Multimedia Laboratory

SENSETIME
商湯集團

SenseTime Group

Object Detection in Videos with Tubelets and Multi-context Cues

Kai Kang
CUvideo Team
The Chinese University of Hong Kong, SenseTime Group

Wanli Ouyang

Kai Kang

Junjie Yan

Xingyu Zeng

Hongsheng Li

Bin Yang

Tong Xiao

Cong Zhang

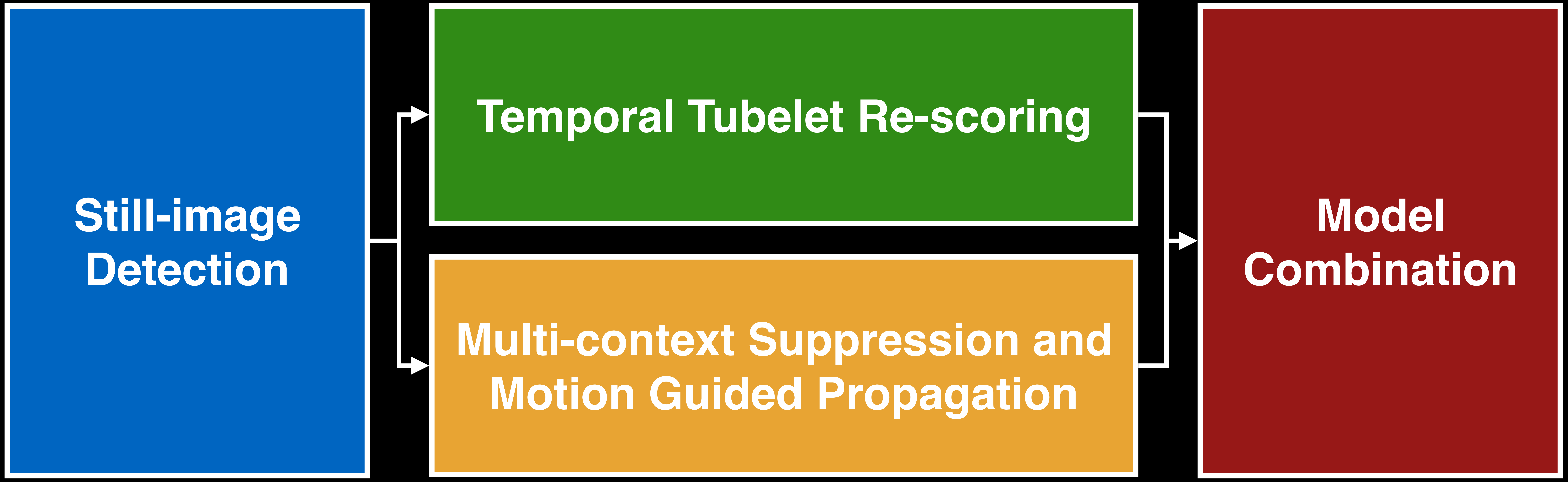
Zhe Wang

Ruohui Wang

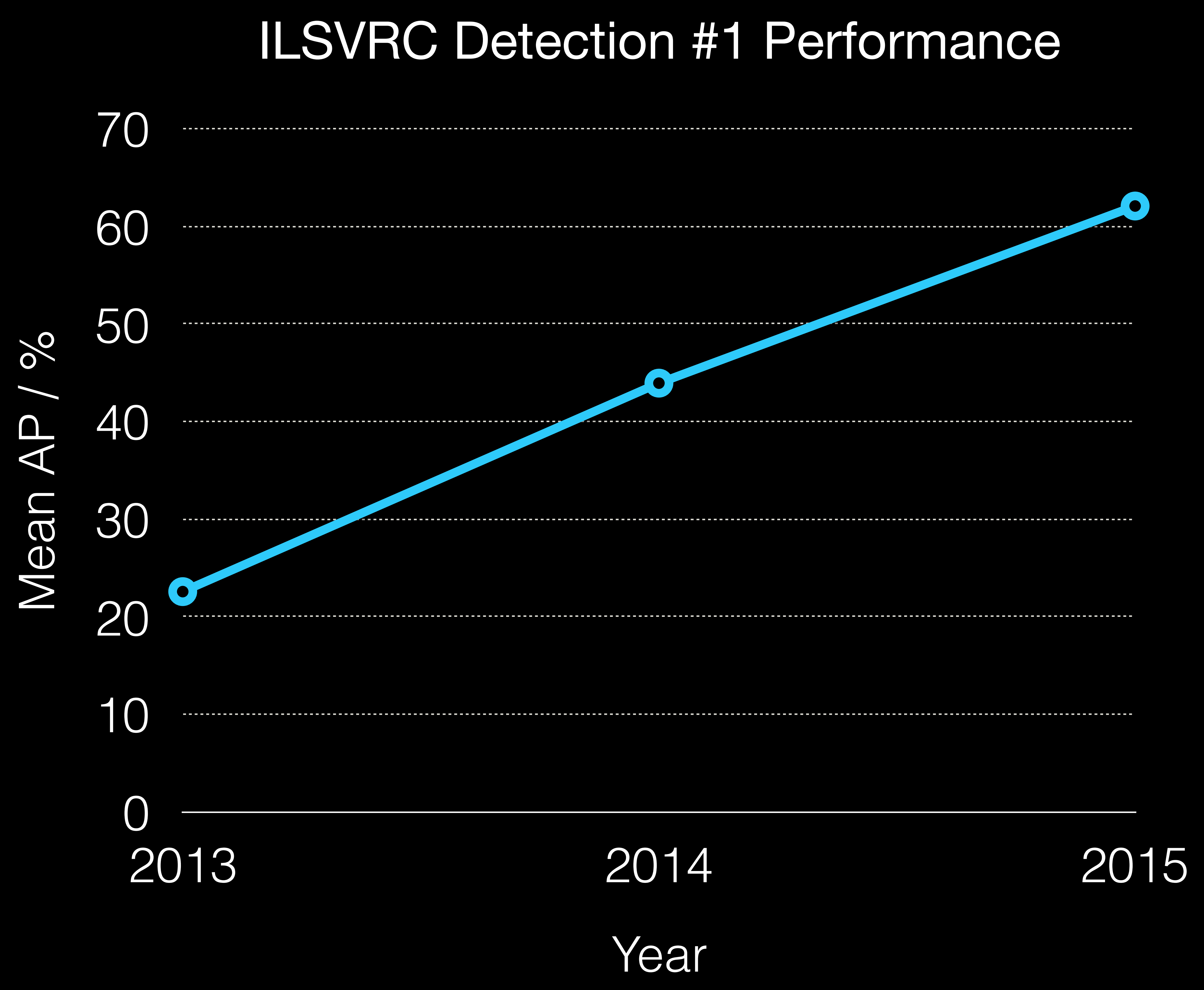
Xiaogang Wang

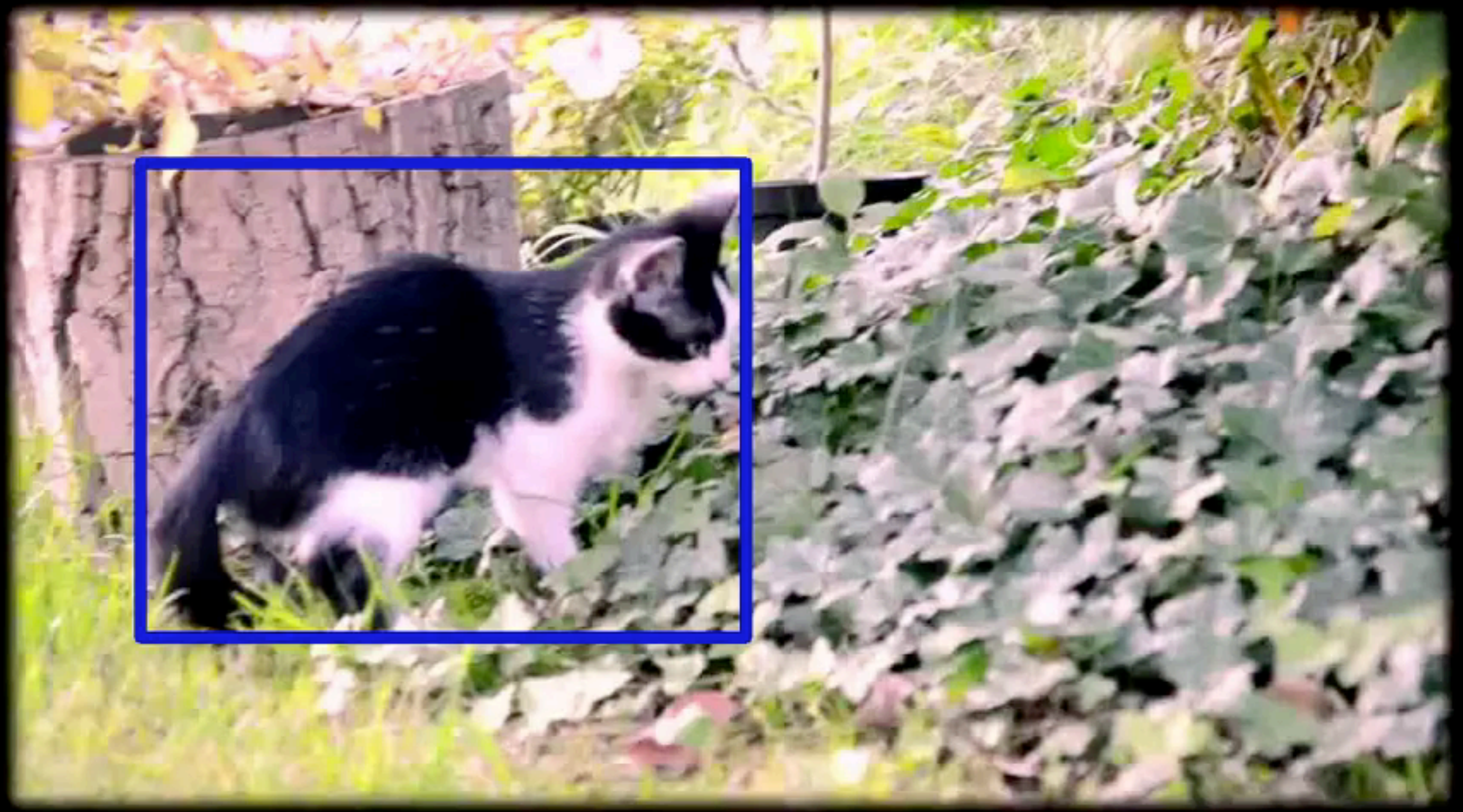
■ CUHK ■ SenseTime

Proposed Framework



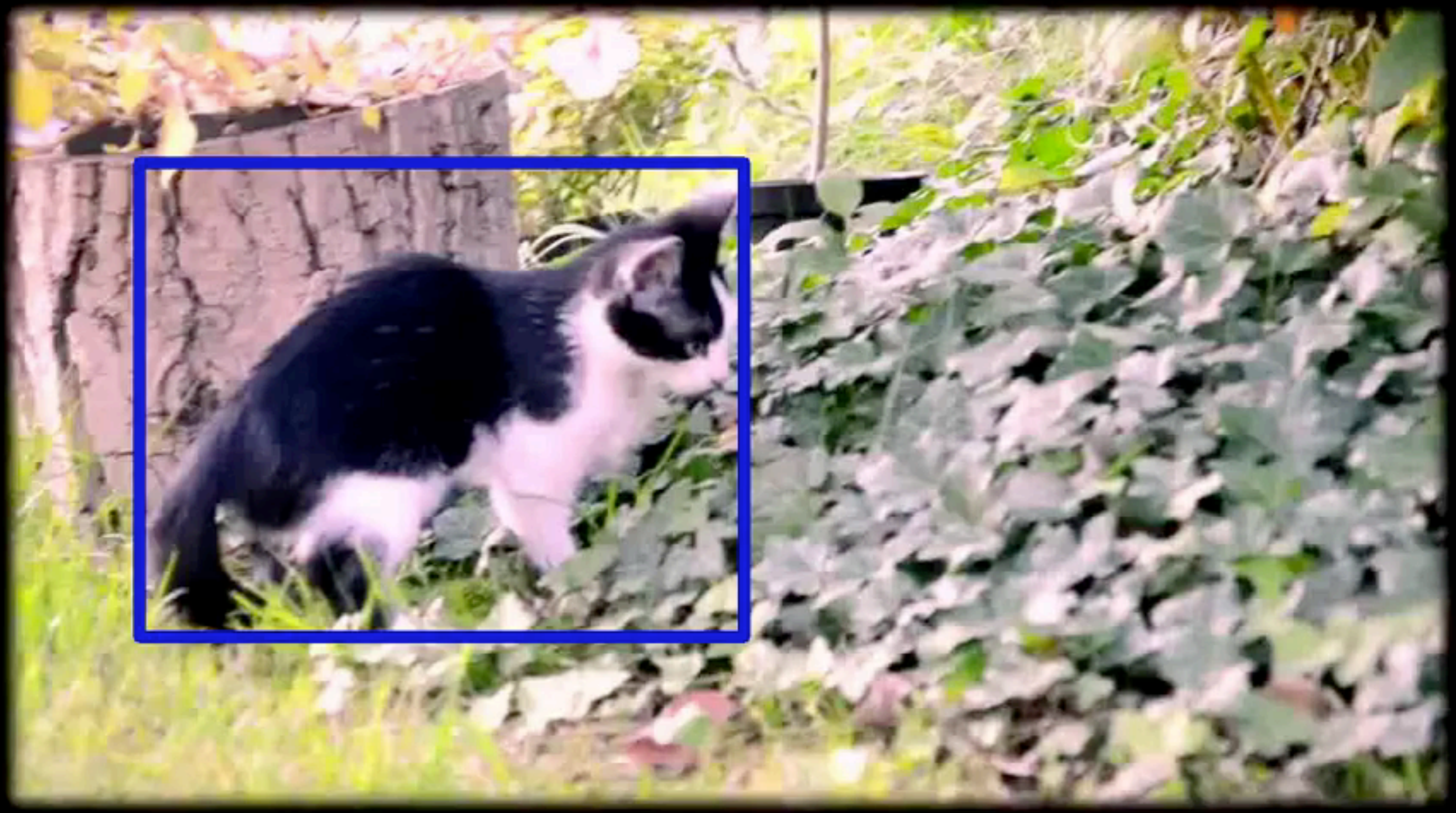
Still-image Detection

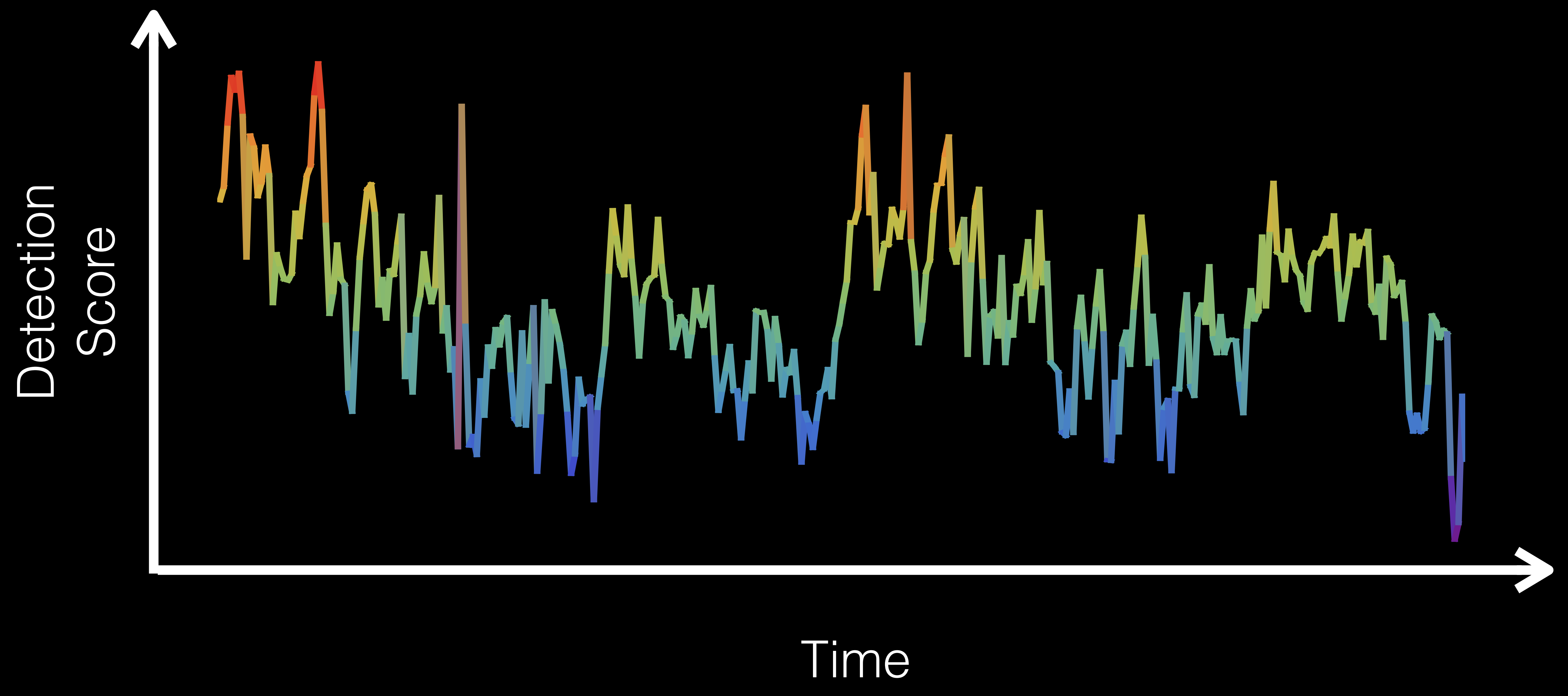




Still-image Detection: Limitation I

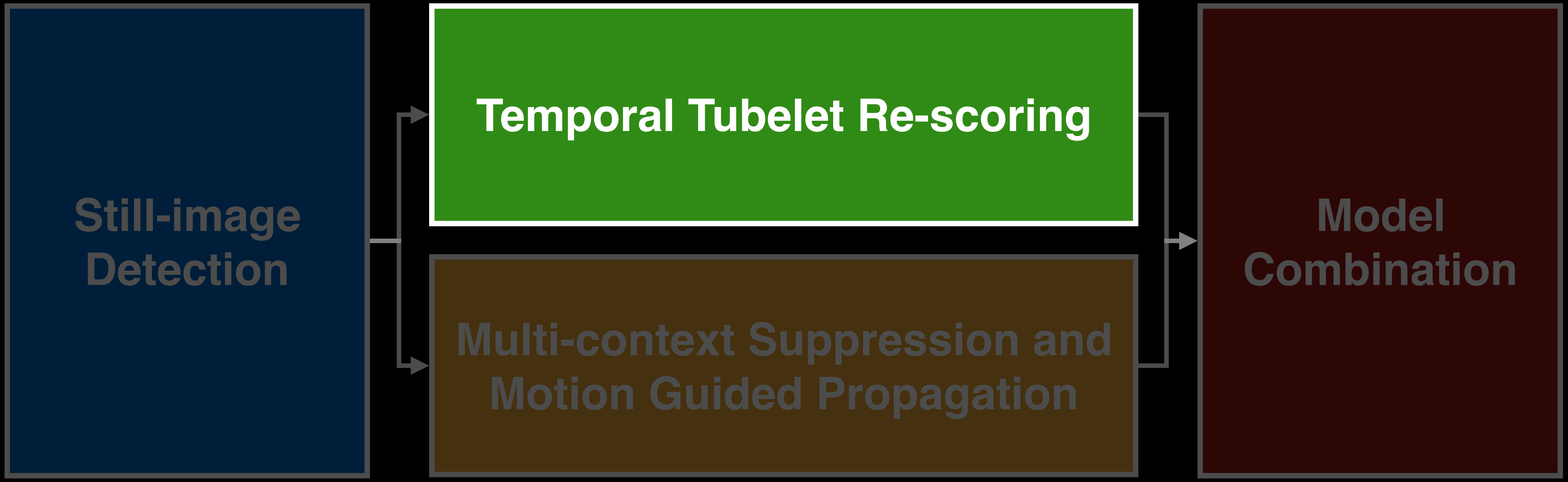
Large Temporal Variations



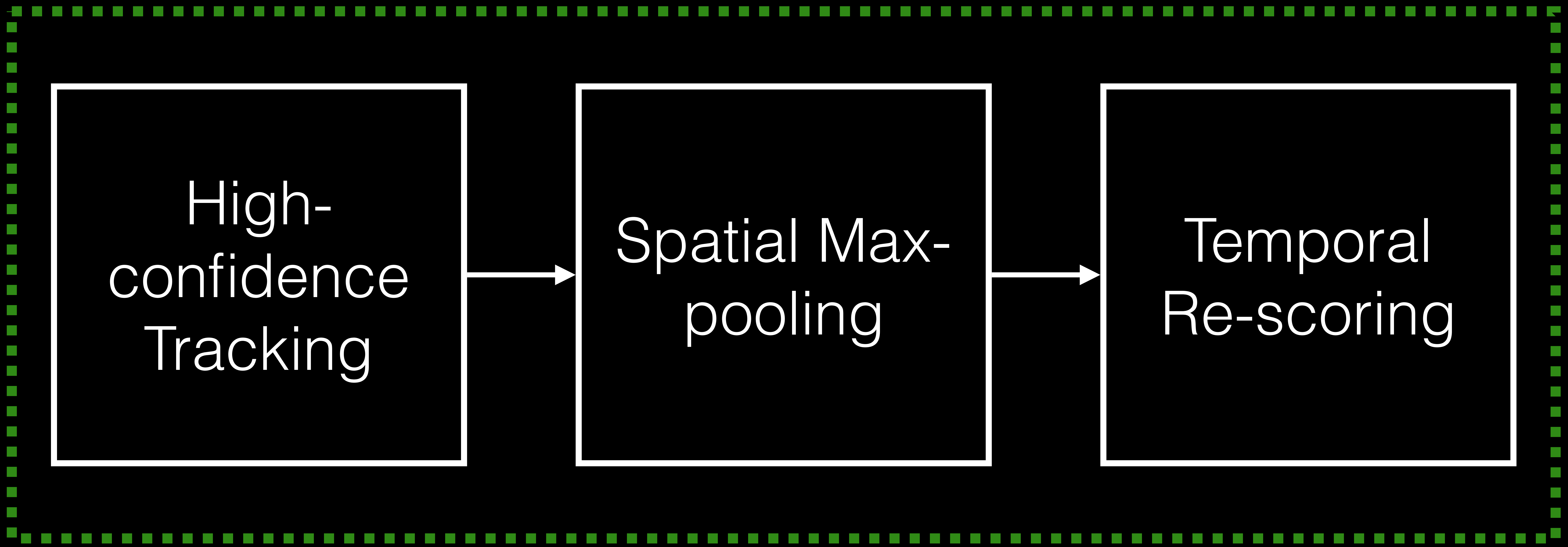


Solution - Tubelets

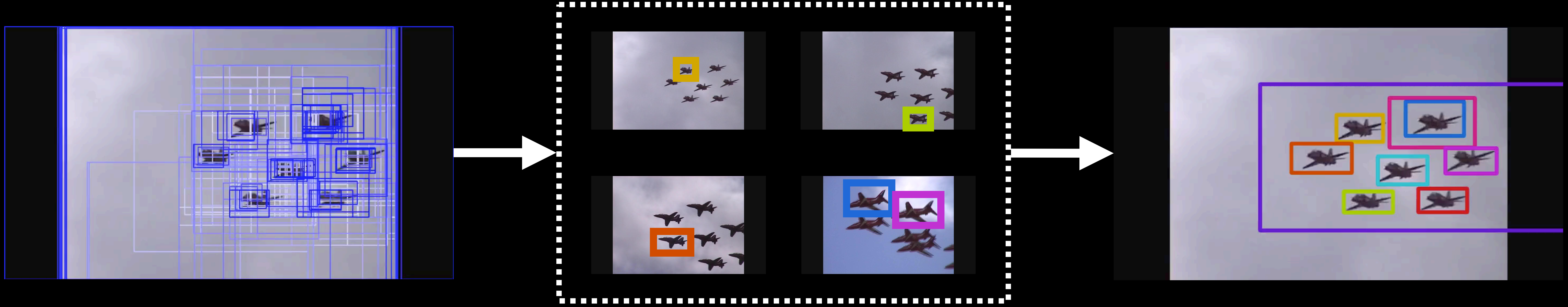
Proposed Framework



Temporal Tubelet Re-scoring



High-confidence Tracking



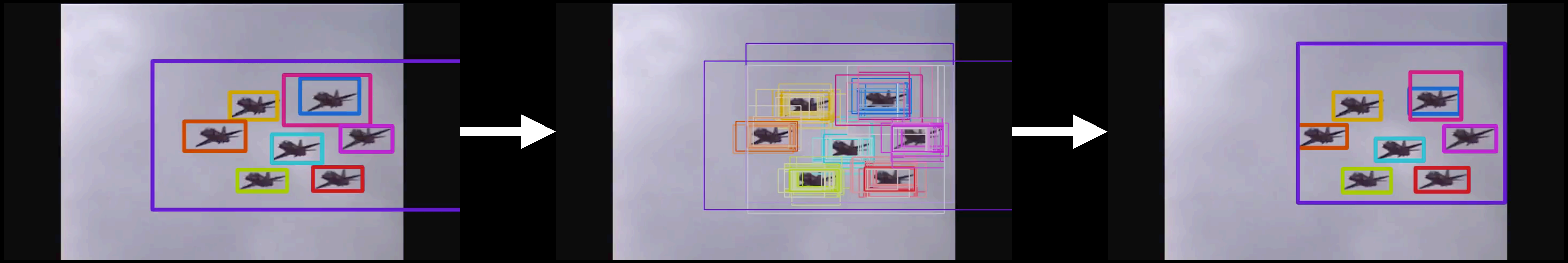
- Obtain detection results from still-image detectors
- Choose high-confidence detections as starting points (anchors) for tracking
- Obtain tubelets, which are bounding box sequences generated from tracking algorithms [1]

[1] Wang, Lijun et al. Visual Tracking with Fully Convolutional Networks. ICCV 2015

Spatial Max-pooling: Why?

- The detection scores on the tracked tubelets are not satisfactory
  - Boxes from tracked tubelets and those from still-image detection have **different statistics**
  - Tracked box locations are not optimal due to **tracking failures**
- Neighboring high-confidence detections are utilized to improve tubelet detection scores, which is called **spatial max-pooling**

Spatial Max-pooling



- Still-image detection results that have **large overlaps with tubelet boxes** are chosen for each tubelet

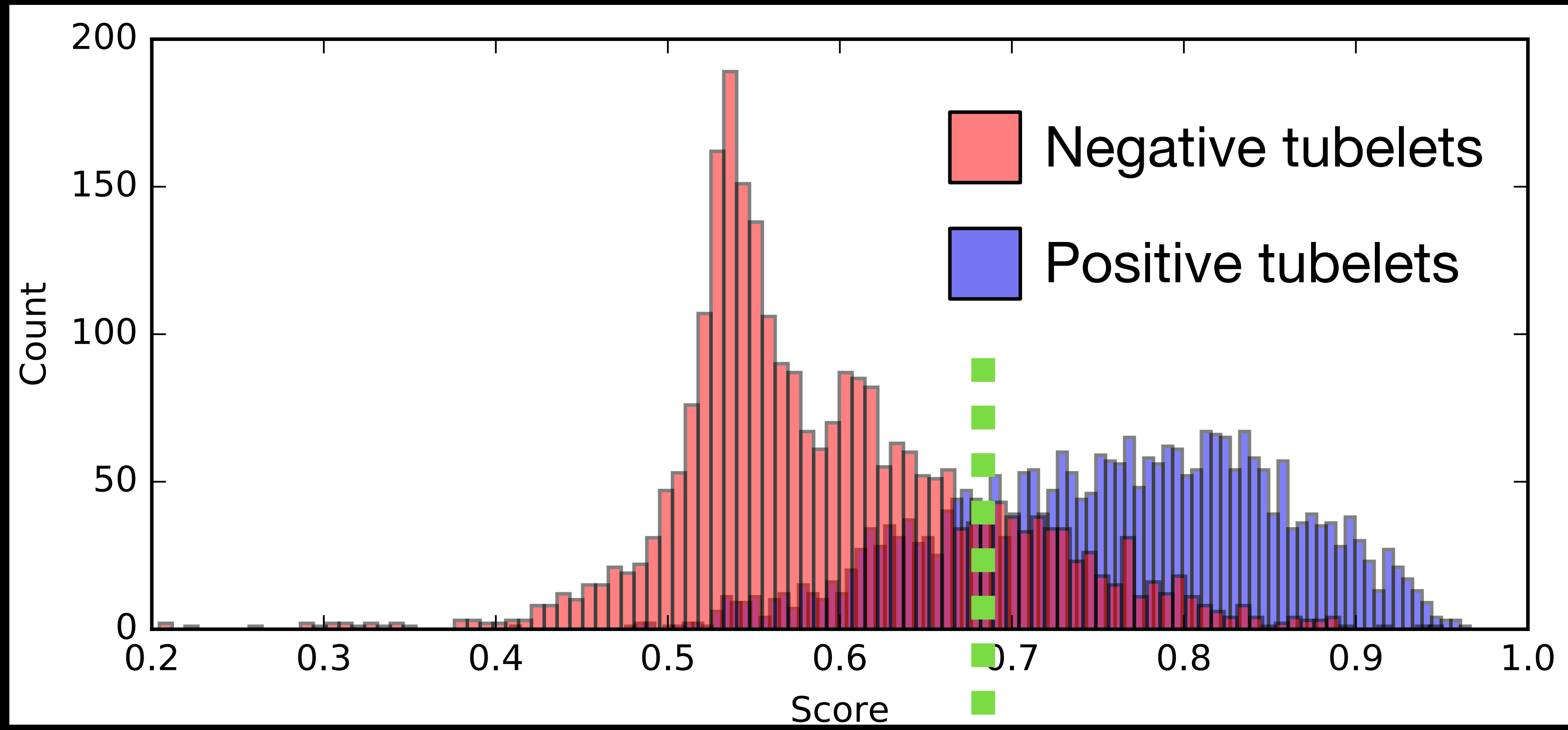
- Only detections with **maximum detection scores** are left after spatial max-pooling

- Use the **Kalman Filter** to smooth the bounding box locations.

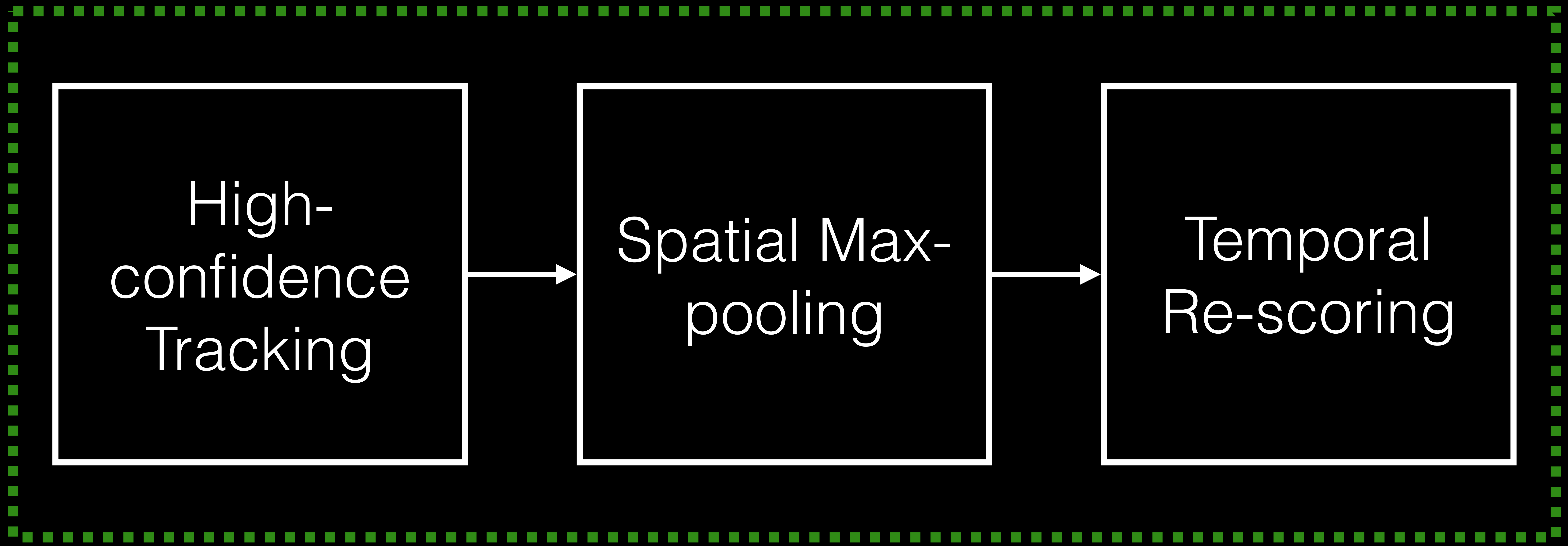
Temporal Re-scoring

- **Tubelet Classification**. Classify tubelets based on statistics of detection scores (mean, median, top-k). A linear classifier is learnt based on the statistics.

- **Tubelet Re-scoring**. Map detection scores of positive tubelets to [0.5, 1], negative ones to [0, 0.5].



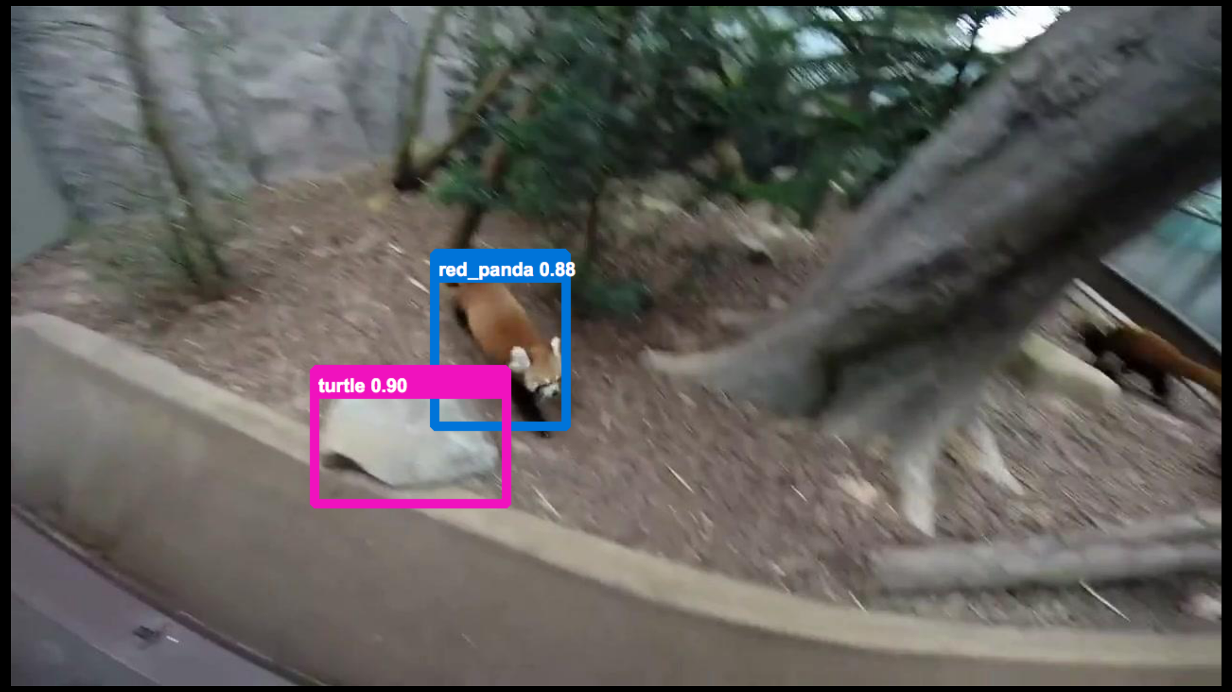
Temporal Tubelet Re-scoring



Still-image Detection: Limitation II

Ignored Context

red panda turtle



red panda turtle

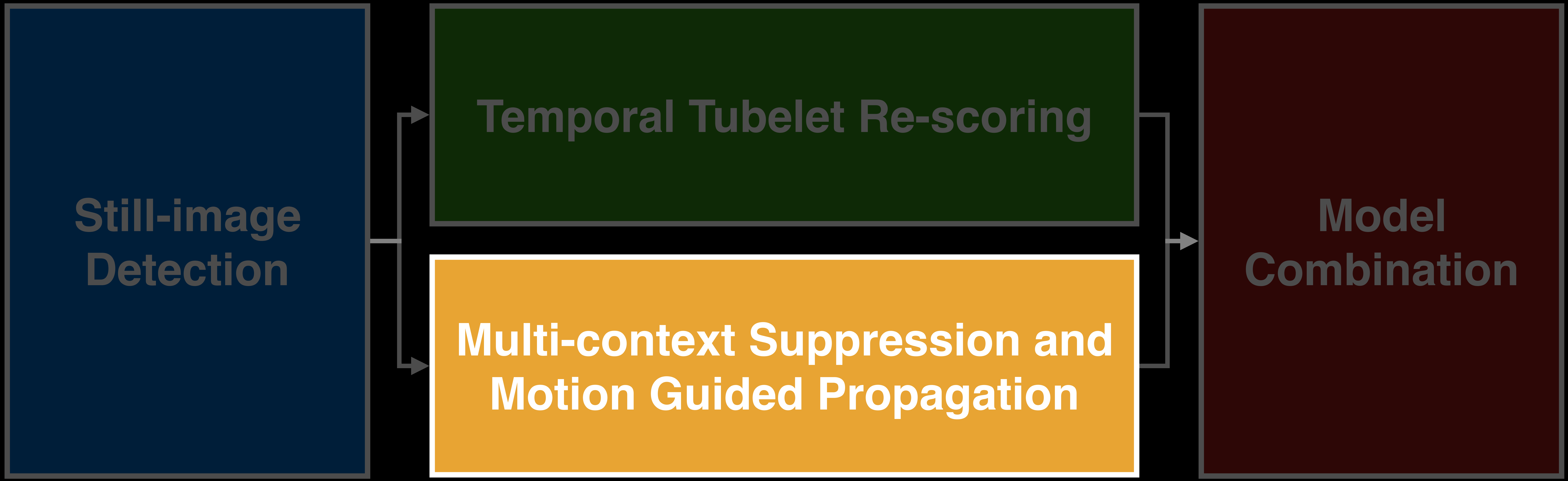
red panda

red panda

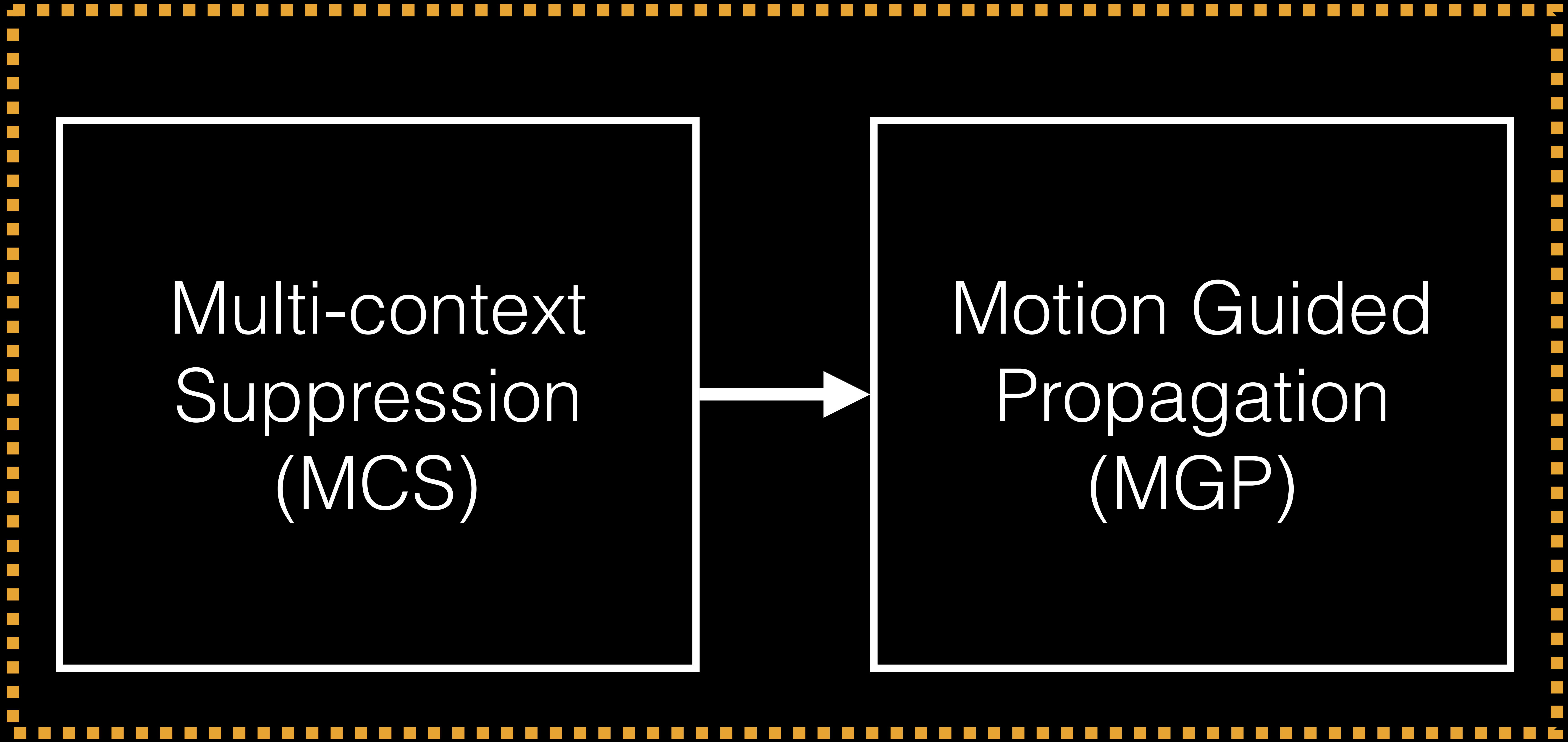
red panda

red panda

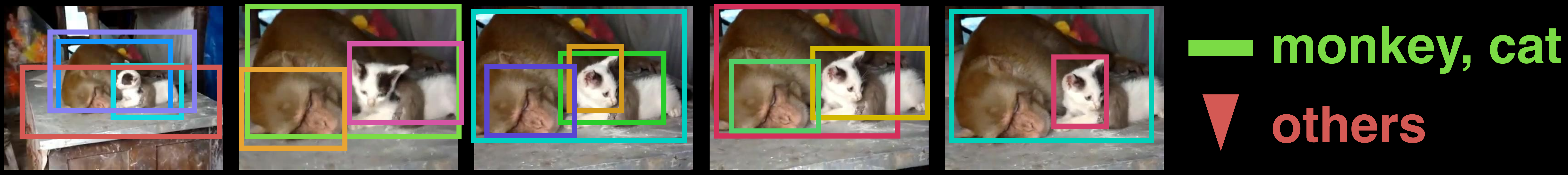
Proposed Framework



Multi-context Suppression and Motion Guided Propagation

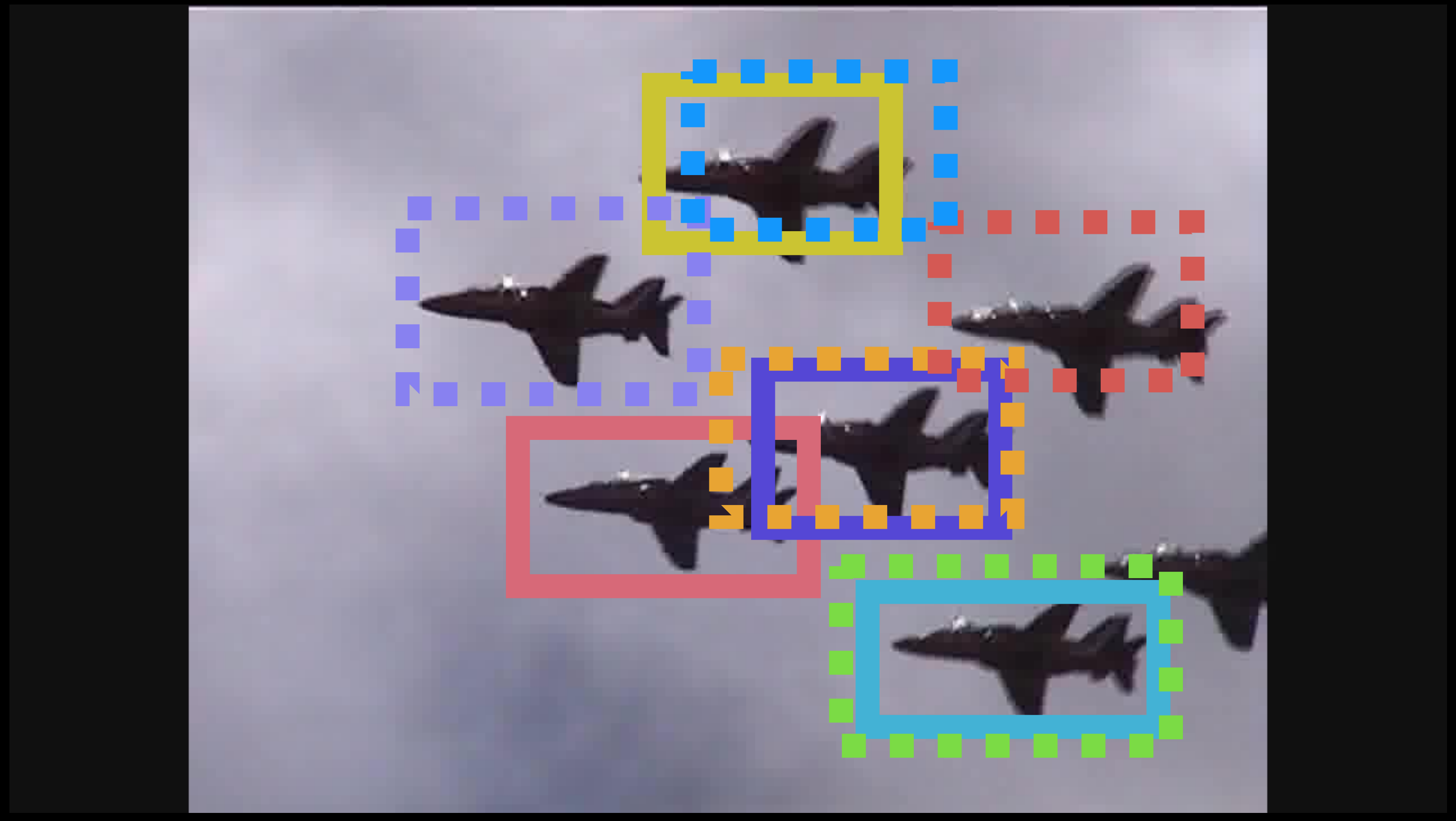


Multi-context Suppression (MCS)

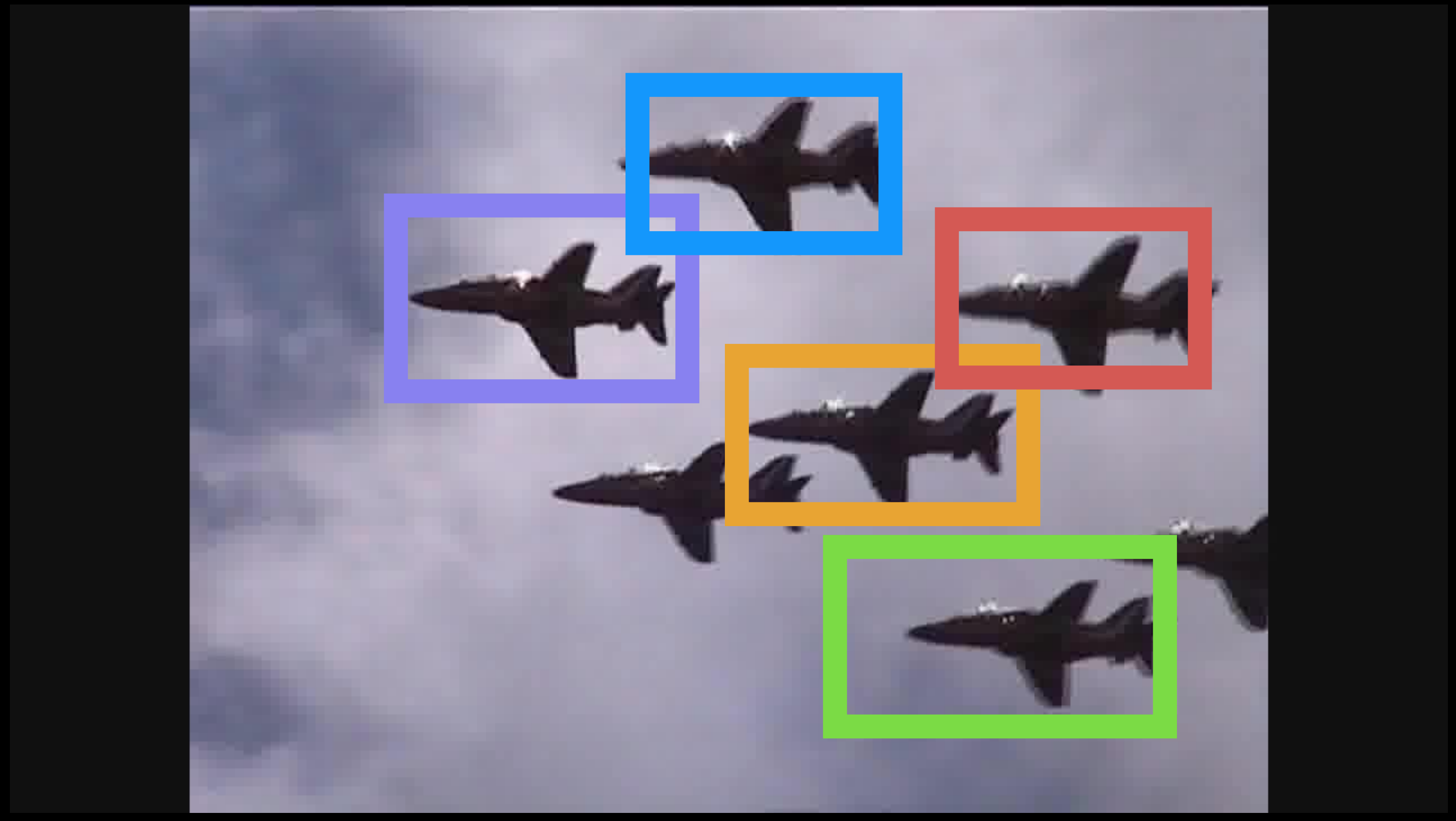


- **Sort** all detection scores of all proposals in a video in **descending order**
- The classes of the **high rankings** are denoted as the confident classes
- The scores of **classes with low rankings** are suppressed, while the scores of confident classes remain unchanged

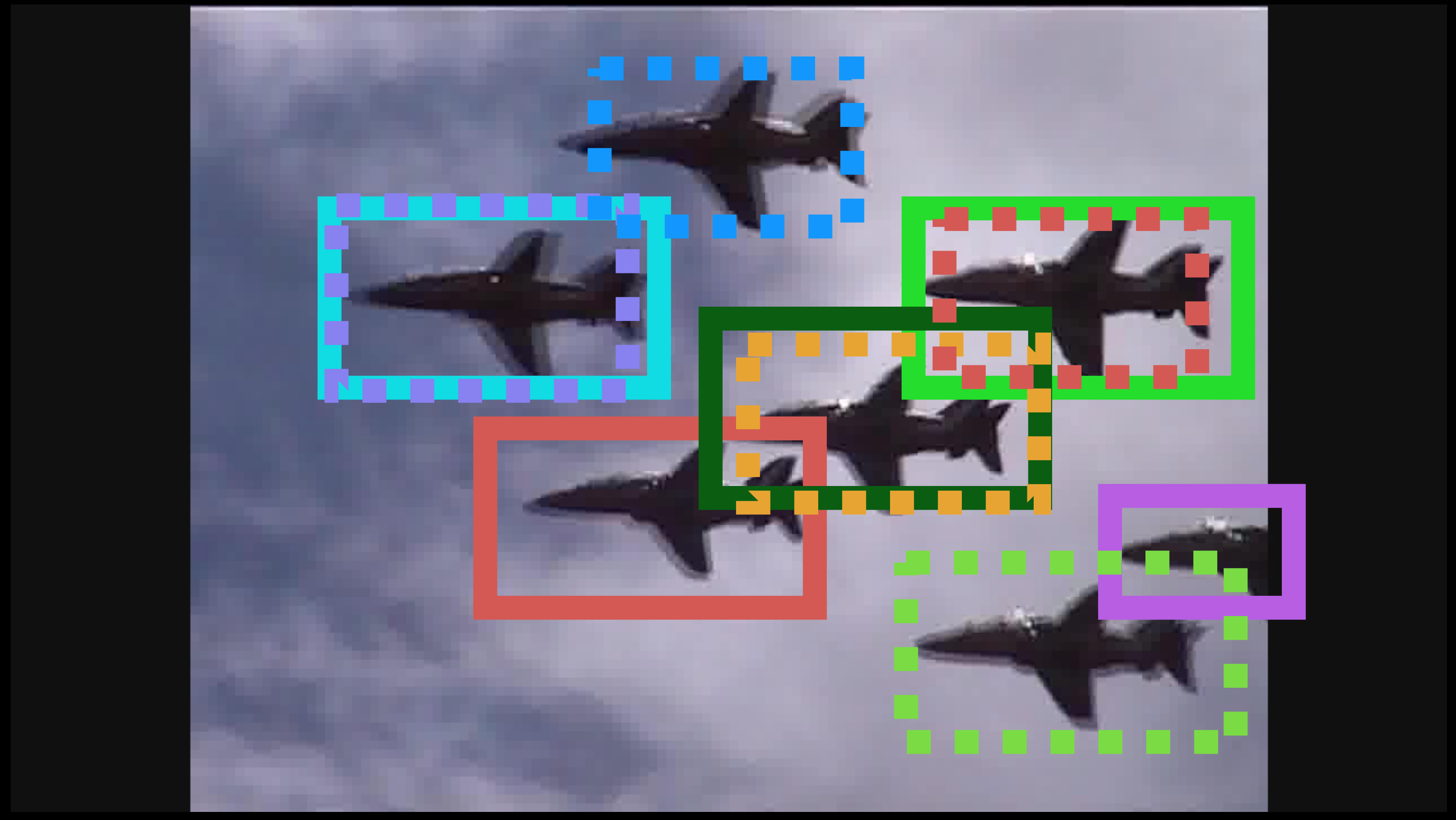
Motion Guided Propagation (MGP)



Frame t-1



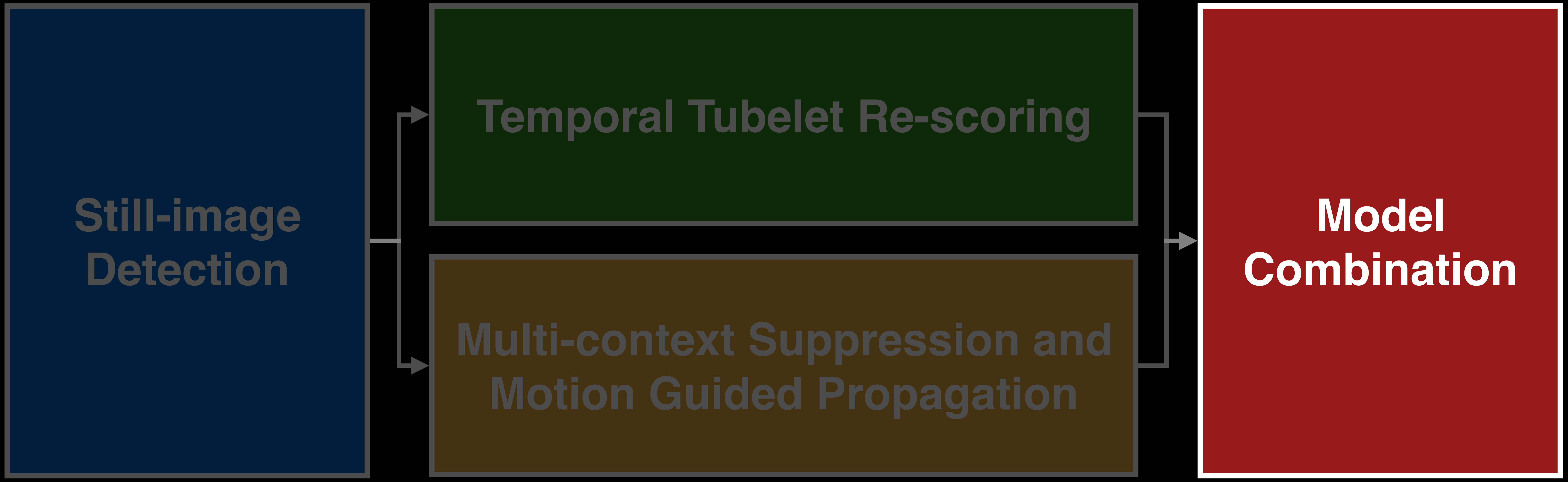
Frame t



Frame t+1

- In each frame, some objects are **not found by detector**. However, detections on adjacent frames are **complementary** to each other.
- Detections are **propagated to adjacent** frames. Optical flow is used for guiding the propagation.
- Propagation results in redundant boxes, which can be **easily handled** by non-maximum suppression (NMS)

Proposed Framework

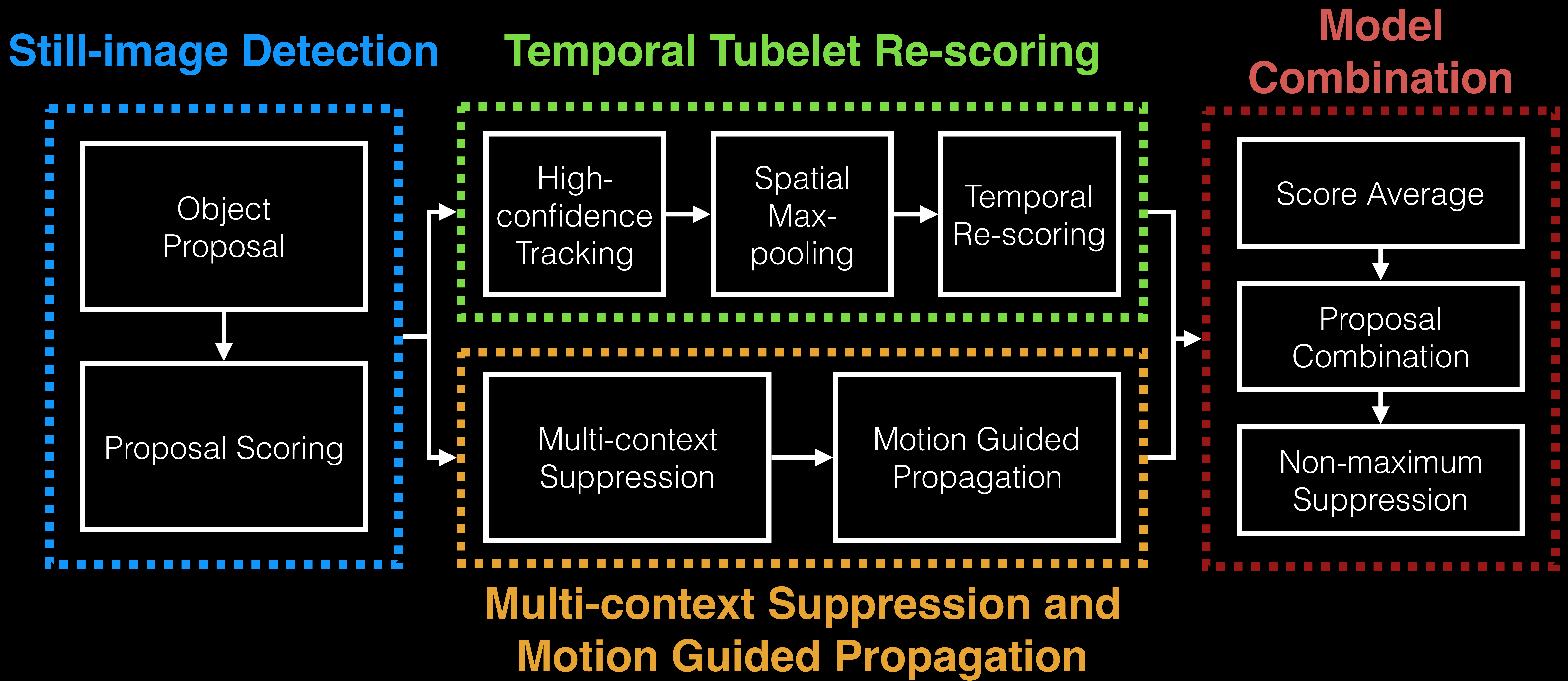


Model Combination

- Two groups of proposals:
  - 1) Proposals from CRAFT [1]: scores from CRAFT
  - 2) Selective Search + EdgeBox: scores from DeepID-net [2]
- Given a group of proposals, their detection scores can be obtained by averaging several models.
- NMS is used for combining multiple groups of proposals

[1] J. Yan, et al. CRAFT Objects from Images, axiv preprint.
[2] W. Ouyang, et al. Deepid-net: Deformable deep convolutional neural networks for object detection. CVPR, 2015.

Proposed Framework



Component Analysis

Training Data Configuration

CNN Training Data

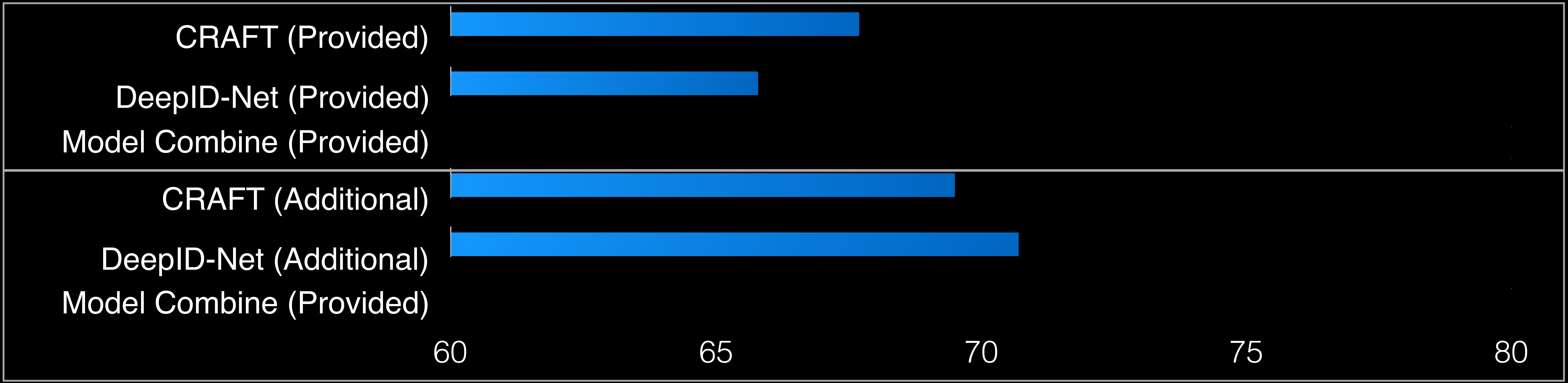
DET:VID Ratio	1:0	3:1	2:1	1:1	1:3
MeanAP / %	49.8	56.9	**58.2**	57.6	57.1

SVM Training Data

DET Positive	✅	✅	❌	❌	❌	✅
VID Positive	❌	✅	✅	✅	✅	✅
DET Negative	✅	✅	✅	✅	❌	✅
VID Negative	❌	❌	❌	✅	✅	✅
MeanAP / %	49.8	47.1	35.8	51.6	52.3	**53.7**

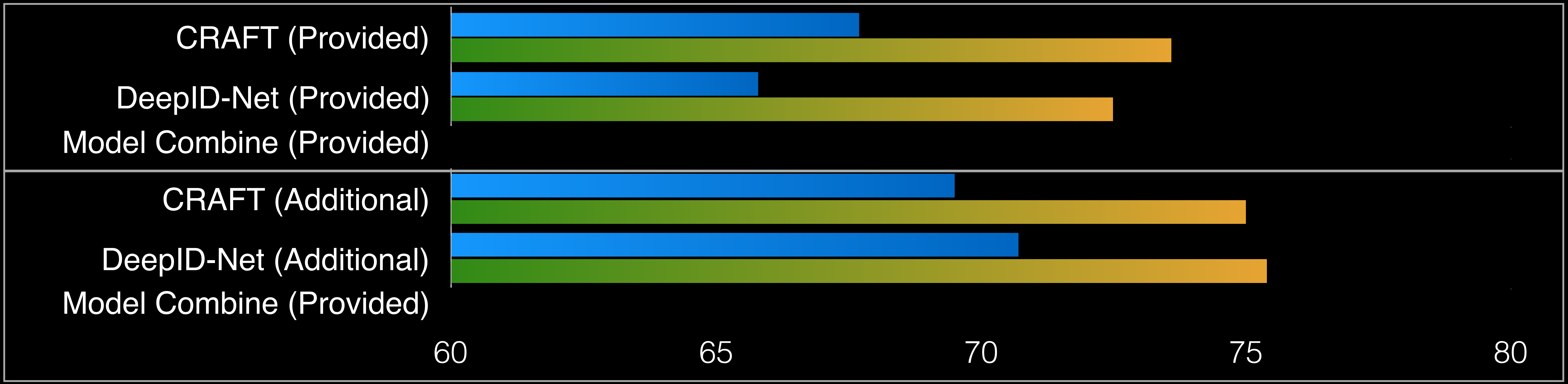
Framework Components

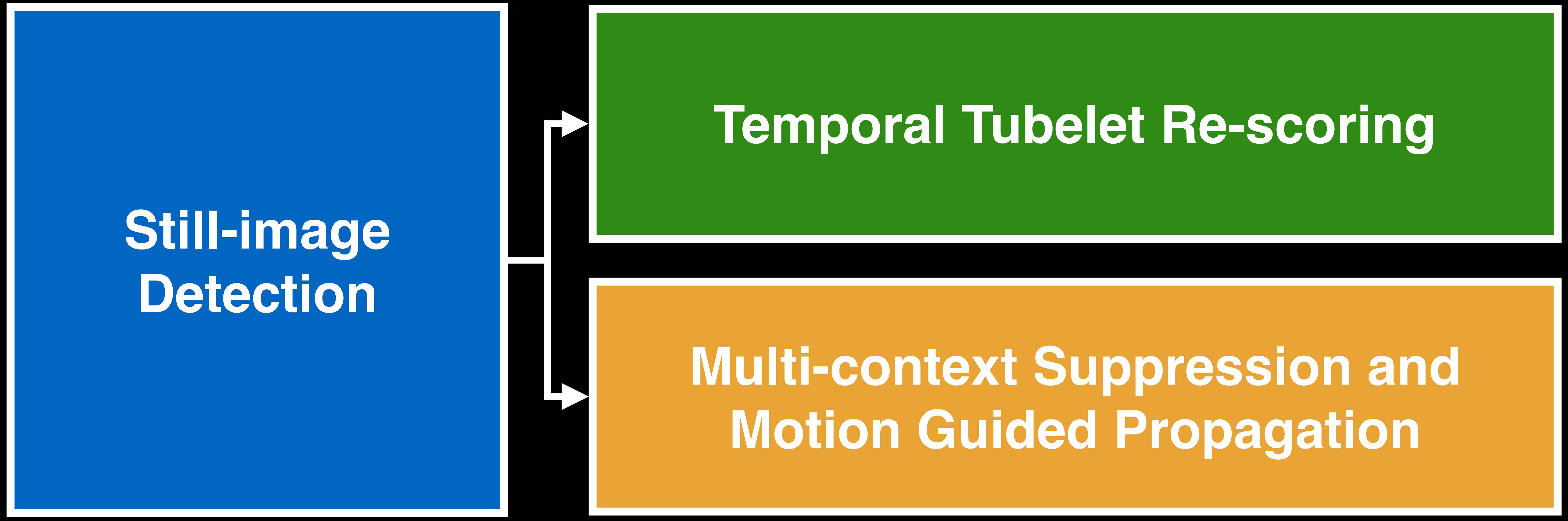
Framework Components



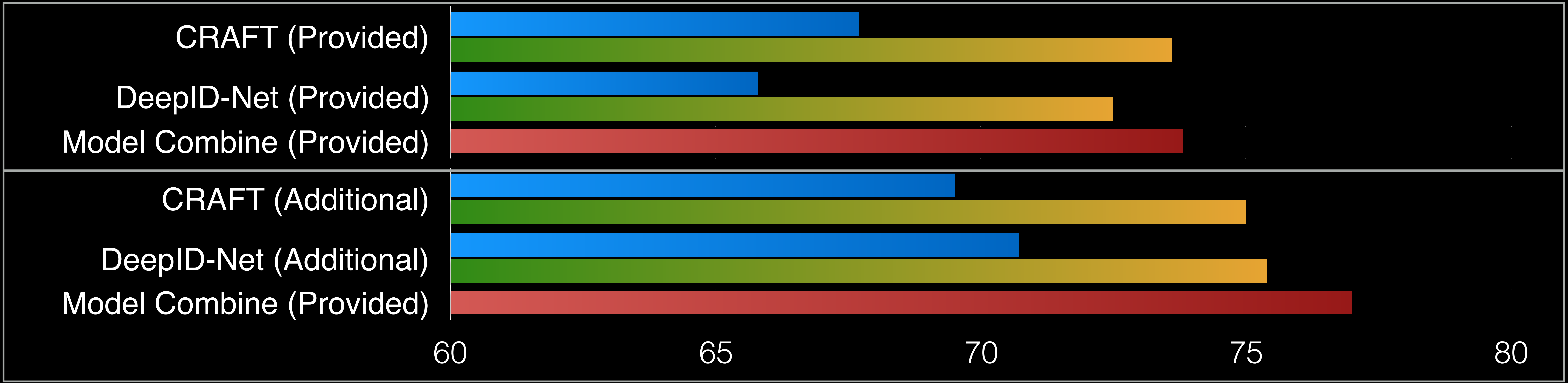
Still-image Detection

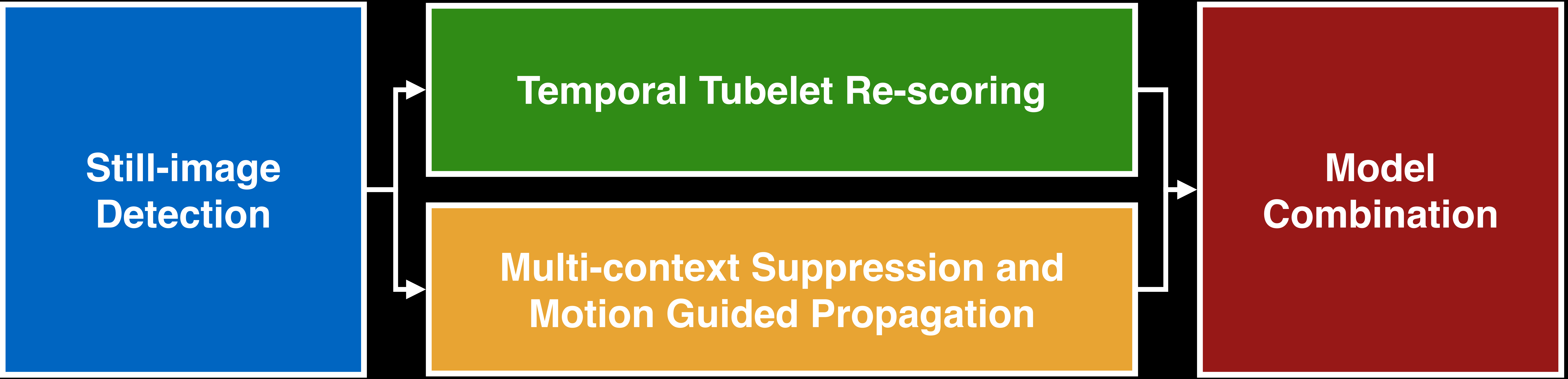
Framework Components





Framework Components





Results

Data	Model	Still-image	MCS+MGP +Rescoring	Model Combine	Test Set (official results)	Rank in ILSVRC 2015	#win
Provided	CRAFT [1]	67.7	73.6	73.8	67.8	#1	28/30
	DeepID-net [2,3,4]	65.8	72.5				
Additional	CRAFT [1]	69.5	75.0	77.0	69.7	#2	11/30
	DeepID-net [2,3,4]	70.7	75.4				

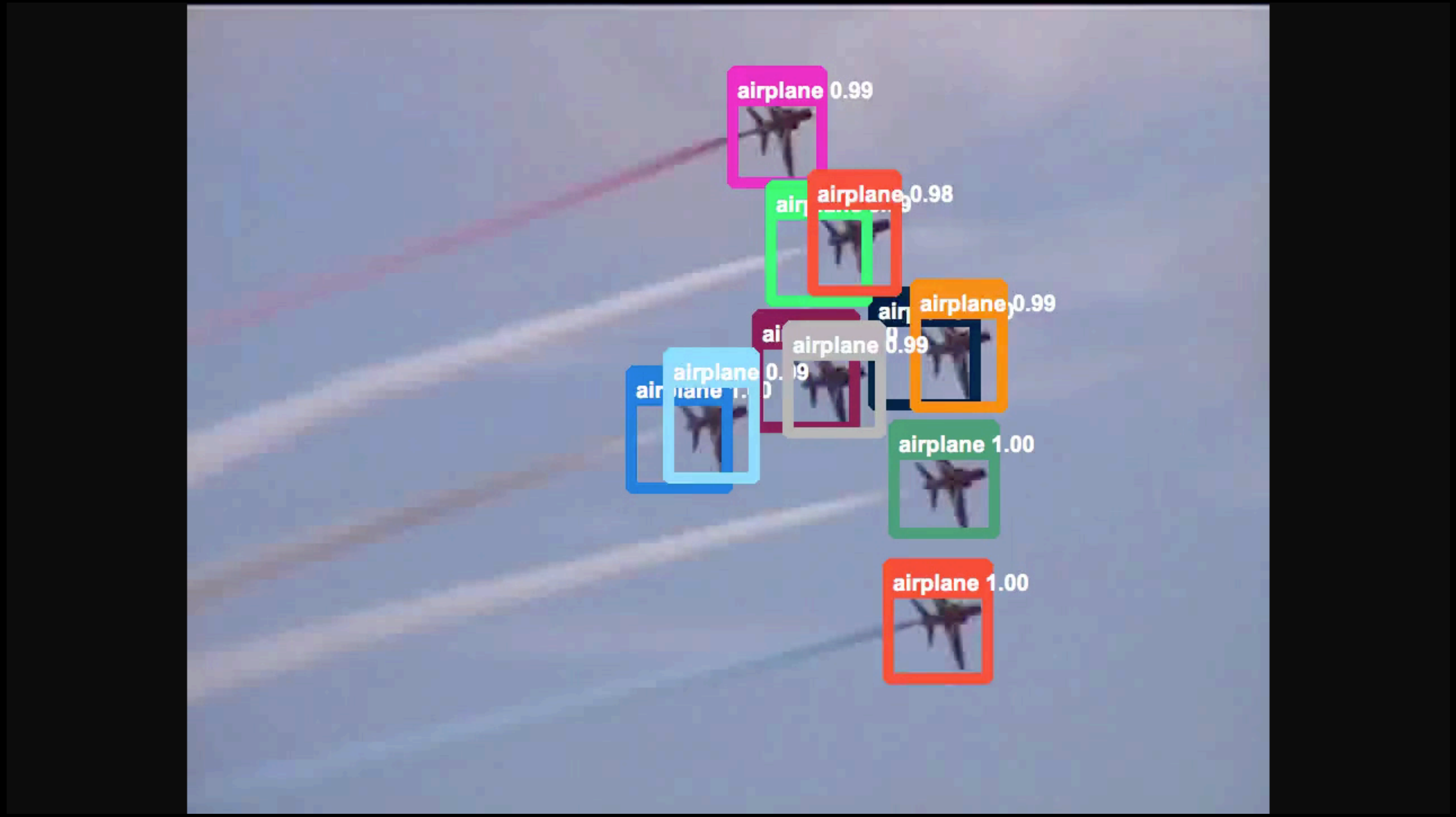
Validation set   Test set

[1] J. Yan, et al. CRAFT Objects from Images, axiv preprint.
[2] W. Ouyang, et al. Deepid-net: Deformable deep convolutional neural networks for object detection. CVPR, 2015.
[3] X. Zeng, et al. Window-Object Relationship Guided Representation Learning for Generic Object Detections , axiv preprint.
[4] W. Ouyang, et al. Factors in Finetuning Deep Model for object detection, axiv preprint.

Our Team in ILSVRC2015

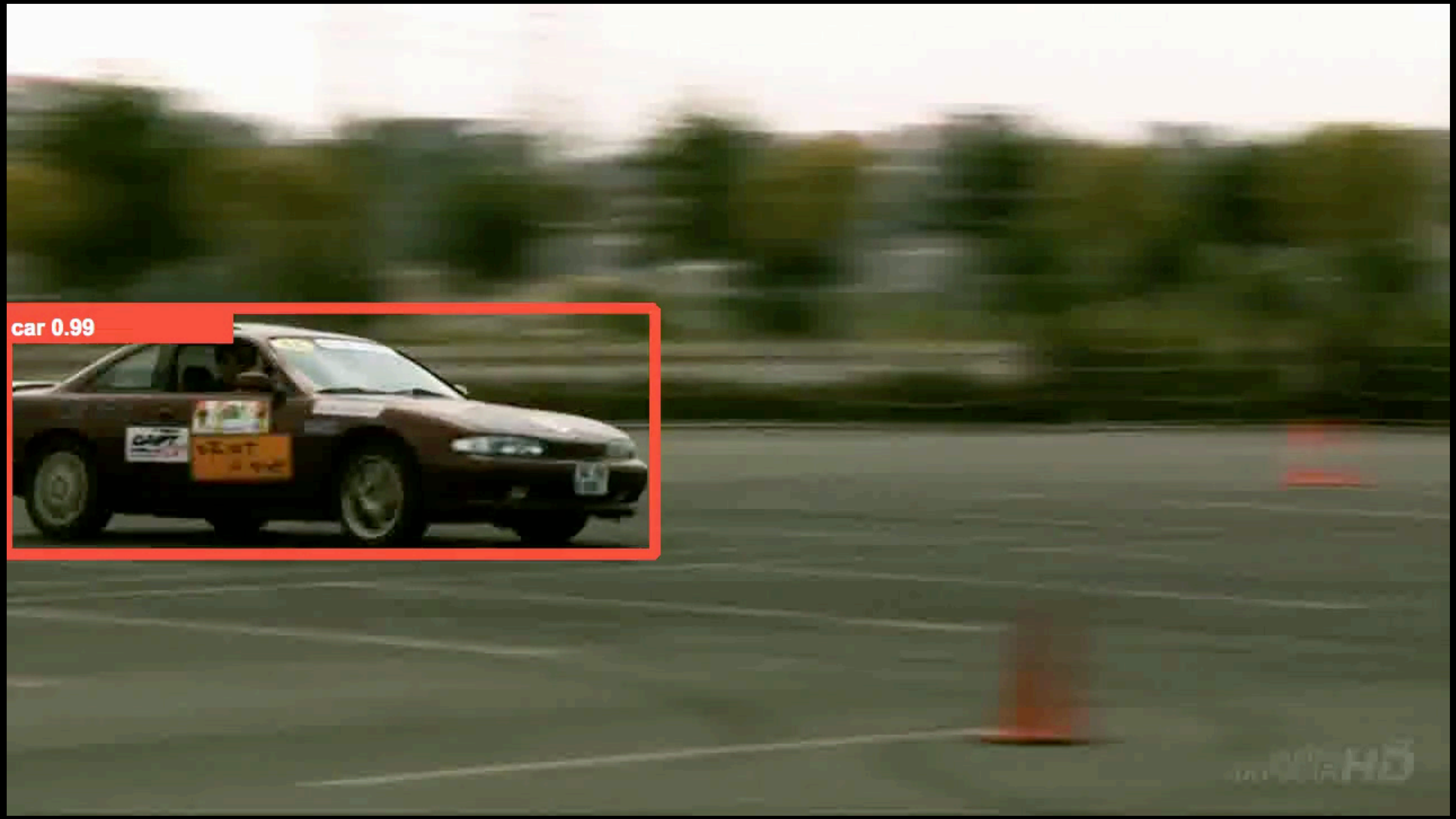
Team	Task	Track	Rank
CUimage	DET	Provided	#3
		Additional	#2
CUvideo	VID	Provided	#1
		Additional	#2

Results

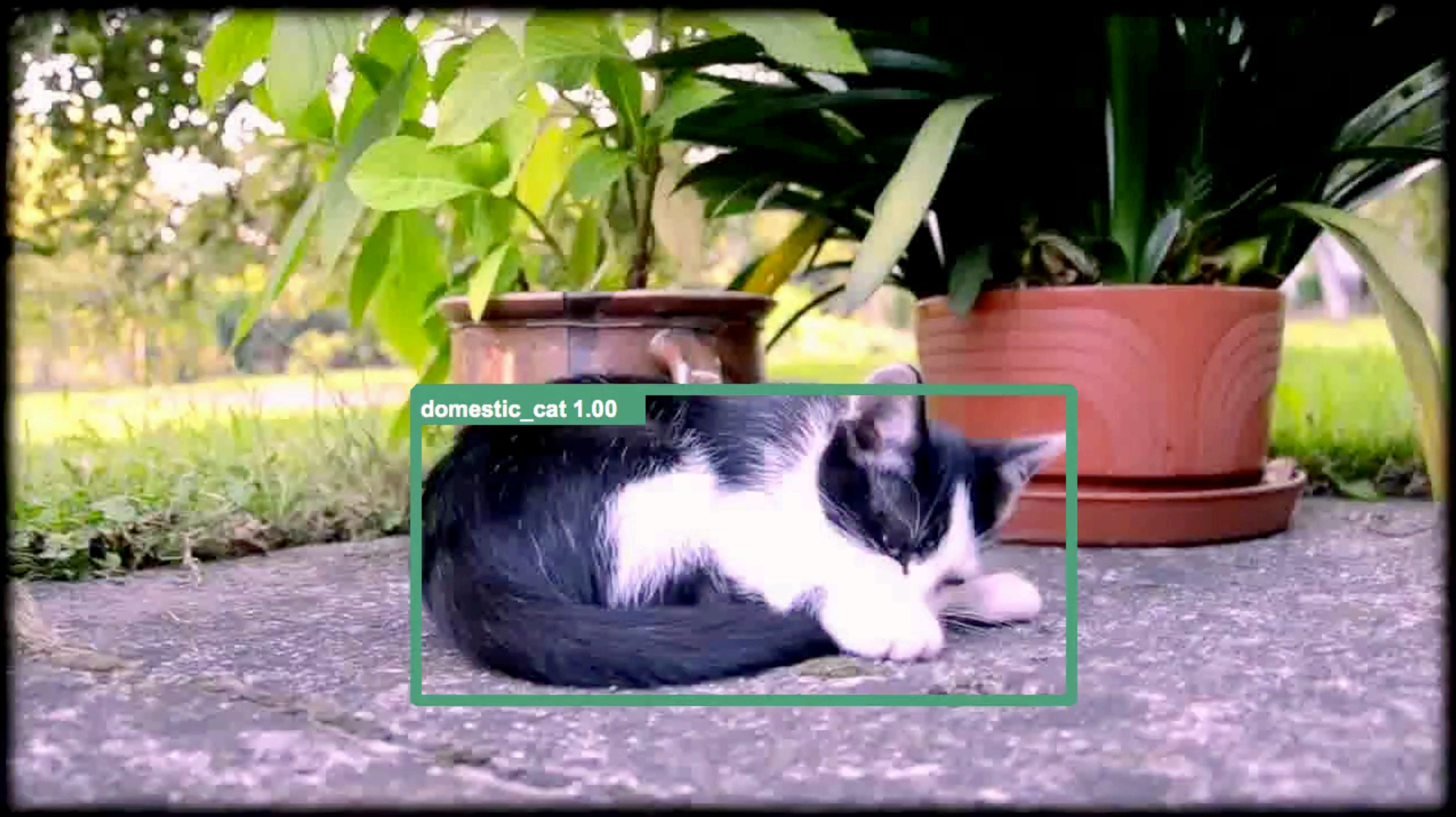

airplane 0.99
airplane 0.98
airplane 0.99
airplane 0.99
airplane 1.00
airplane 1.00

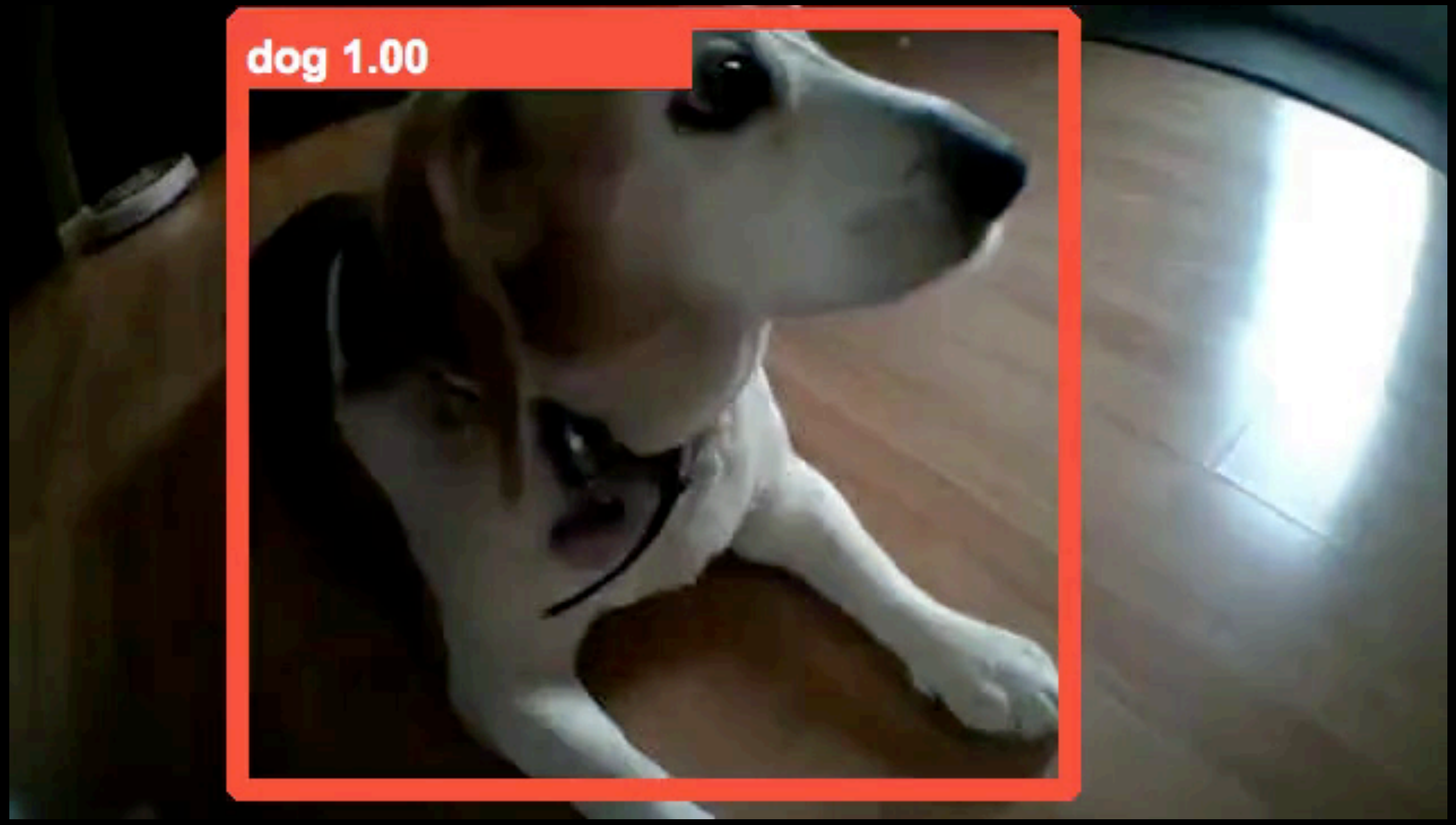

airplane 0.78
airplane 0.85
airplane 0.88
airplane 0.66
airplane 0.72

Results

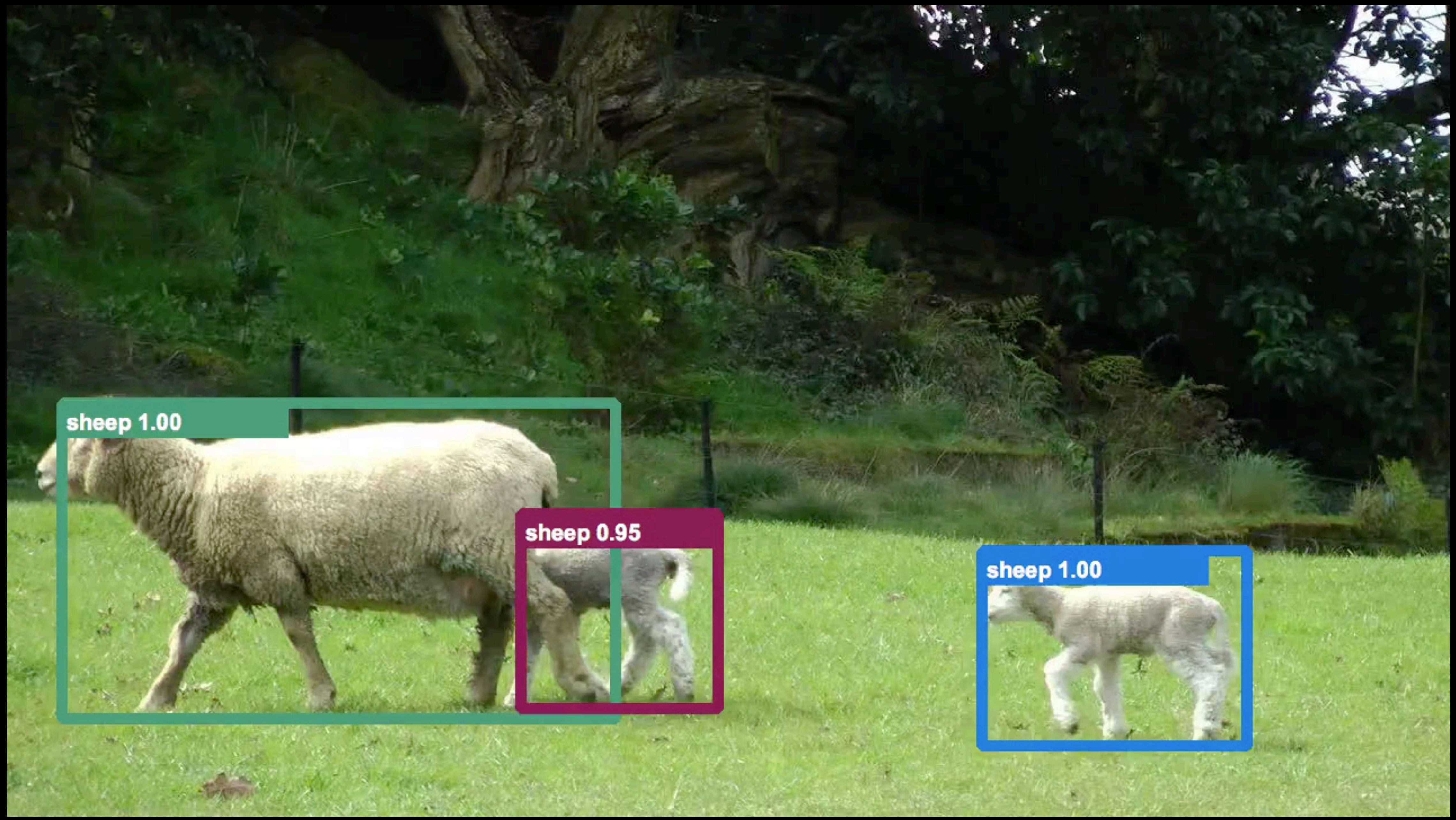


Results





Results



Results

Thank You!

Questions?