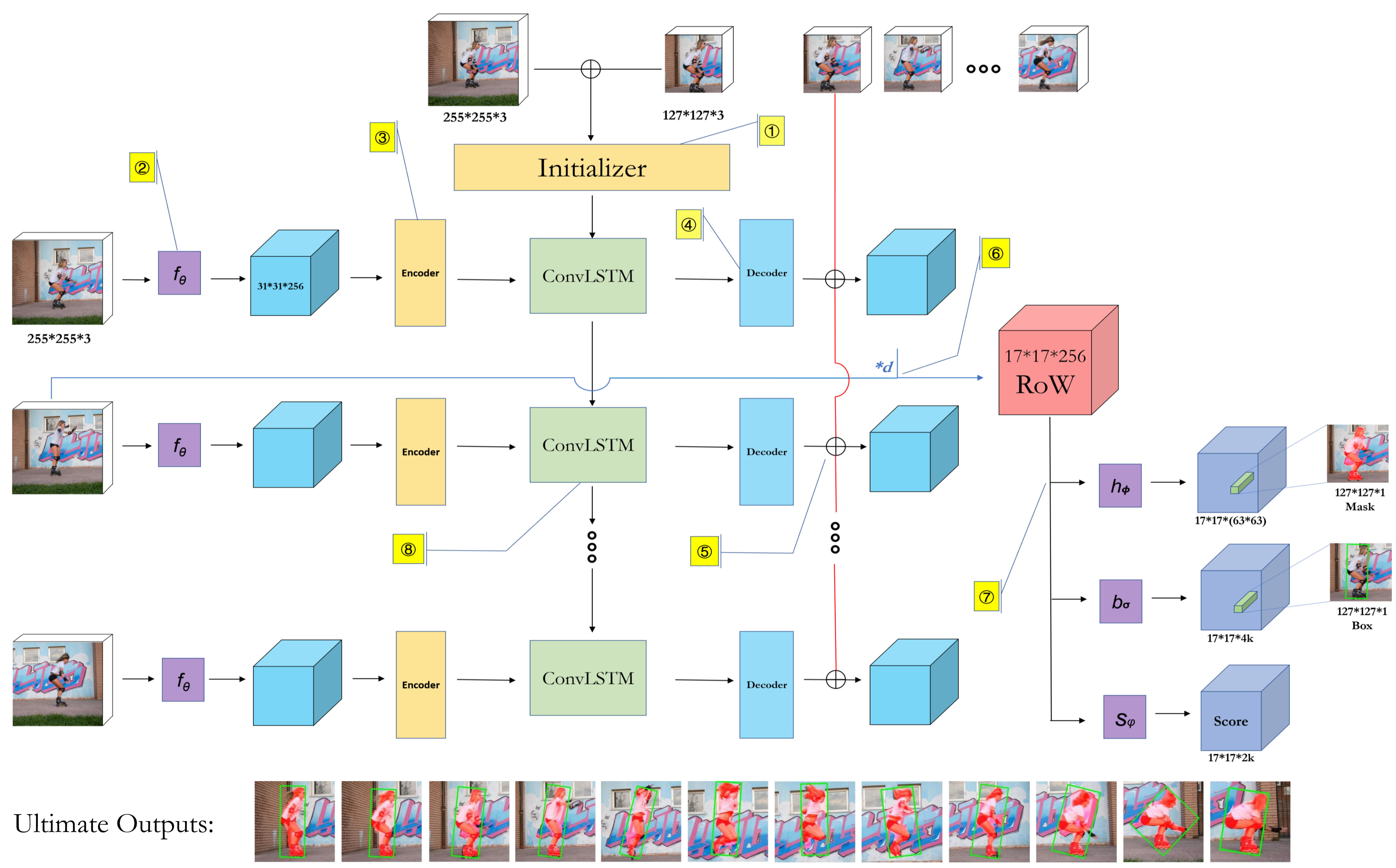


Different from LSTM, ConvLSTM exchanges internal matrix multiplications with convolution operations. ConvLSTM is preferred while handling images sequences because the inputs will remain its dimensions instead of being flattened.

1. We propose a new network architecture named Dynamic Template Module (DTM), which builds a dynamic template to enhance the robustness of the model. DTM includes a dynamic template construction module based on the ConvLSTM to use for constructing dynamic templates using search map of each frame.
2. Our model architecture uses ConvLSTM to construct dynamic template construction module, which fully utilizes timing information while retaining computation capacity that can run in real-time.



OUR MODEL



① $c_1, h_1 = Initializer(x_1, u_1)$

② $x_t = f_\theta(X_t), x_1 \in R^{W_z' \times H_z' \times C_z}$

③ $\hat{x}_t = Encoder(x_t)$

④ $\hat{z}_t = Decoder(h_t)$

⑤ $z_t = \gamma_1 \cdot \hat{z}_t + \gamma_2 \cdot z_{t-1}$

⑥ $g(z_t, x_t) = z_t \star x_t$

⑦ $Score = f_{score}\left(g_\theta(x_{k'}, z_k)\right)$

$Score \in [0,1]^{W_{out} \times H_{out} \times (2 \times Anchor)}$

$Bbox = f_{bbox}\left(g_\theta(x_{k'}, z_k)\right)$

$Bbox \in R^{W_{out} \times H_{out} \times (4 \times Anchor)}$

$Mask = f_{mask}\left(g_\theta(x_{k'}, z_k)\right)$

$Mask \in [0,1]^{W_{out} \times H_{out} \times (64 \times 64)}$

⑧ $c_t, h_t = ConvLSTM\left(\hat{x}_t\right)$

EXPERIMENTS

	Template Output Size	Search Output Size	Search+ Mask Size	Details		
Conv1	61x61	125x125	125x125	7x7, 64, stride2		
Conv2_x	31x31	63x63	63x63	3x3 maxpool, stride2		
Conv3_x	15x15	31x31	31x31	[1x1, 128		
Conv4_x	15x15	31x31	31x31	[1x1, 256		
Adjust	15x15	31x31	31x31	1x1, 256		
head_1			25x25	7x7, 256		
head_2			21x21	5x5, 256		
head_3			15x15	7x7, 256		
Convlstm_1			15x15	3x3, 128		
Convlstm_2			15x15	3x3, 128		
Decoder			15x15	1x1, 256		
Add	Alpha * template		(1-alpha) * template			
xcorr	17x17			Depth-wise		
Block				Score	Box	Mask
Conv5				1x1, 256	1x1, 256	1x1, 256
Conv6				1x1, 2k	1x1, 4k	1x1,(63x63)

f_θ -- ResNet50

Initializer

Encoder

ConvLSTM

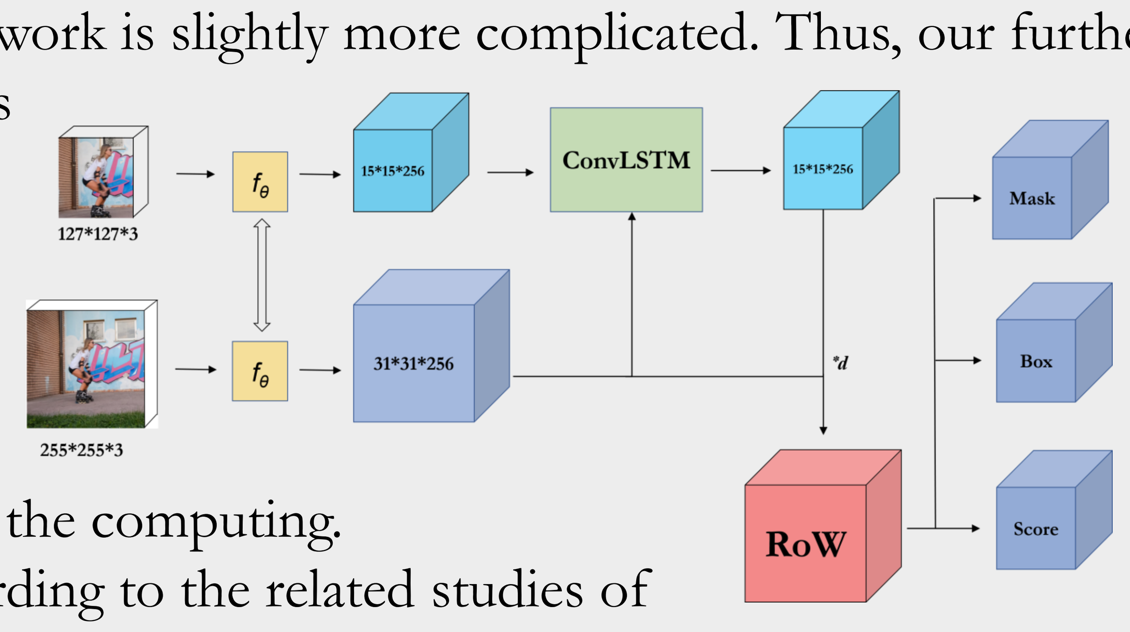
Concatenate

**d -- Correlation*

Decoder

FURTHER WORK

DTM model improves the robustness of the model by taking advantage of timing sequence under the condition of small computation. However, the question raised by this study is that the initializer is challenging to train and converge, and encoder is encoded on size, the decoder is decoded on the dimension, which does not match each other in the model. Besides, the network is slightly more complicated. Thus, our further study could be conducted as follows in the future:



Using the template to be the initialized state in Conv-LSTM can save training time, cropping the search template can also reduce the computing.

We try to predict the template according to the related studies of feature video prediction, and the overall model can be more sensible and concise.