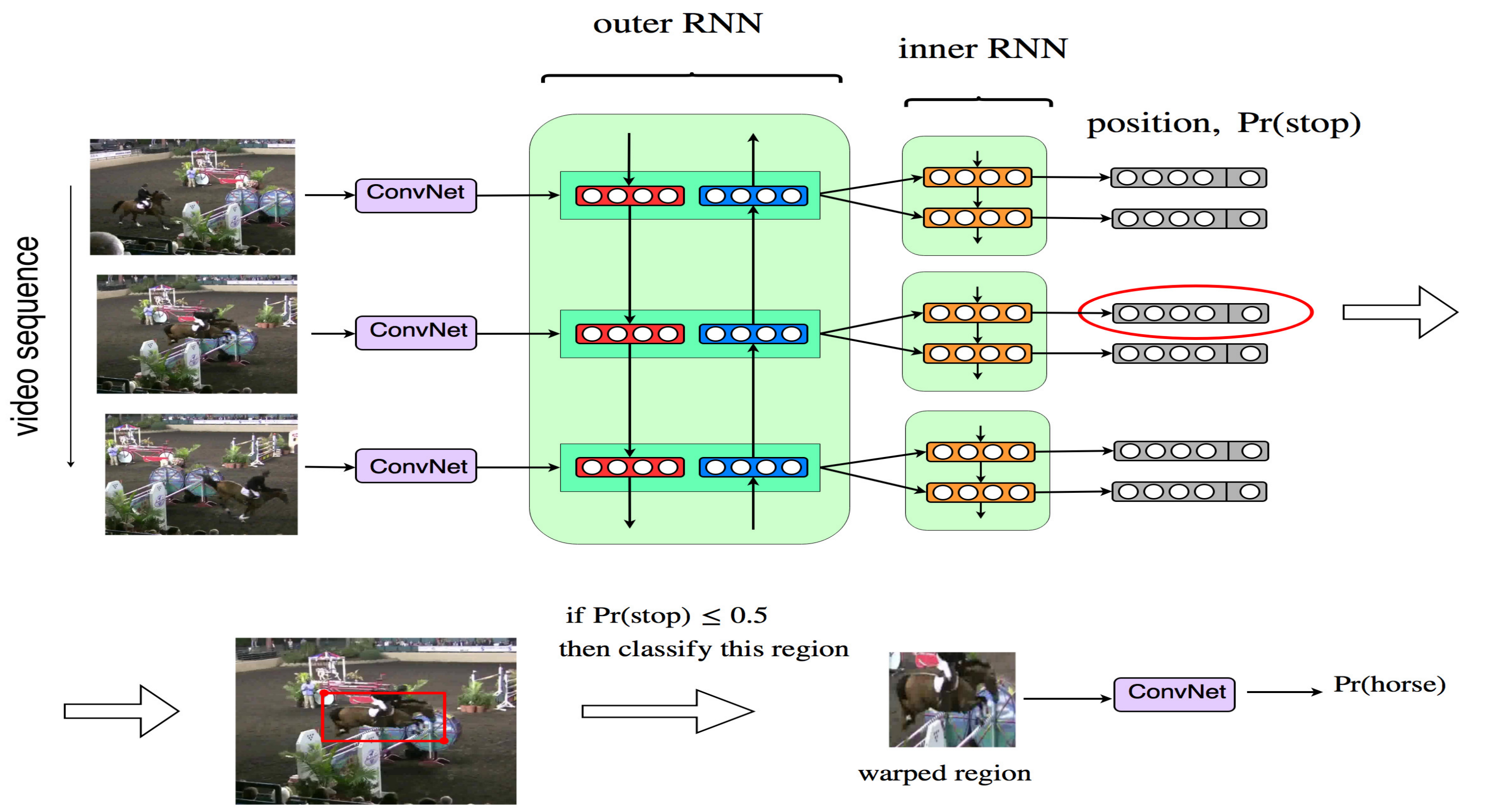
Object Detection from Video with Nested Recurrent Neural Networks

Noriki Nishida, Jan Zdenek, Hideki Nakayama
Machine Perception Group, The University of Tokyo, {nishida, jan, nakayama}@nlab.ci.i.u-tokyo.ac.jp

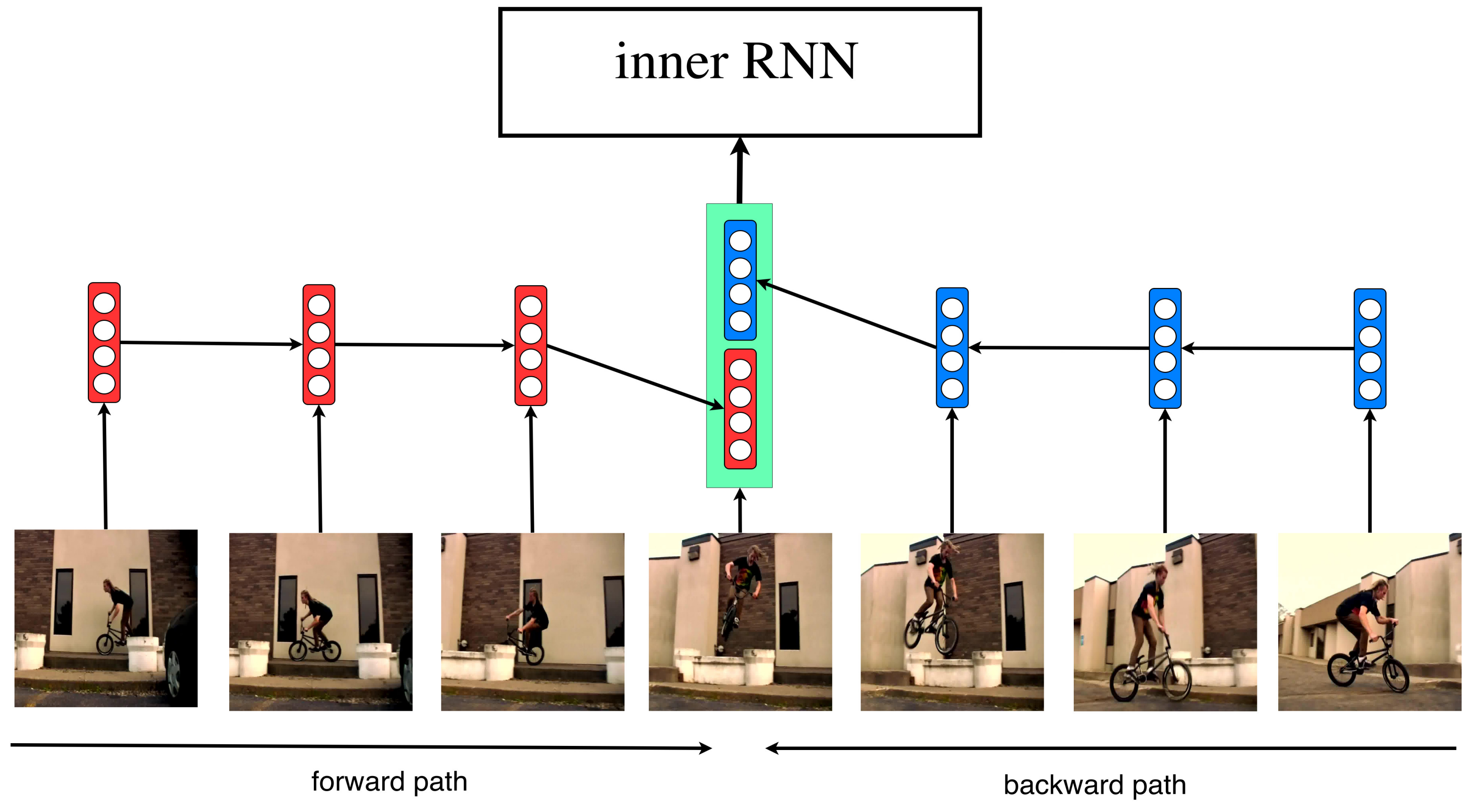
Summary

- Our goal is to propose a completely data-driven model for object detection from video.
- We develop a neural network model with nested recurrent structure that makes it possible to consider wide-range context.
- Our Nested Recurrent Neural Network consists of:
  - a bidirectional LSTM ("outer" RNN) for extracting temporal dynamics of objects from surrounding frames, and
  - an unidirectional LSTM ("inner" RNN) for predicting multiple bounding boxes sequentially in every frame.

Complete Model

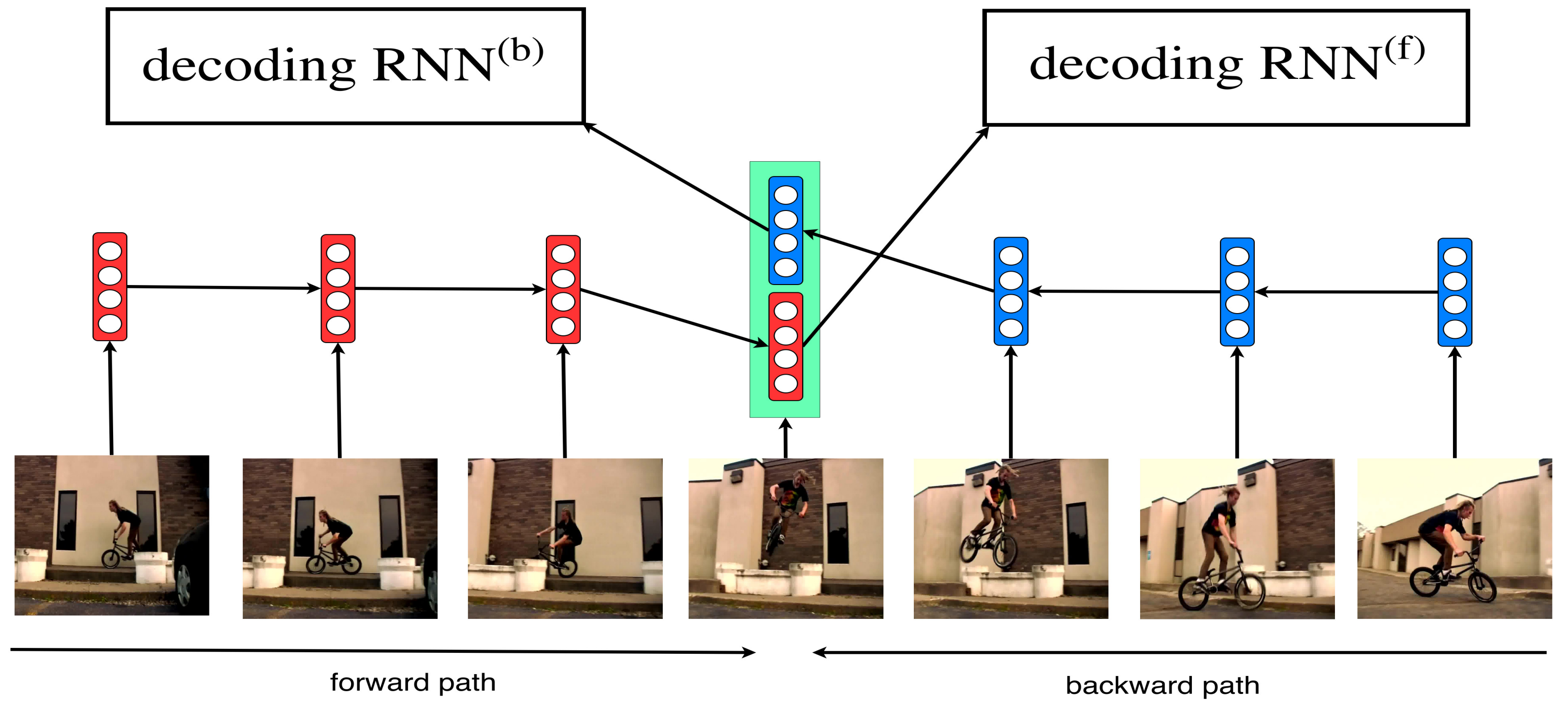


Outer RNN for Considering Surrounding Frames

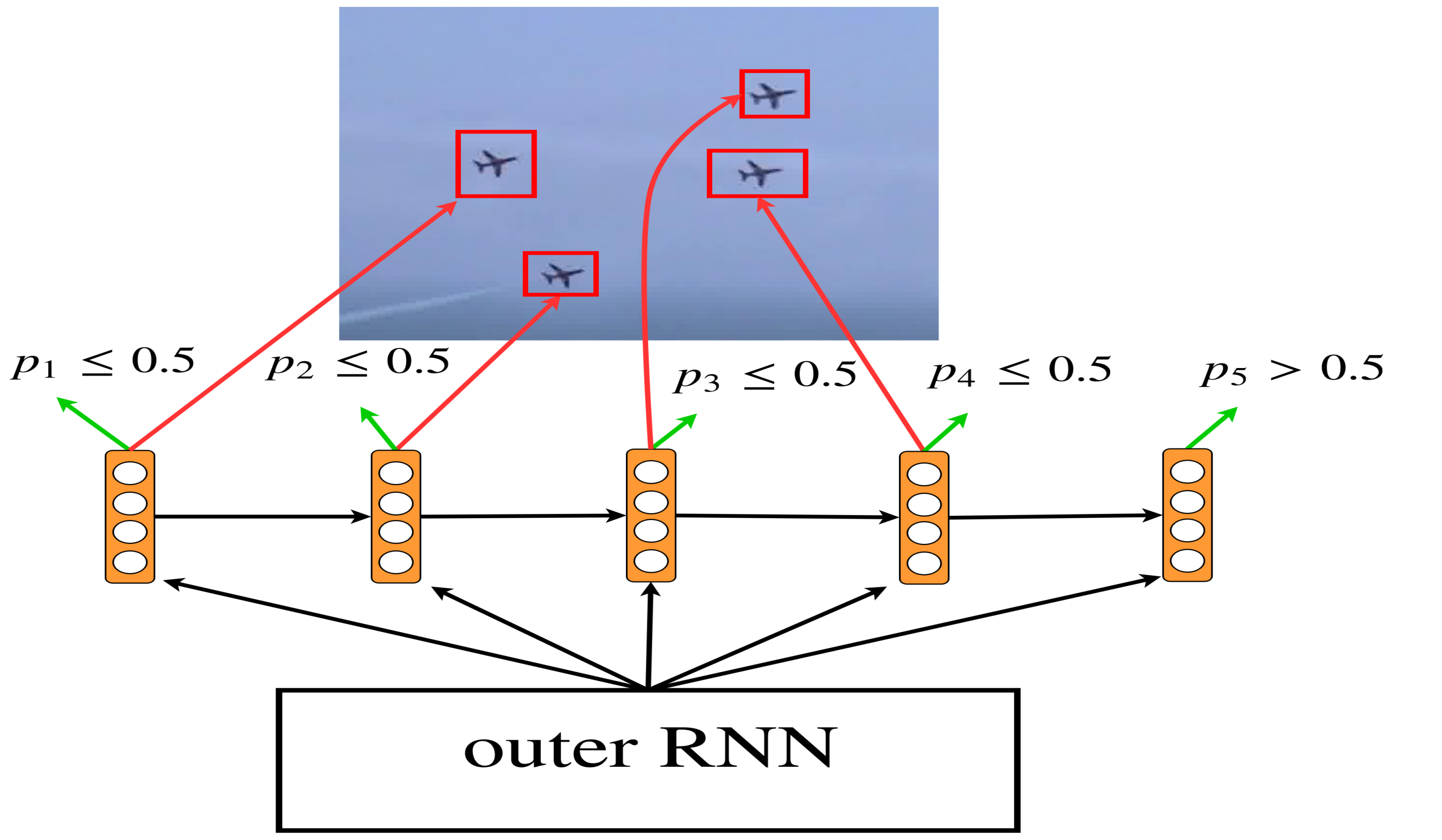


Auxiliary Training for Outer RNN

Reconstruct the original frames (resized to 100×100)



Inner RNN for Sequential Detection



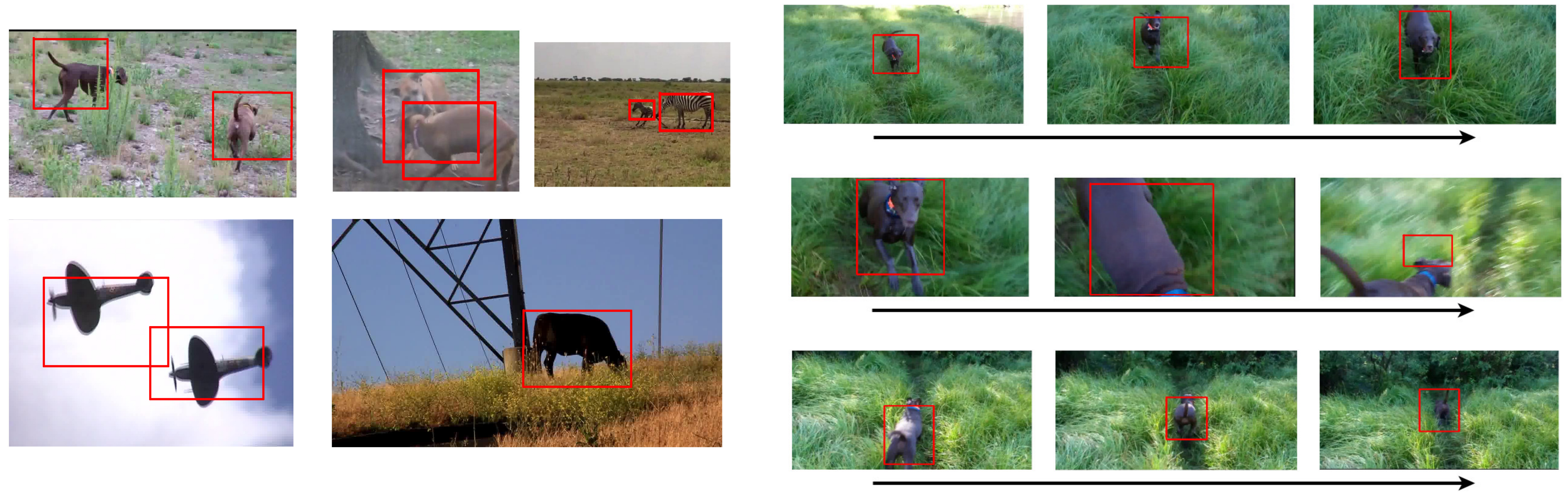
Training

- Loss function:

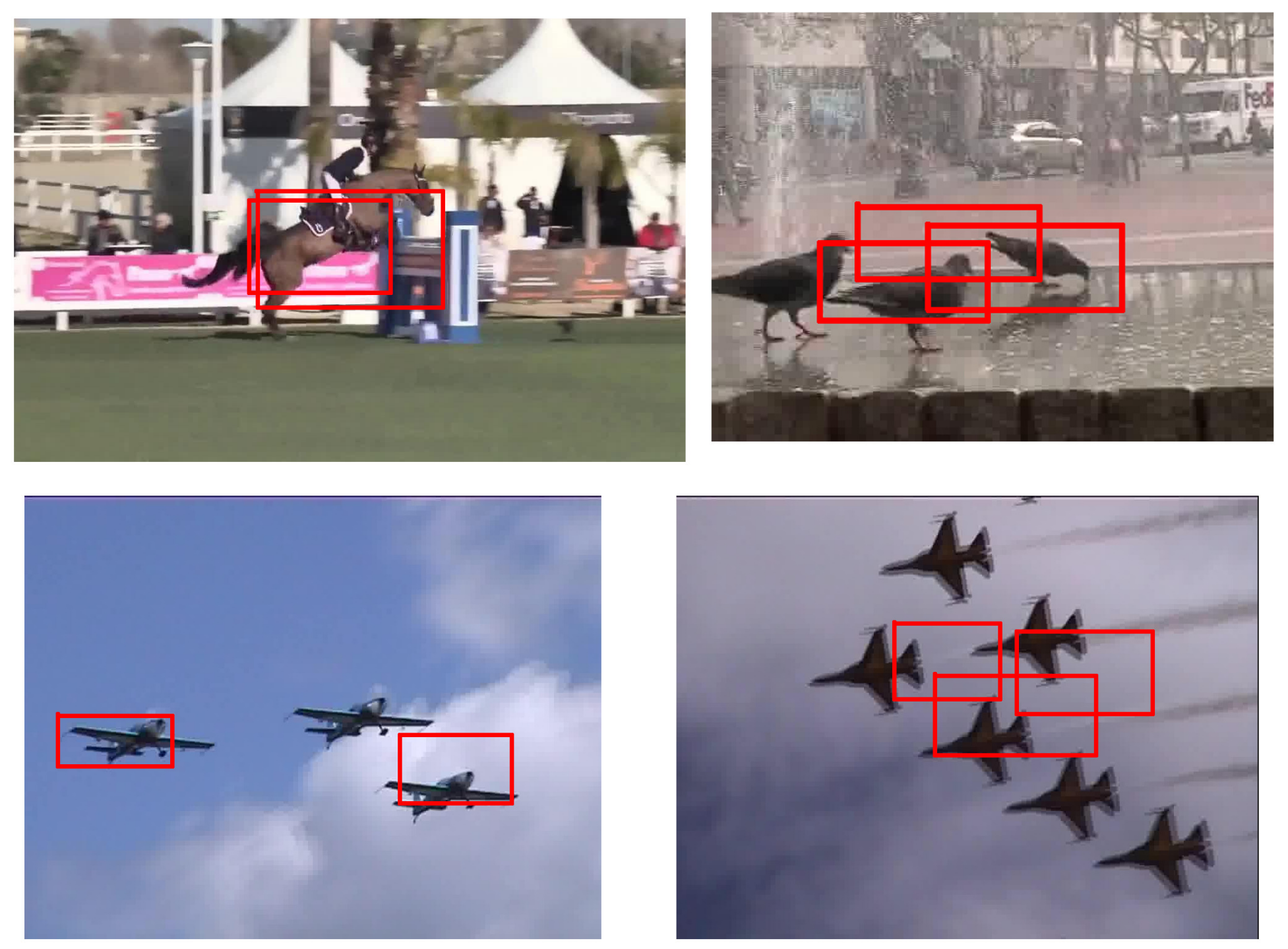
$$L = \alpha L_{\text{position}} + (1-\alpha) L_{\text{stop}} + \beta L_{\text{reconstruction}}$$

- L_{position} and $L_{\text{reconstruction}}$ are mean squared errors, and L_{stop} is binary cross entropy error.
- BPTT and SGD
- We gradually increase the maximum number of objects in every frame in the training set from 1 to 5 during training (curriculum learning).

Positive Results



Negative Results



- There are many negative results.
- Especially, detecting more than three objects is difficult for our current model.
- In addition, our model mistakenly predicts multiple bounding boxes to a single object in some cases.
- We need more efficient architecture (e.g., attention mechanism) and training methods.