香港中文大學
The Chinese University of Hong Kong

Learning effective deep models for object detection and using Multi-Context Cues for video object detection

Wanli Ouyang （欧阳万里）
香港中文大学

我们团队在ImageNet Challenge

Task	Track	Rank
CLS+LOC	Additional	3
DET	Provided	3
DET	Additional	2
VID	Provided	1
VID	Additional	2


Wanli Ouyang


Hongsheng Li


Xiaogang Wang


Junjie Yan


Xingyu Zeng


Kai Kang


Hongyang Li


Zhe Wang


Bin Yang


Cong Zhang


Tong Xiao


Ruohui Wang


Yubin Deng


Xuanyi Dong


Buyu Li

Sihe Wang

纲要

- 图像中的物体检测

欧阳万里

- 视频中的物体检测

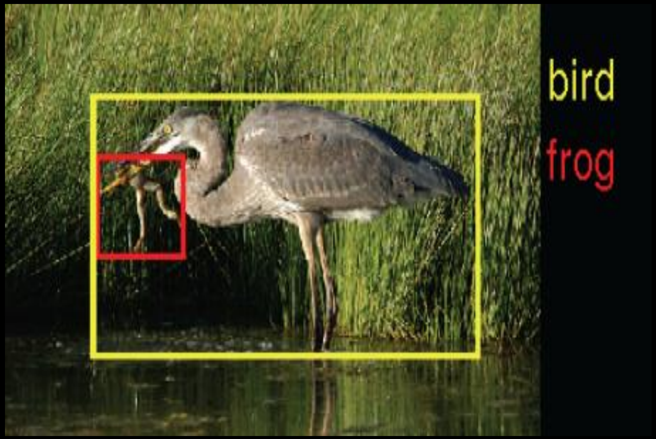
李鸿升

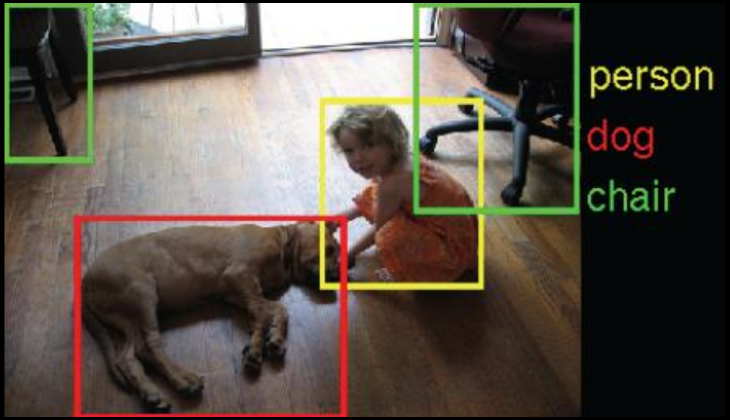
纲要

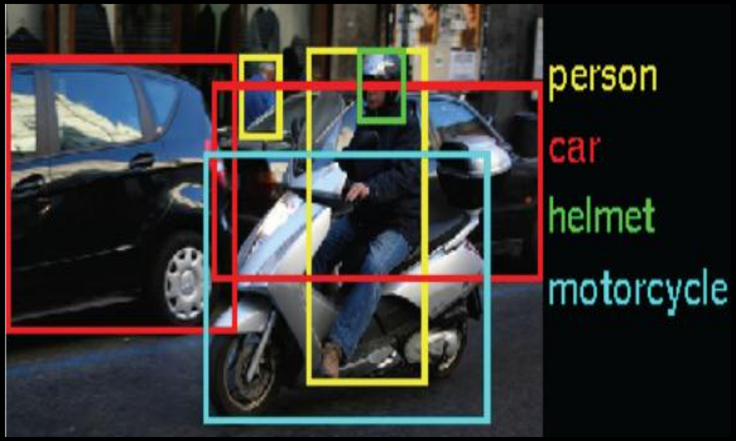
- 图像中的物体检测
  - 简要介绍
  - 基于多个上下文的框和物体关系学习 arXiv:1512.02736
  - 考虑物体长尾性质的分层级联学习 arXiv:1601.05150
  - 框生成和框分类多级级联学习
- 视频中的物体检测

物体检测

- 200 类，~56万训练图片，～5万测试图片

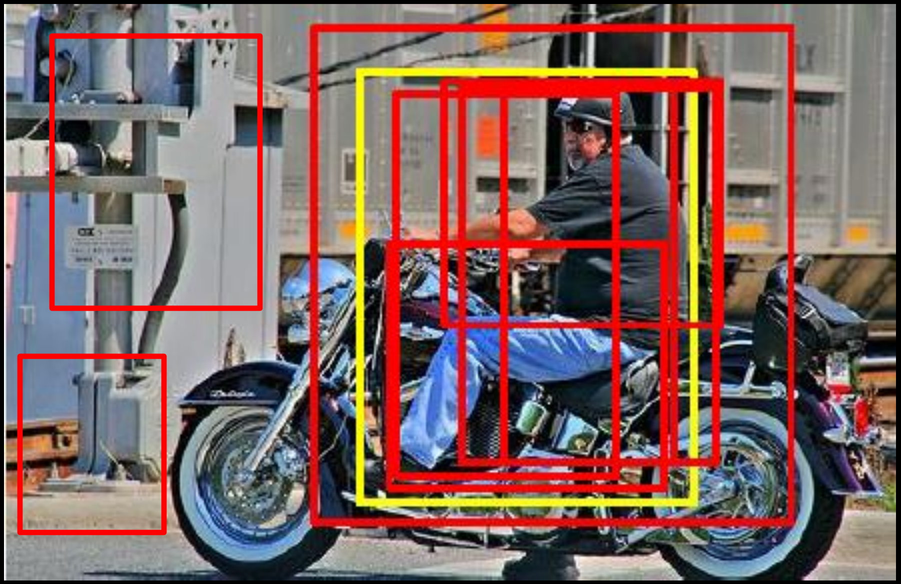






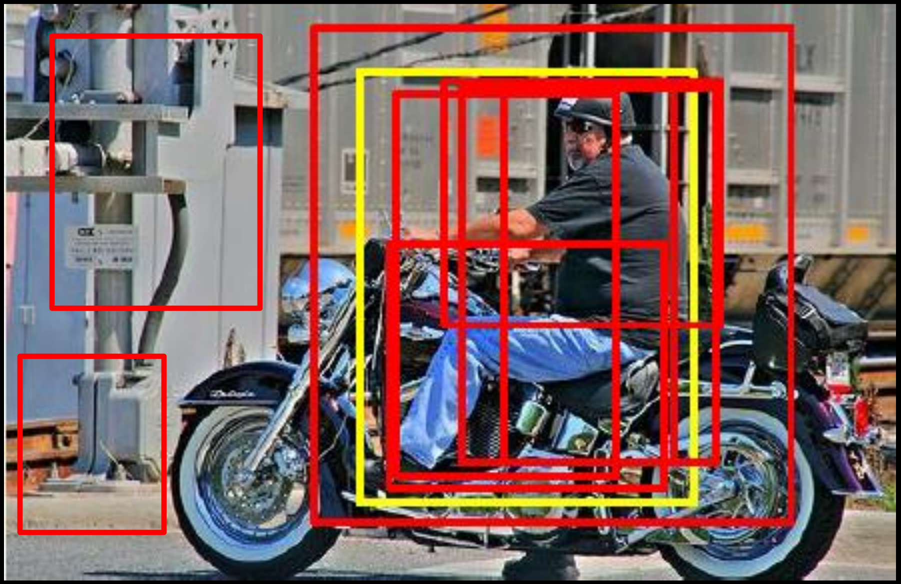
物体检测基本步骤

- 生成框
  - 生成可能有物体的框



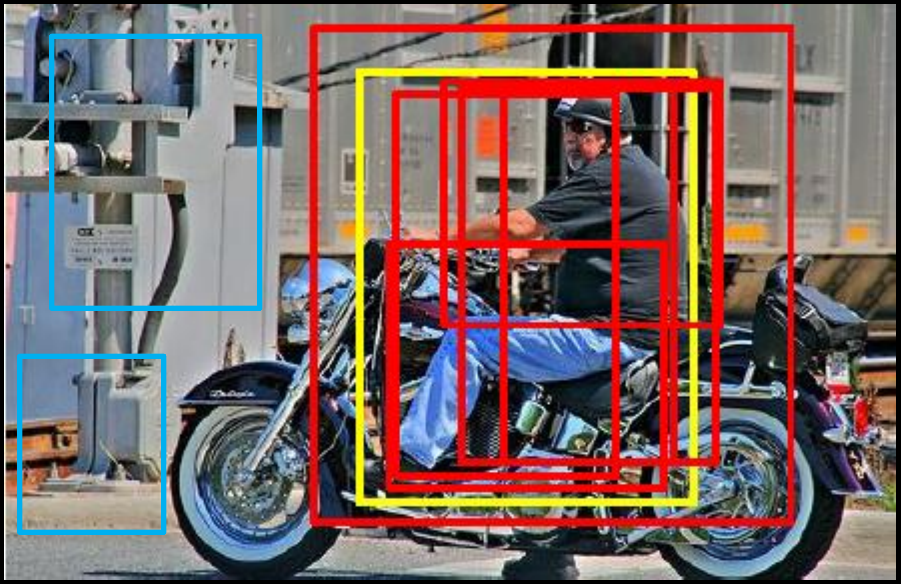
物体检测基本步骤

- 生成框
  - 生成可能有物体的框
- 分类
  - 判断这些区域是属于哪一类



物体检测基本步骤

- 生成框
  - 生成可能有物体的框
- 分类
  - 判断这些区域是属于哪一类

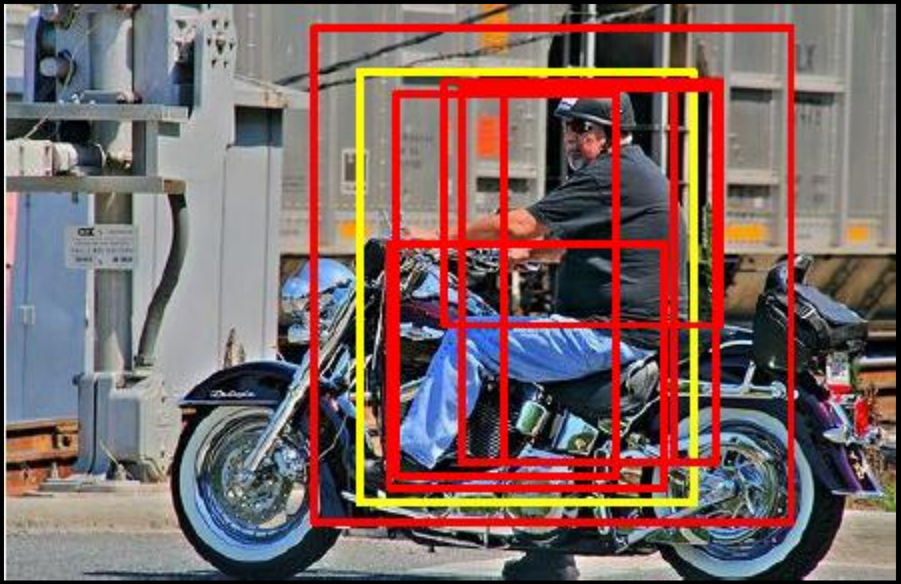


纲要

- 图像中的物体检测
  - 简要介绍
  - 基于多个上下文的框和物体关系学习 arXiv:1512.02736
  - 考虑物体长尾性质的分层级联学习 arXiv:1601.05150
  - 框生成和框分类多级级联学习
- 视频中的物体检测

基于多个上下文的框和物体关系学习

- 生成框
  - 生成可能有物体的框
- 分类
  - 判断这些区域是属于哪一类
- 不同区域得到的视觉信息不同



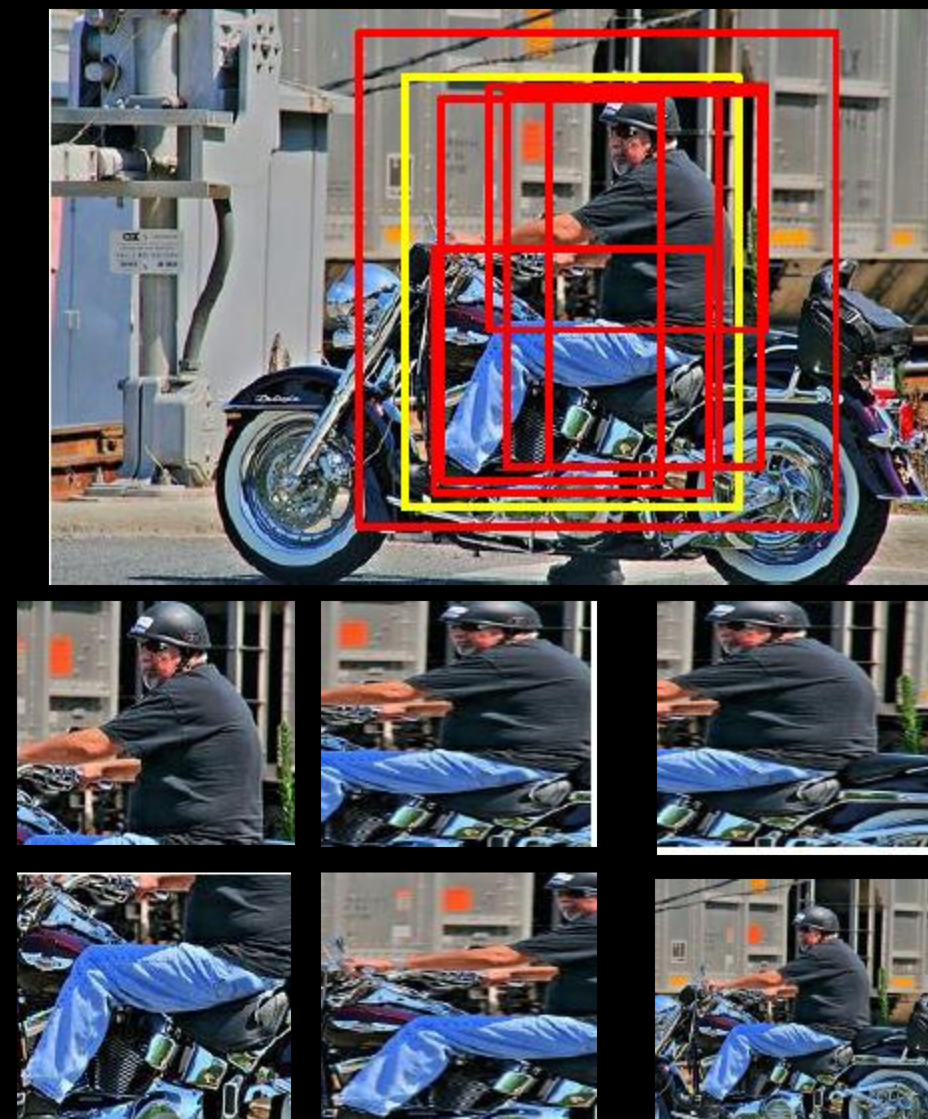


基于多个上下文的框和物体关系学习

- 生成框
  - 生成可能有物体的框
- 分类
  - 判断这些区域是属于哪一类
- 不同区域得到的视觉信息不同

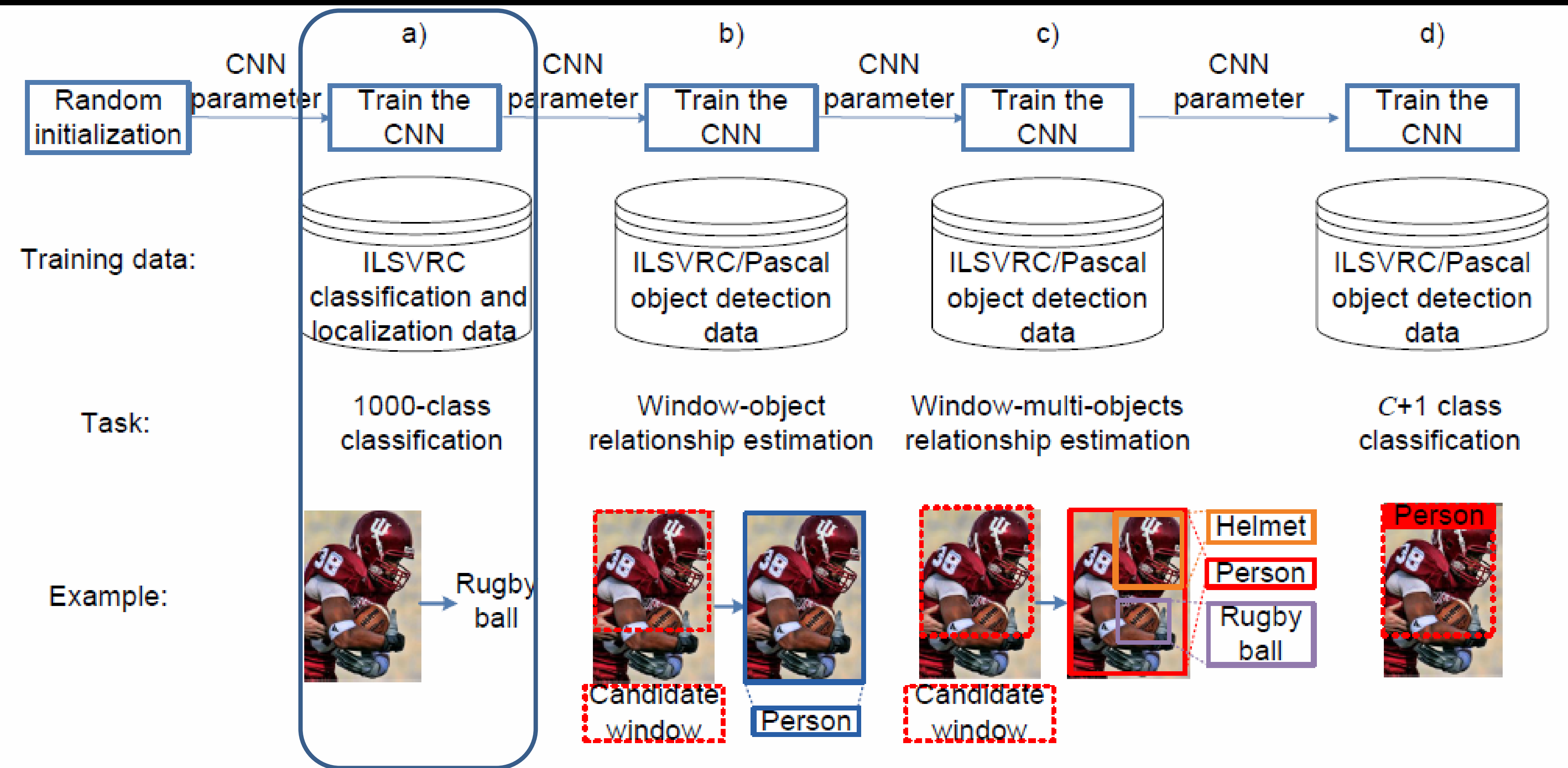
基于多个上下文的框和物体关系学习

- 生成框
  - 生成可能有物体的框
- 分类
  - 判断这些区域是属于哪一类
- 不同区域得到的视觉信息不同
  - 这些不同被忽视
  - 利用这些信息



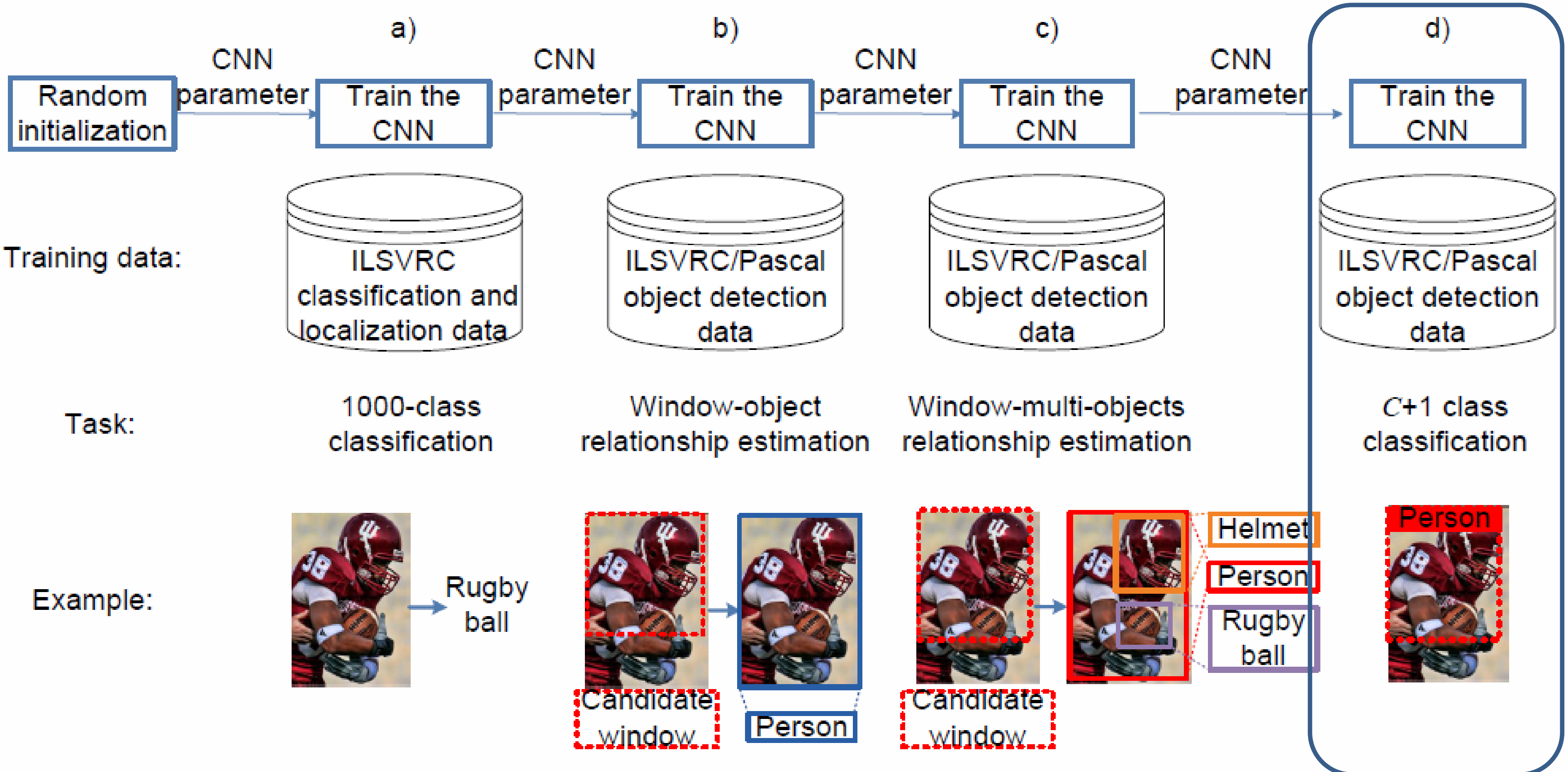
基于多个上下文的框和物体关系学习

- 1000 类ImageNet classification 数据预训练



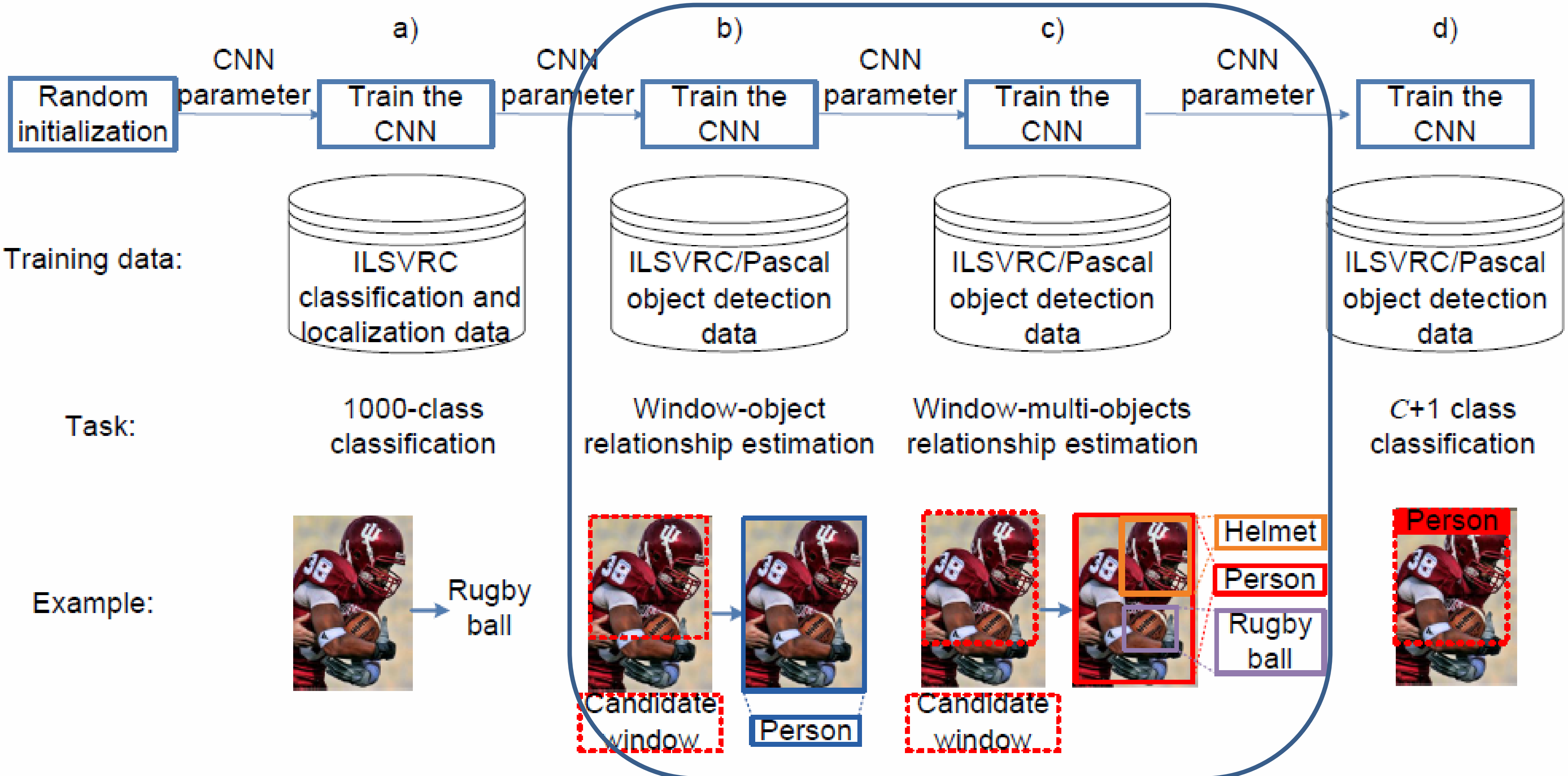
基于多个上下文的框和物体关系学习

- 学习物体检测（分类框）



基于多个上下文的框和物体关系学习

- 学习框与真实物体之间的位置关系

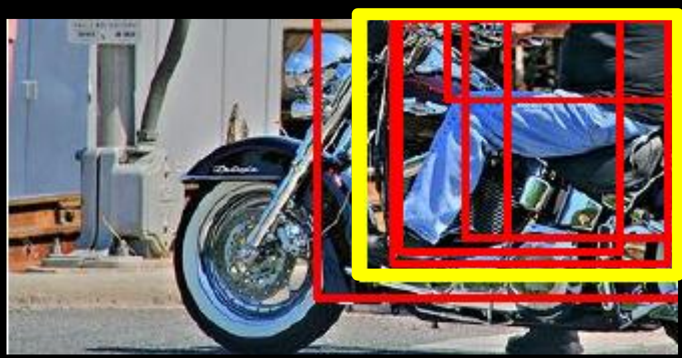


多上下文的框

- 框与真实框(ground truth)之间的关系是否正确的歧义性

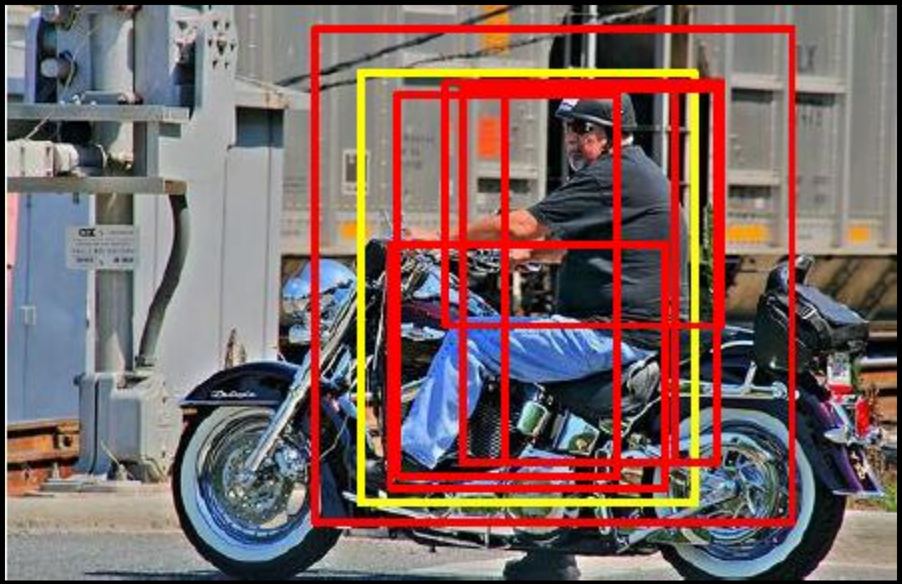
多上下文的框

- 框是否正确的歧义性



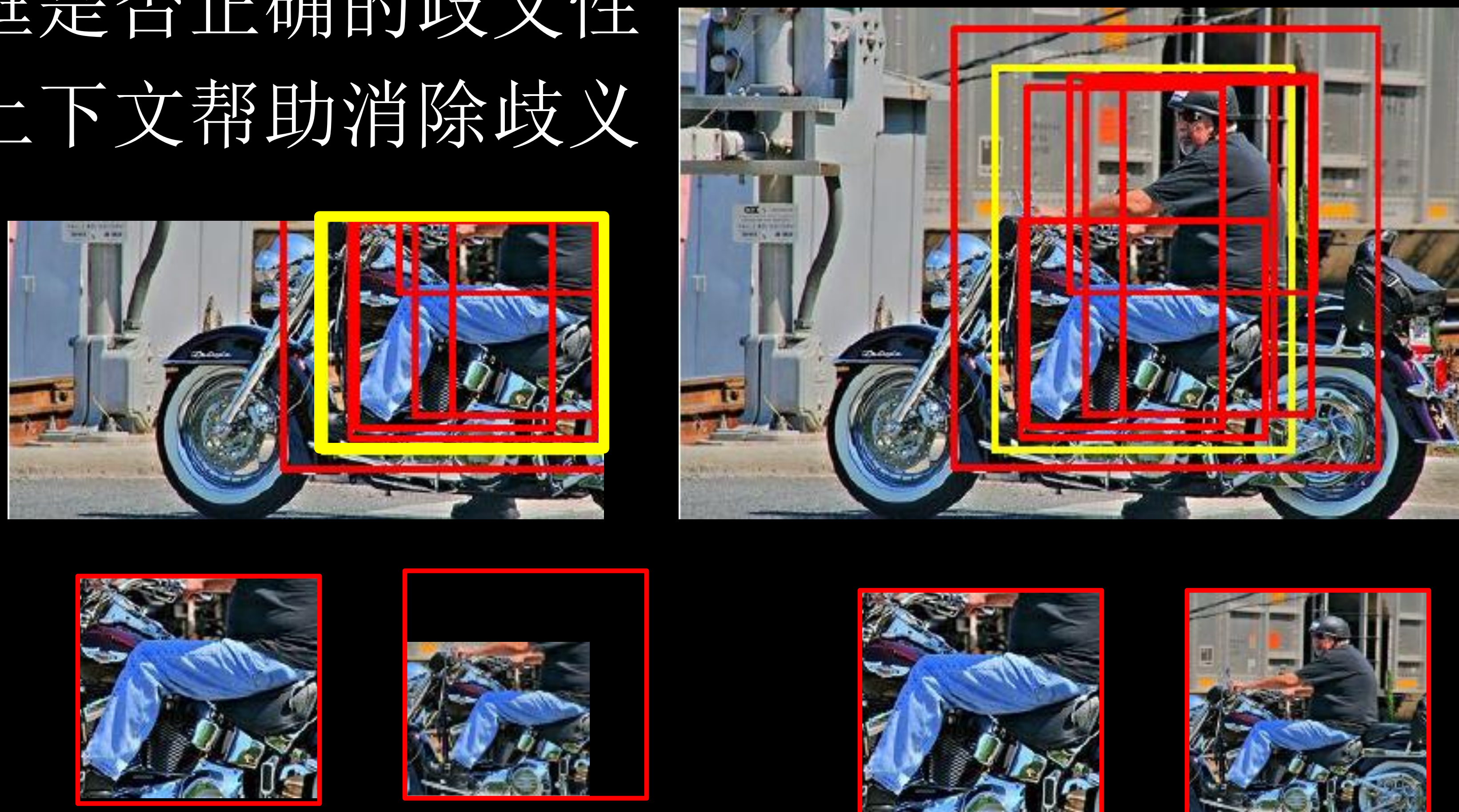
多上下文的框

- 框是否正确的歧义性



多上下文的框

- 框是否正确的歧义性
- 上下文帮助消除歧义



实验结果

GoogLeNet [1]	+上下文	+框与物体关系学习	+更好的预训练[2] +bounding box regression
39.9	42.1	46.3	49.1

[1] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. CVPR, 2015.

[2] W. Ouyang, P. Luo, X. Zeng, S. Qiu, Y. Tian, H. Li, S. Yang, Z. Wang, Y. Xiong, C. Qian, et al. Deepid-net: deformable deep convolutional neural networks for object detection. CVPR 15

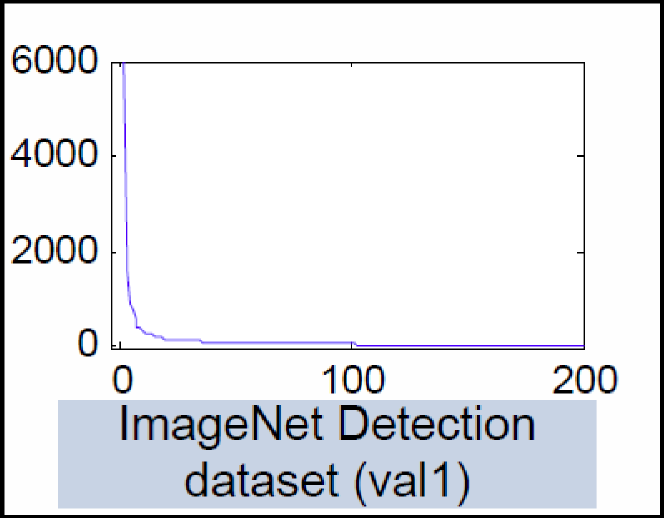
Code and model available on
www.ee.cuhk.edu.hk/~wlouyang/projects/imagenetDeepId/index.html

纲要

- 图像中的物体检测
  - 基于多个上下文的框和物体关系学习 arXiv:1512.02736
  - 考虑物体长尾性质的多层级分组级联学习 arXiv:1601.05150
  - Region proposal和框多级级联学习
- 视频中的物体检测

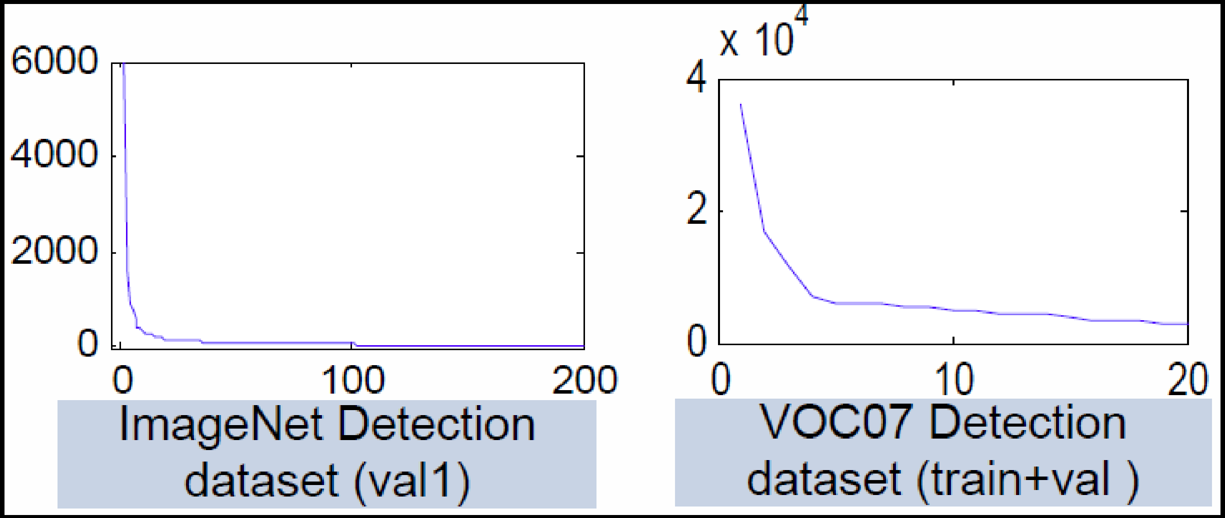
物体检测中的长尾性质

- 在物体检测中，不同类样本数目呈现长尾性质
- ImageNet val1:
  - 人(6,007) 狗(2,142) 鸟(1643)
  - 狮子(19) 蜈蚣(19) 仓鼠(16).



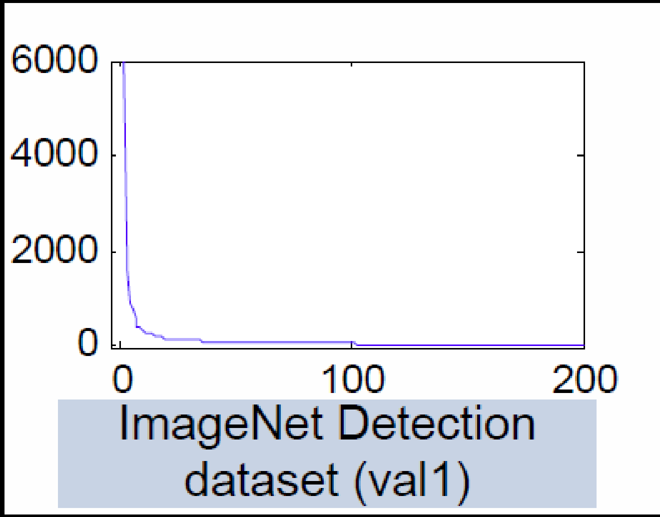
物体检测中的长尾性质

- 在物体检测中，不同类样本数目呈现长尾性质
- ImageNet val1:
  - 人(6,007) 狗(2,142) 鸟(1643)
  - 狮子(19) 蜈蚣(19) 仓鼠(16).



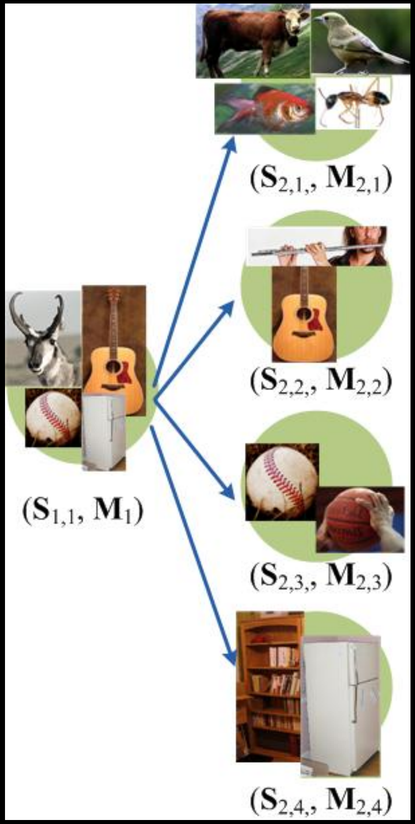
分组学习

- 物体视觉信息不同
- 物体种类太多时，深度学习在trade-off



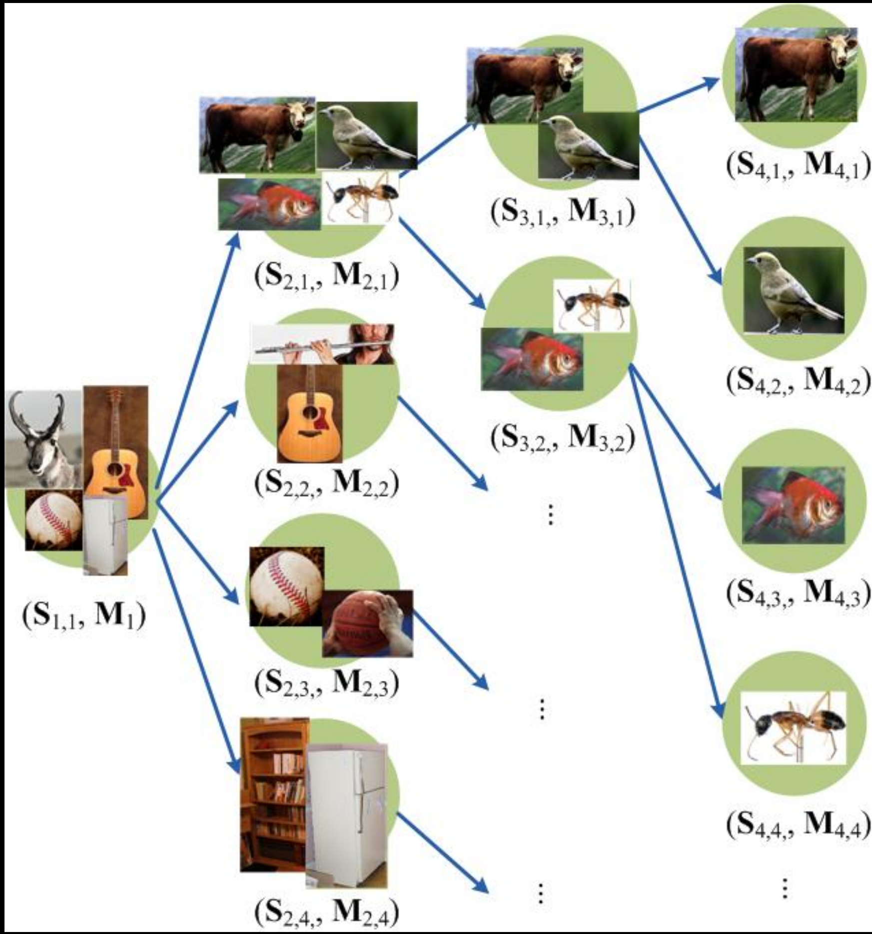
分组学习

- 物体视觉信息不同
- 物体种类太多时，深度学习在trade-off
- 将相似类别分组学习



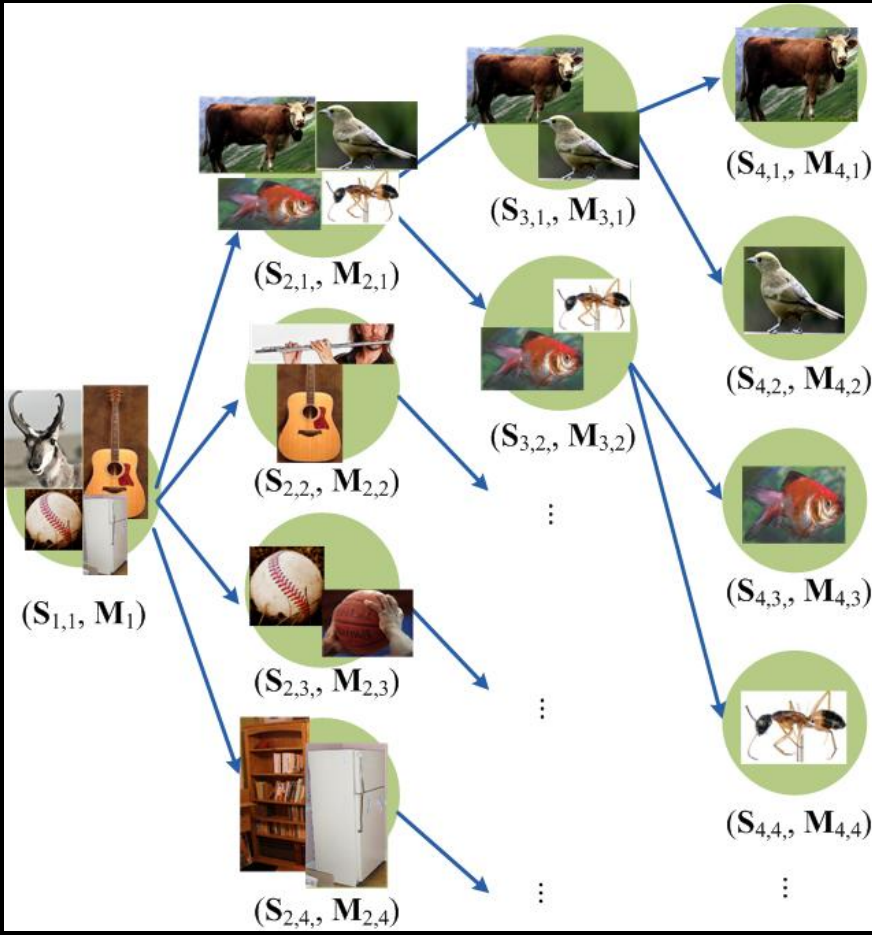
多层次分组学习

- 物体视觉信息不同
- 物体种类太多时，深度学习在trade-off
- 将相似类别分组学习

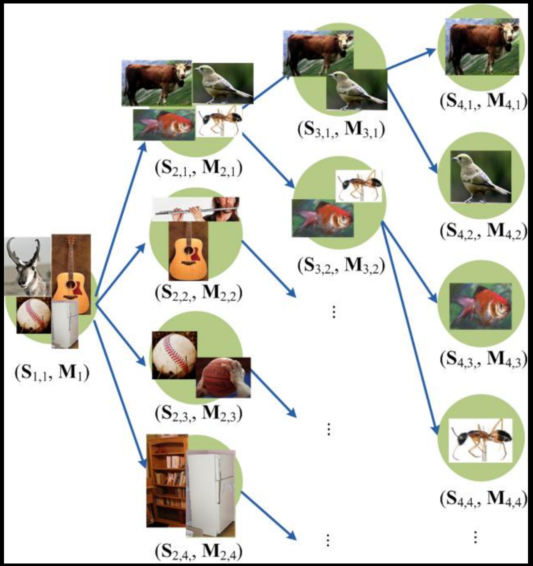


多层次分组学习

- 物体视觉信息不同
- 物体种类太多时，深度学习在trade-off
- 将相似类别分组
- 利用相似性做多层级联(cascasde)以提速



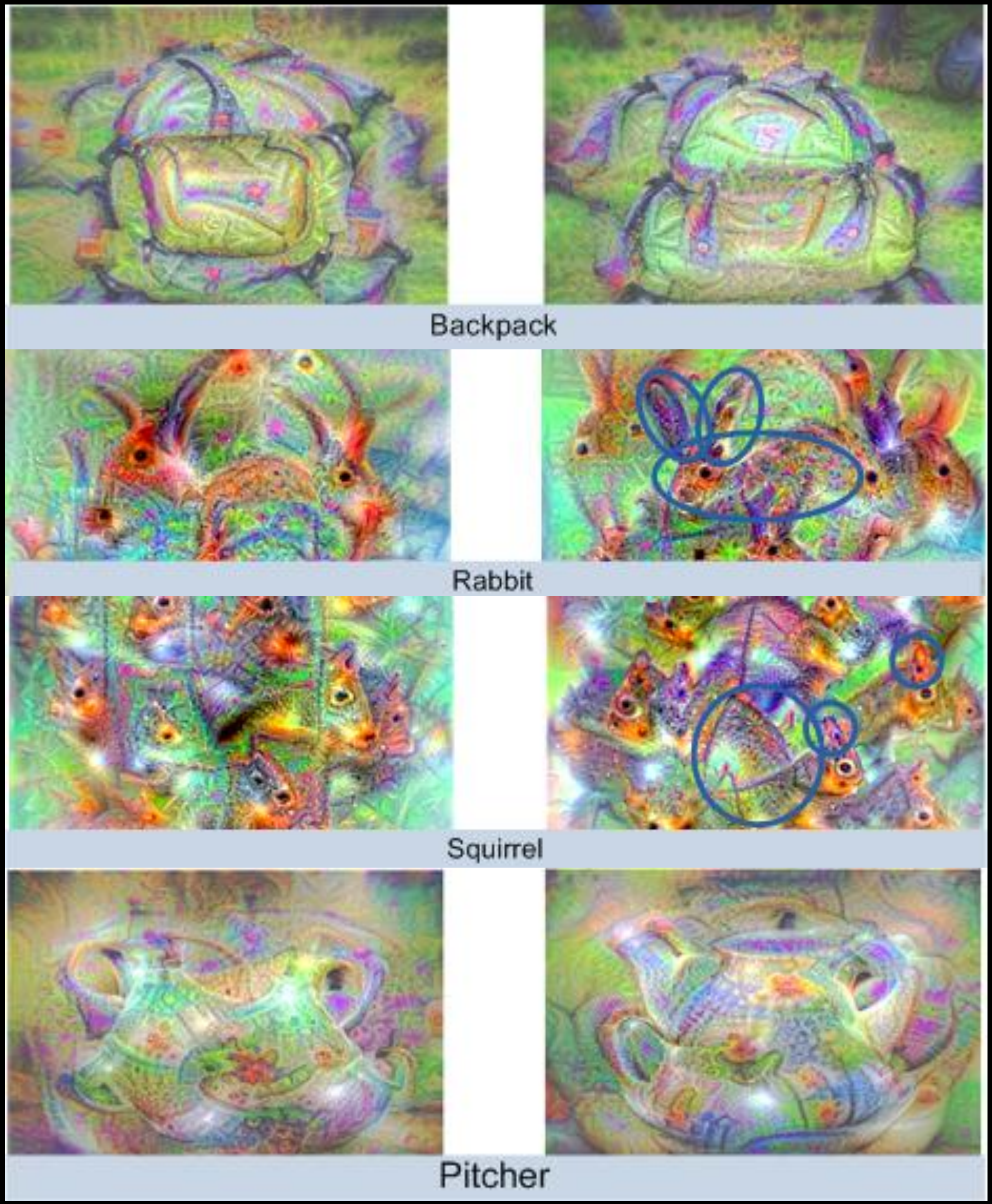
实验结果



层级数	1	2	3	4	新结果
分组数	1	4	7	18	7
每组内平均类别数目	200	50	29	11	29
级联后每组所需考虑的框数	136	25.8	15.2	5.6	
mAP	40.3	41.3	42.5	45	56

Code and model available on
www.ee.cuhk.edu.hk/~wlouyang/projects/imagenetDeepId/index.html

实验结果

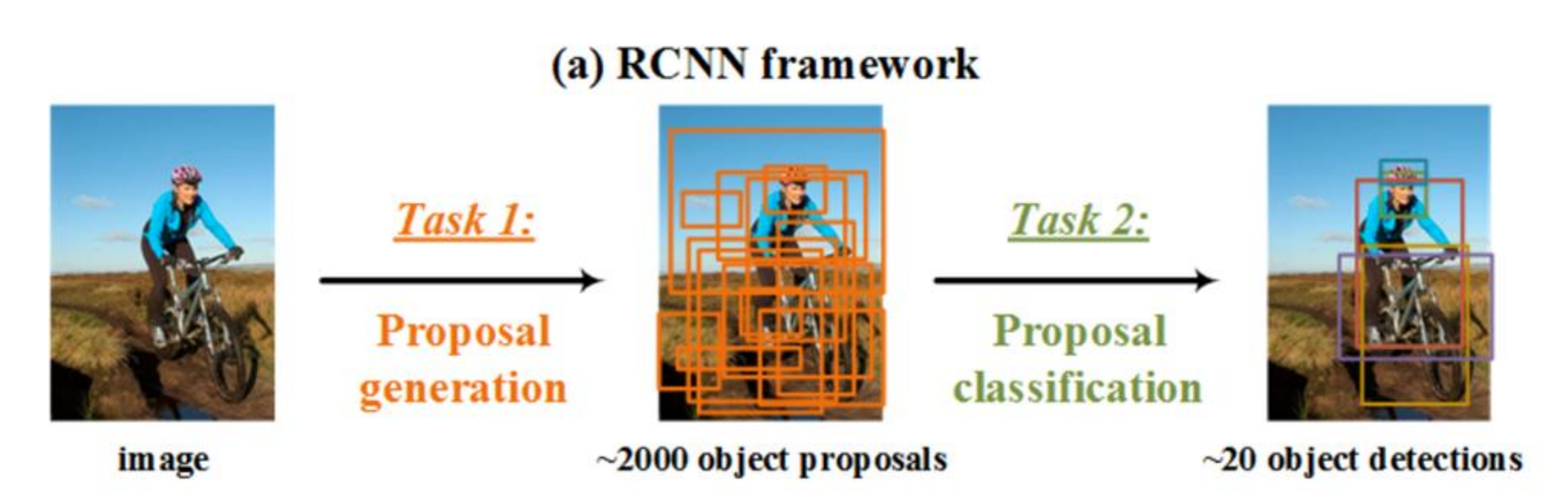


纲要

- 图像中的物体检测
  - 基于多个上下文的框和物体关系学习 arXiv:1512.02736
  - 考虑物体长尾性质的分层级联学习 arXiv:1601.05150
  - 框生成和框分类多级级联学习
- 视频中的物体检测

物体检测两步

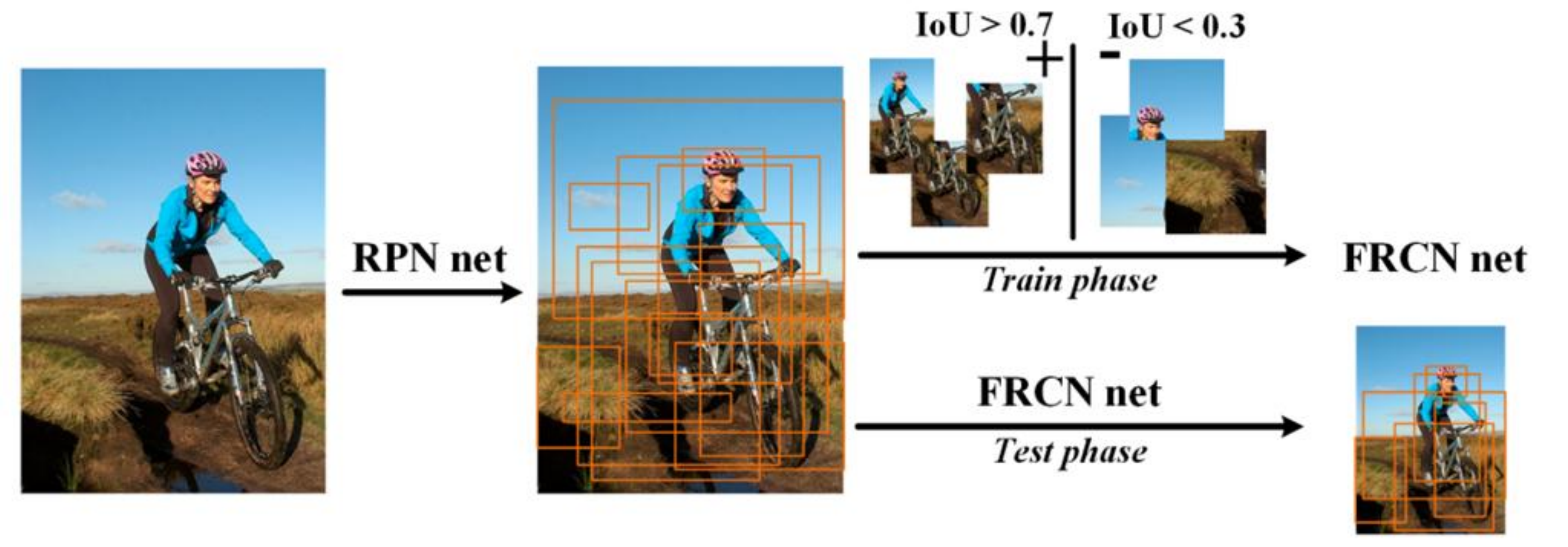
- 生成框(proposal generation/region proposal)
- 对框进行分类 (proposal classification)



Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." NIPS. 2015.

物体检测两步级联

- 生成框(proposal generation/region proposal)
  - 对生成框的深度模型进行级联



Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." NIPS. 2015.

生成框质量实验结果

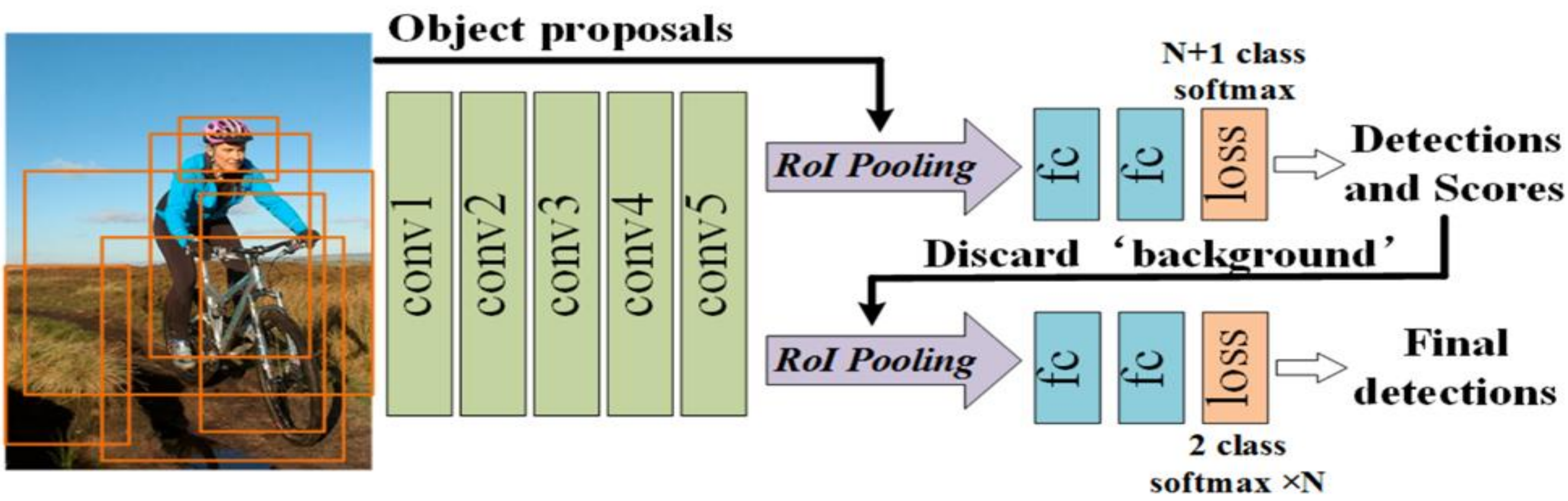
- Selective search 2000 框

Setting	Number of proposals	Recall (%)
Selective Search	2000	92.09
RPN-1 [1]	300	89.94
RPN-2	300	91.83
RPN+FRCN	300	92.38
SS+RPN+FRCN	300	94.13

[1] Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." NIPS. 2015.

物体检测两步级联

- 生成框(proposal generation/region proposal)
- 框分类 (proposal classification)
  - 对框分类的深度模型进行级联



物体检测实验结果

Results on VOC07

Setting	mAP(%)
No cascade	65.0
Single-class re-score	63.5
Multi-class re-score	**68.0**

ILSVRC14 val2

Setting	mAP(%)
GoogLeNet _BN	47.0
Cascade GoogLeNet BN	48.5
Improvement	+1.5

总结

- 设计深度学习方法使得模型更有效
- 思考物体检测存在的问题
  – 框的标签单一，学习框与物体间的关系
  – 长尾，分层级联学习
  – 框生成和框分类的不匹配，多层级联，磨合不匹配
  – 使预训练(pretraining)和微调(fine-tuning)匹配[a]
  – 使得深度模型学习物体形变 [a]

[a] Ouyang, W., Wang, X., Zeng, X., Qiu, S., Luo, P., Tian, Y., ... & Tang, X. Deepid-net: Deformable deep convolutional neural networks for object detection. In *CVPR* 2015.

Multimedia Laboratory

Object Detection in Videos with Tubelets and Multi-context Cues

CUvideo Team
The Chinese University of Hong Kong

Wanli Ouyang

Kai Kang

Junjie Yan

Xingyu Zeng

Hongsheng Li

Bin Yang

Tong Xiao

Cong Zhang

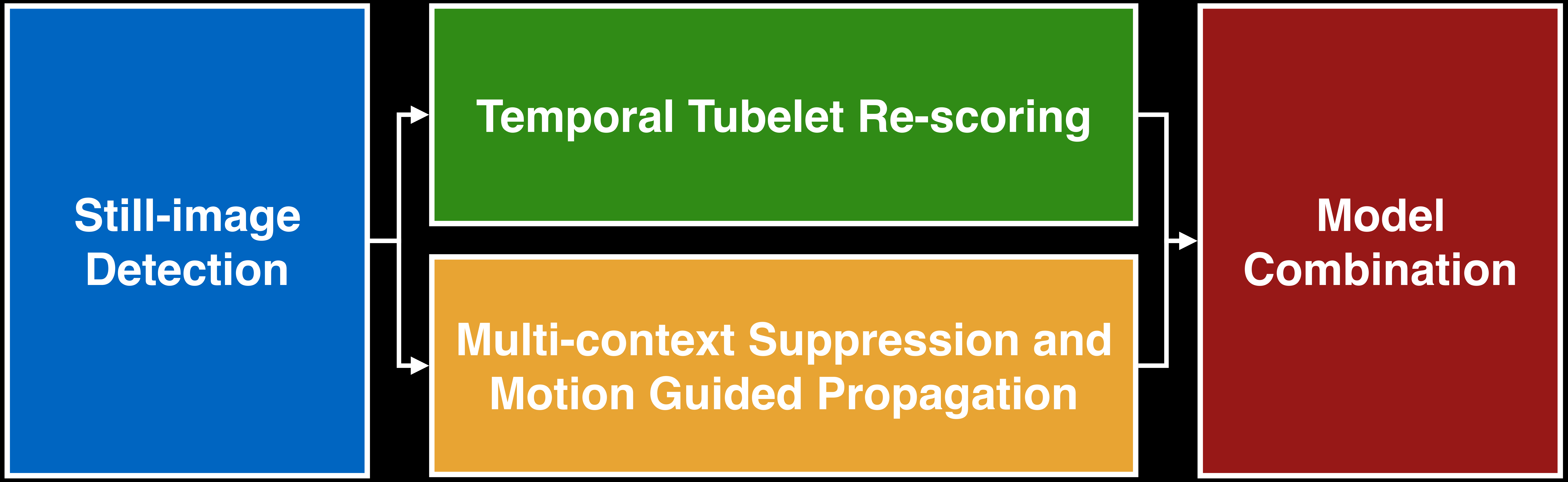
Zhe Wang

Ruohui Wang

Xiaogang Wang

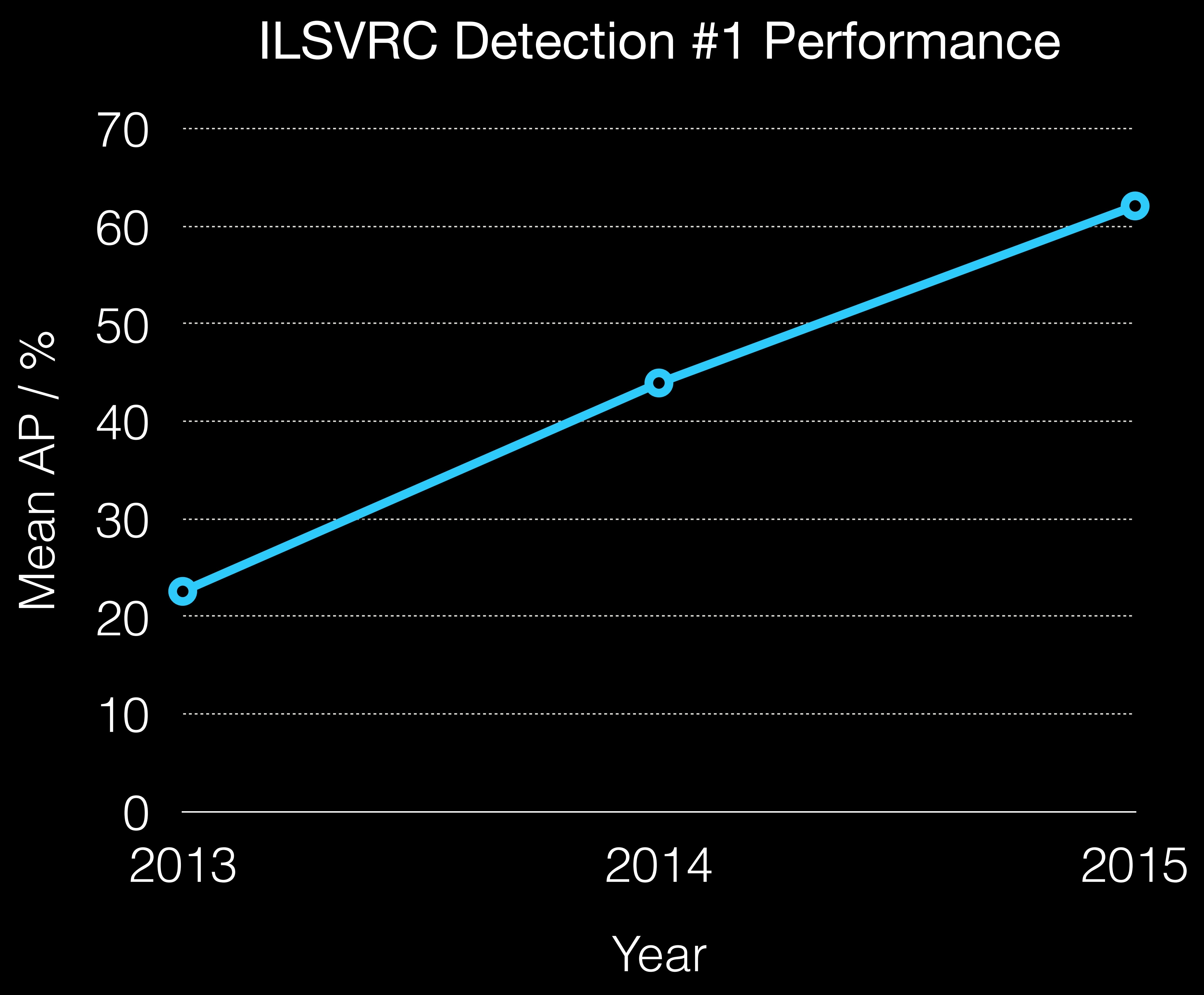
Proposed Framework

Proposed Framework

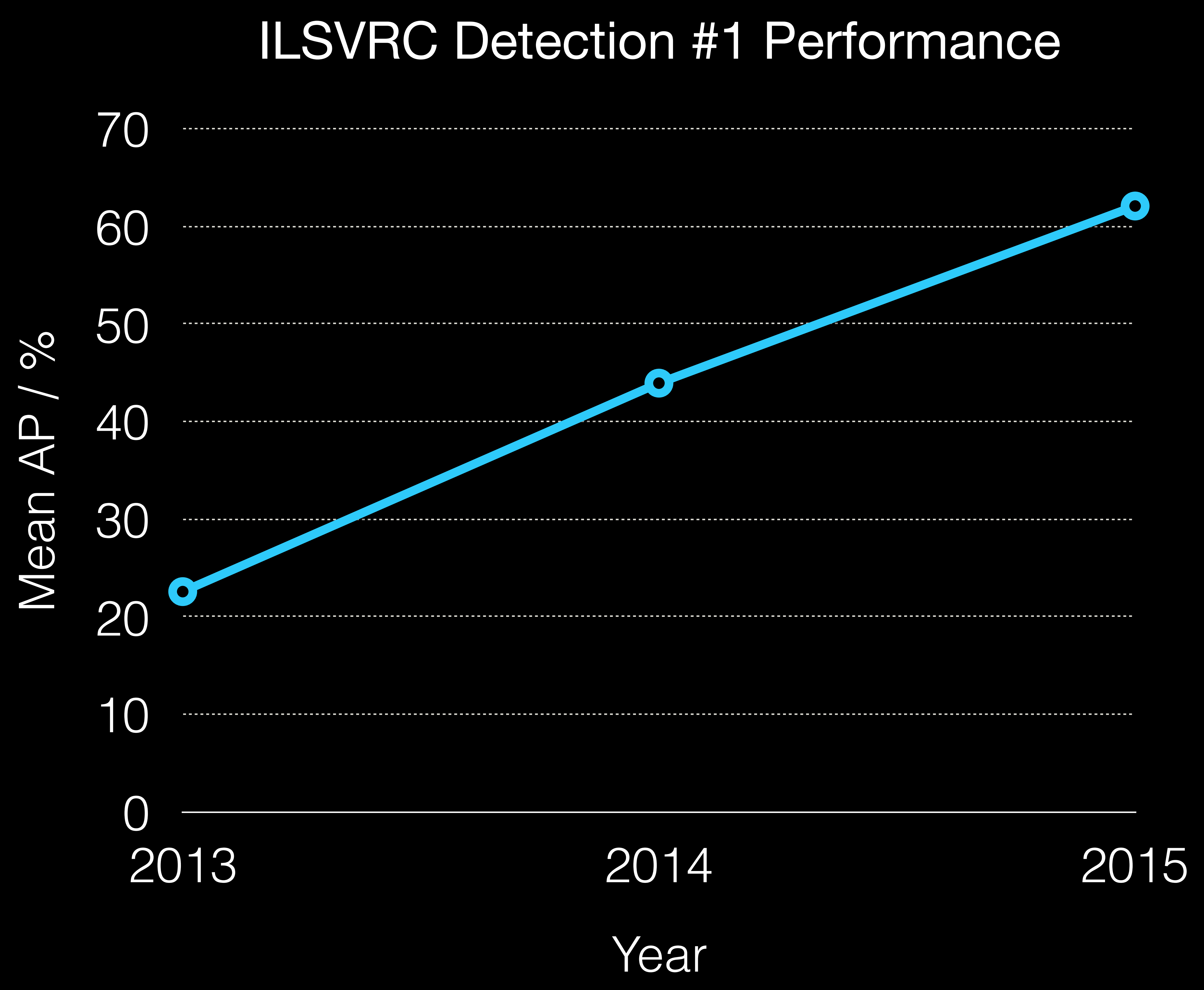


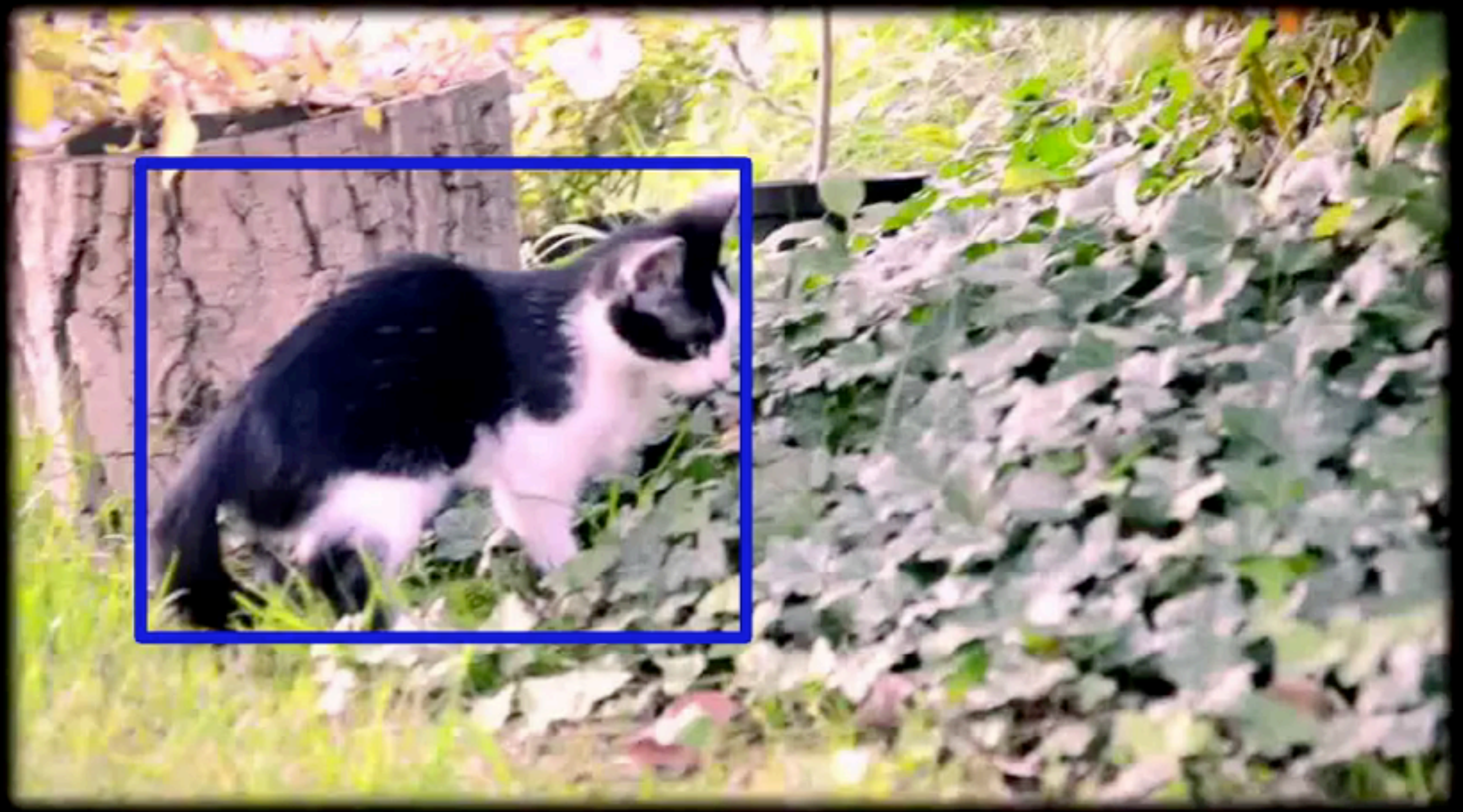
Still-image Detection

Still-image Detection



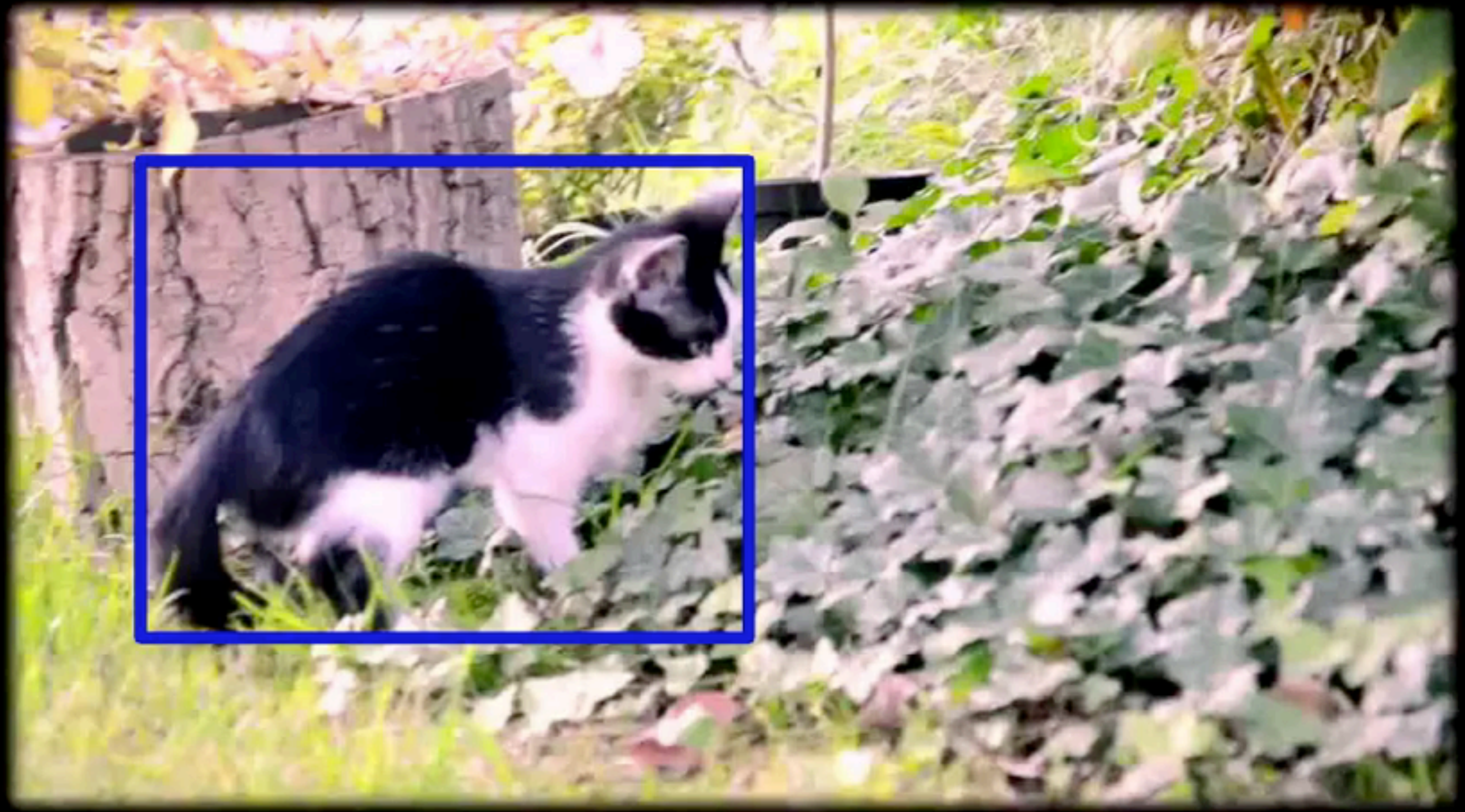
Still-image Detection

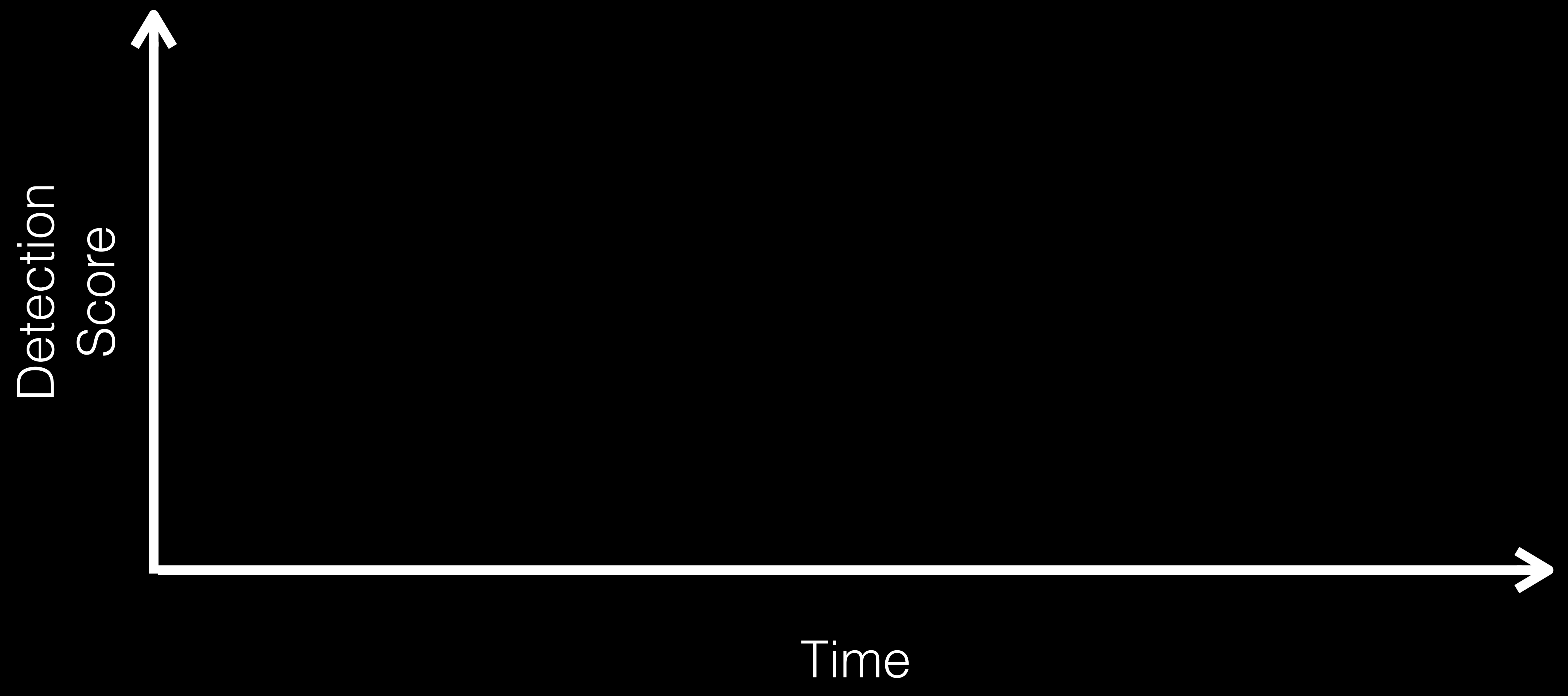




Still-image Detection: Limitation I

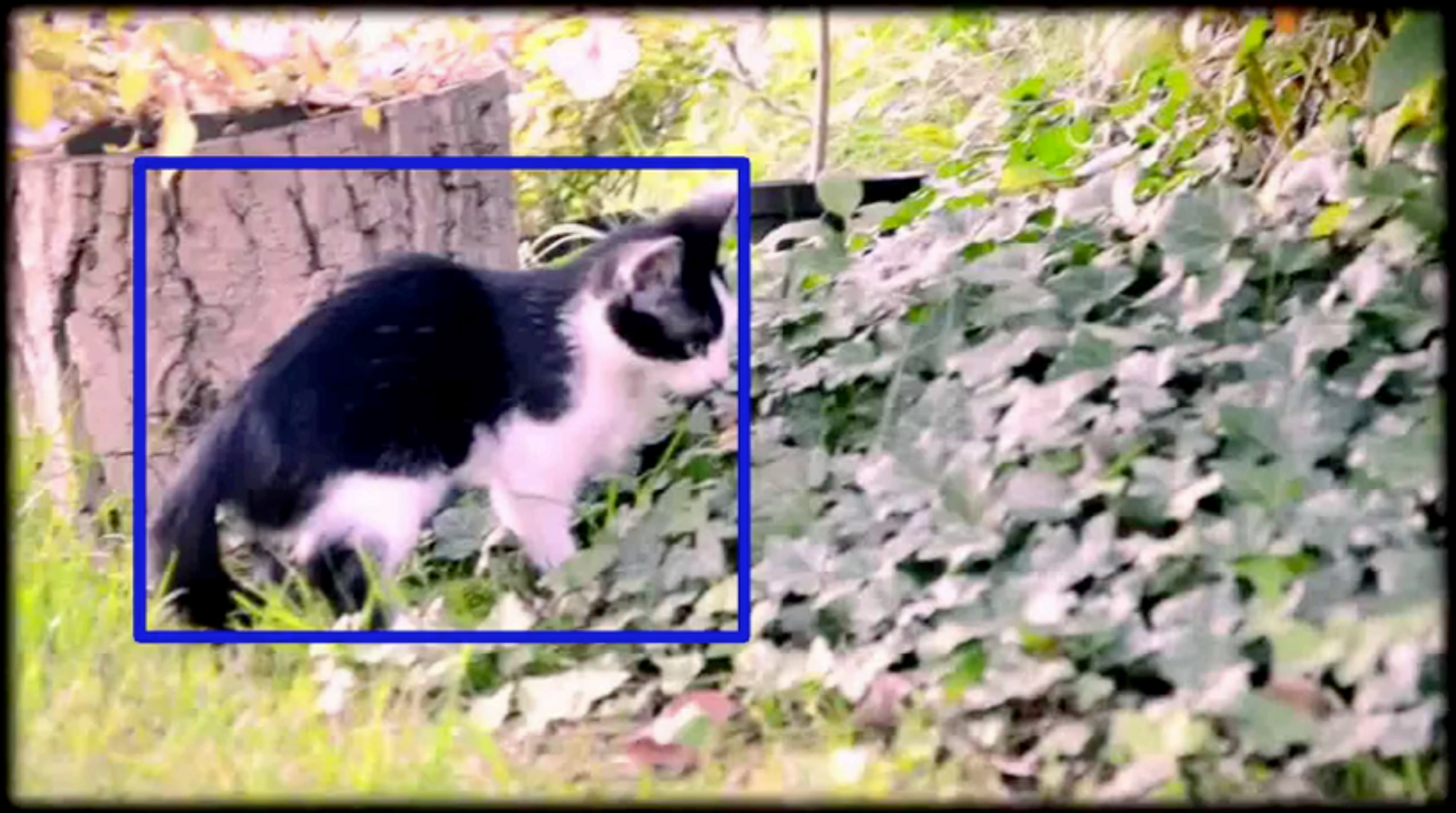
Large Temporal Variations

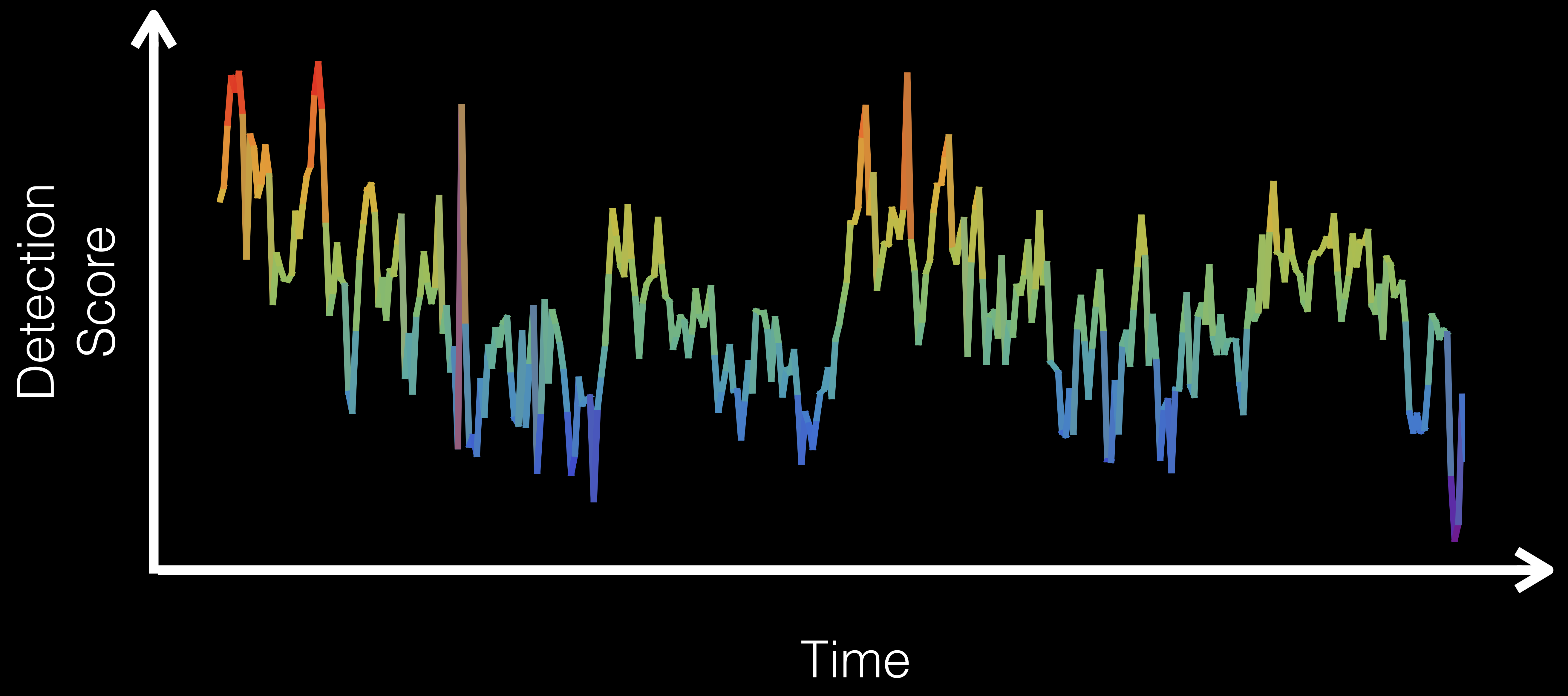




Still-image Detection: Limitation I

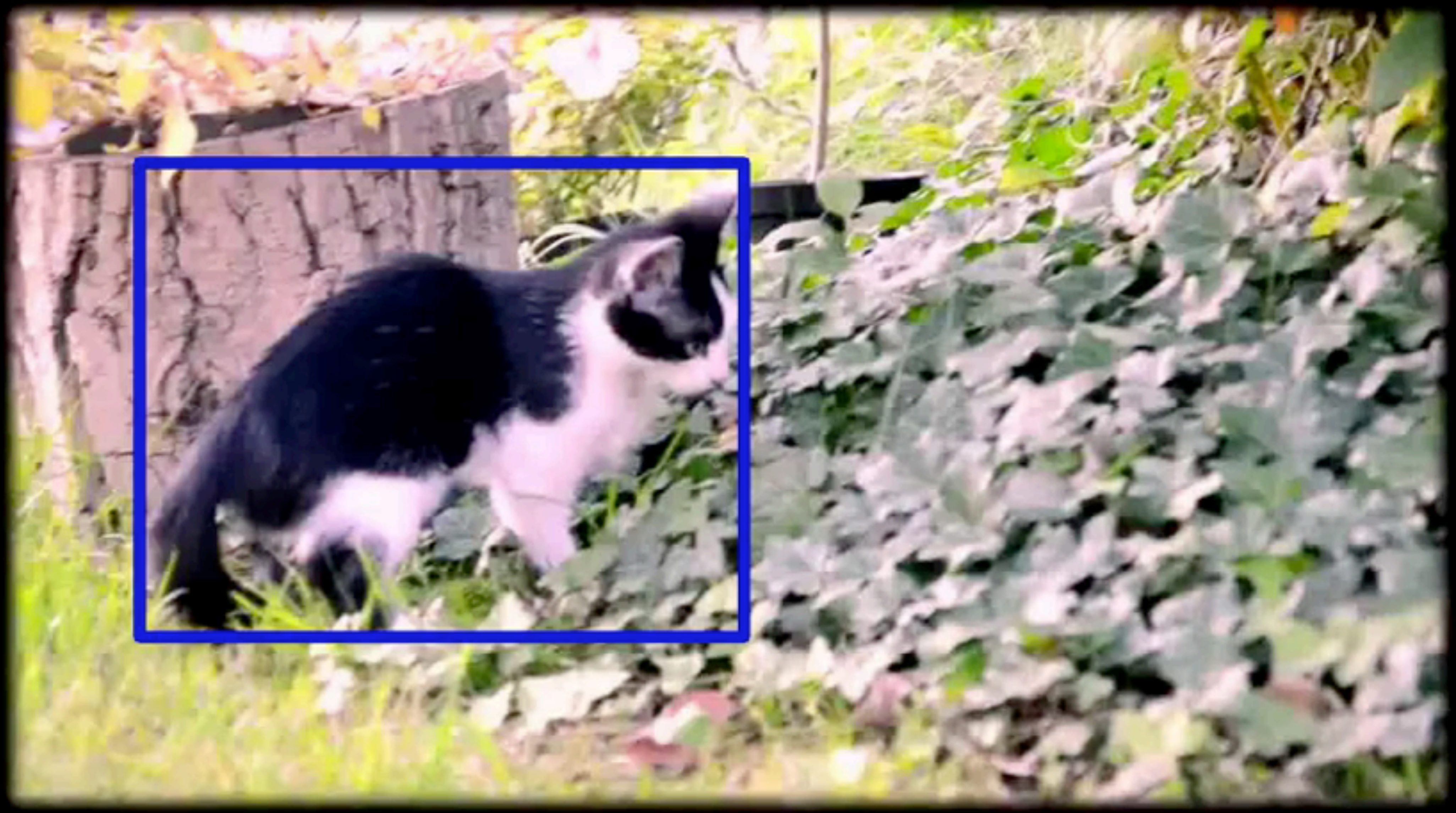
Large Temporal Variations

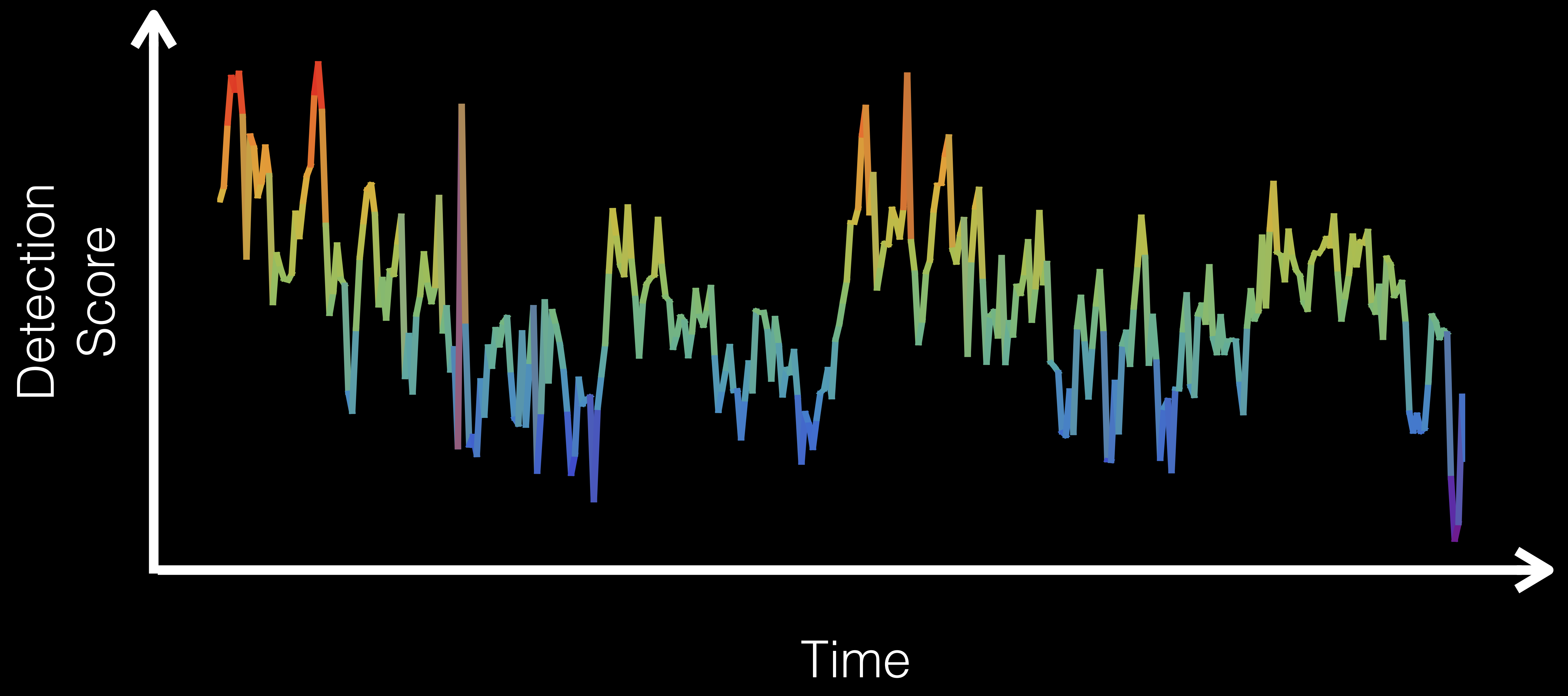




Still-image Detection: Limitation I

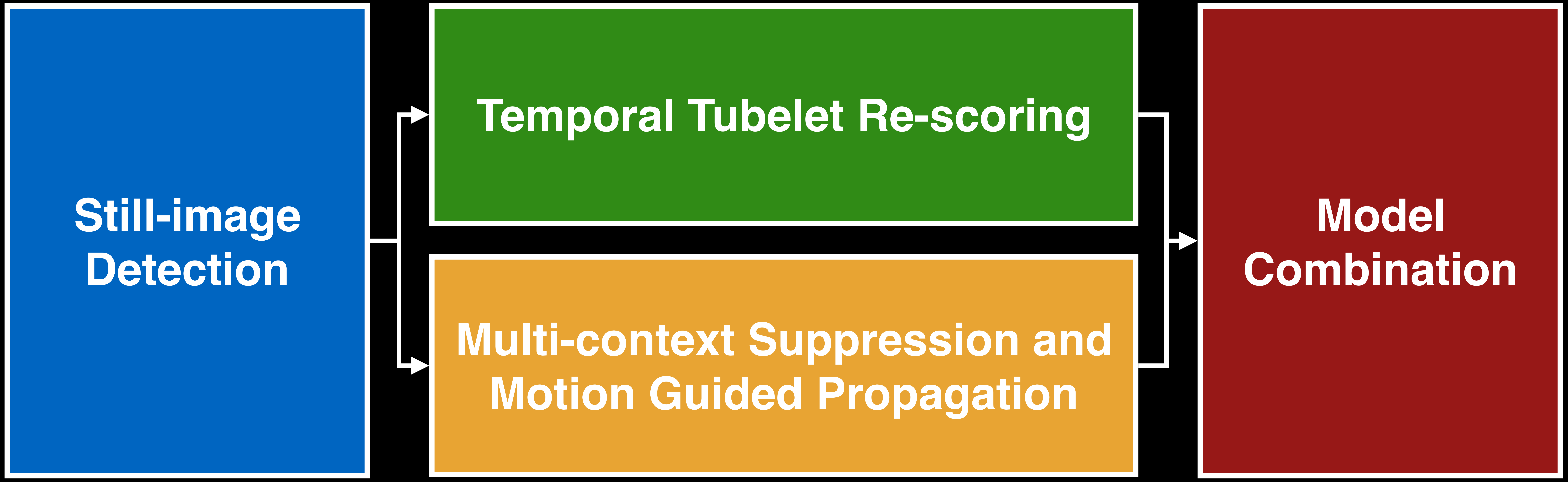
Large Temporal Variations



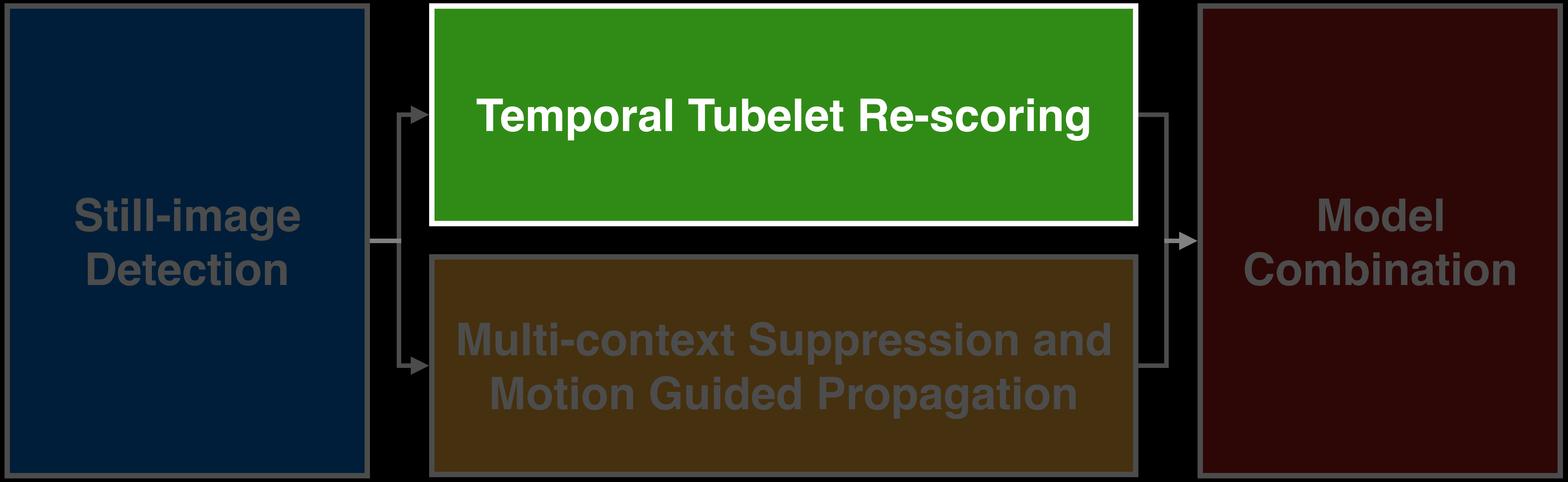


Solution - Tubelets

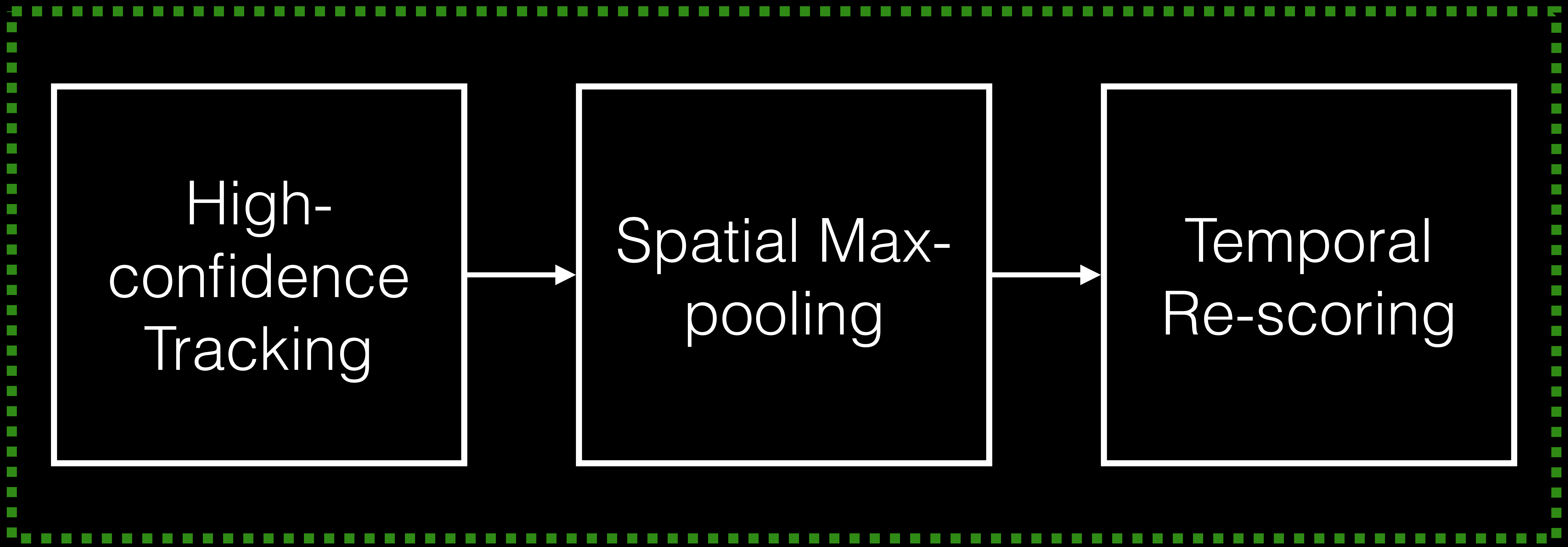
Proposed Framework



Proposed Framework

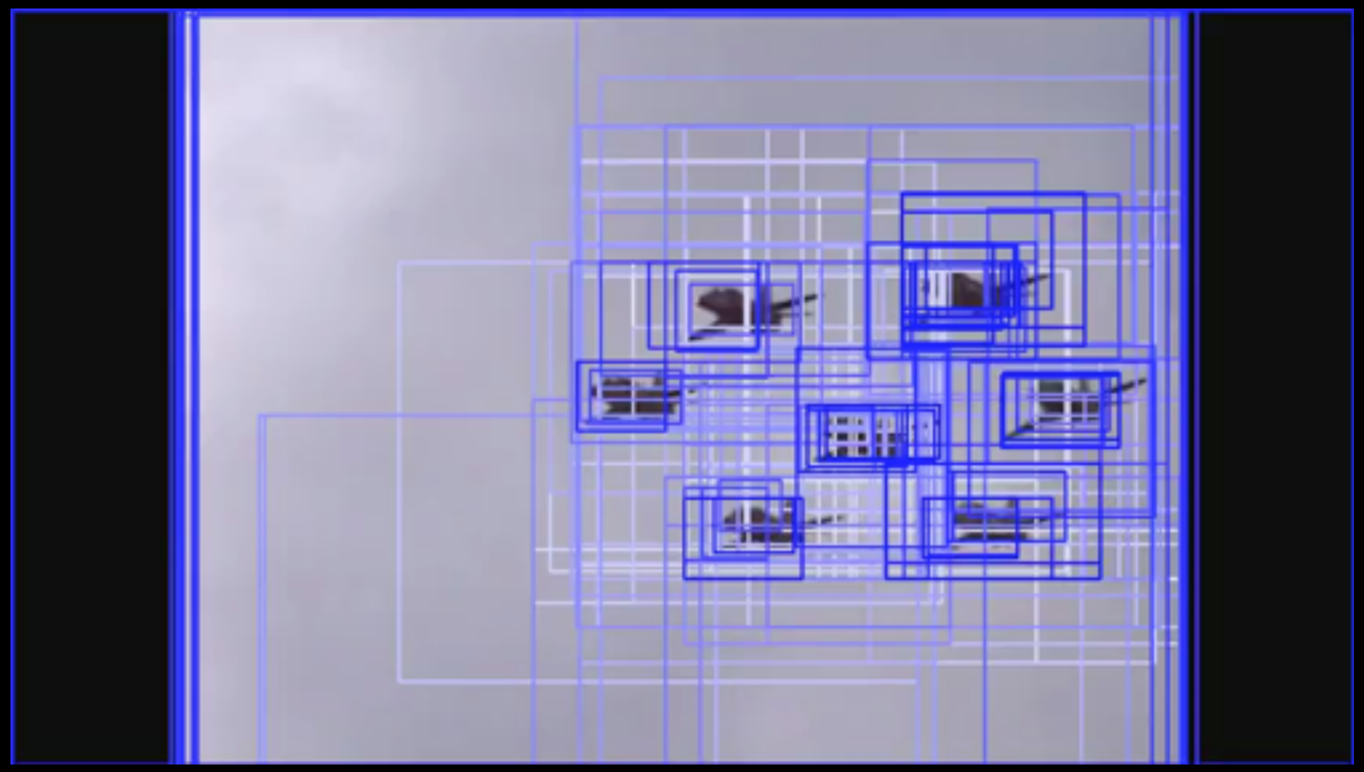


Temporal Tubelet Re-scoring



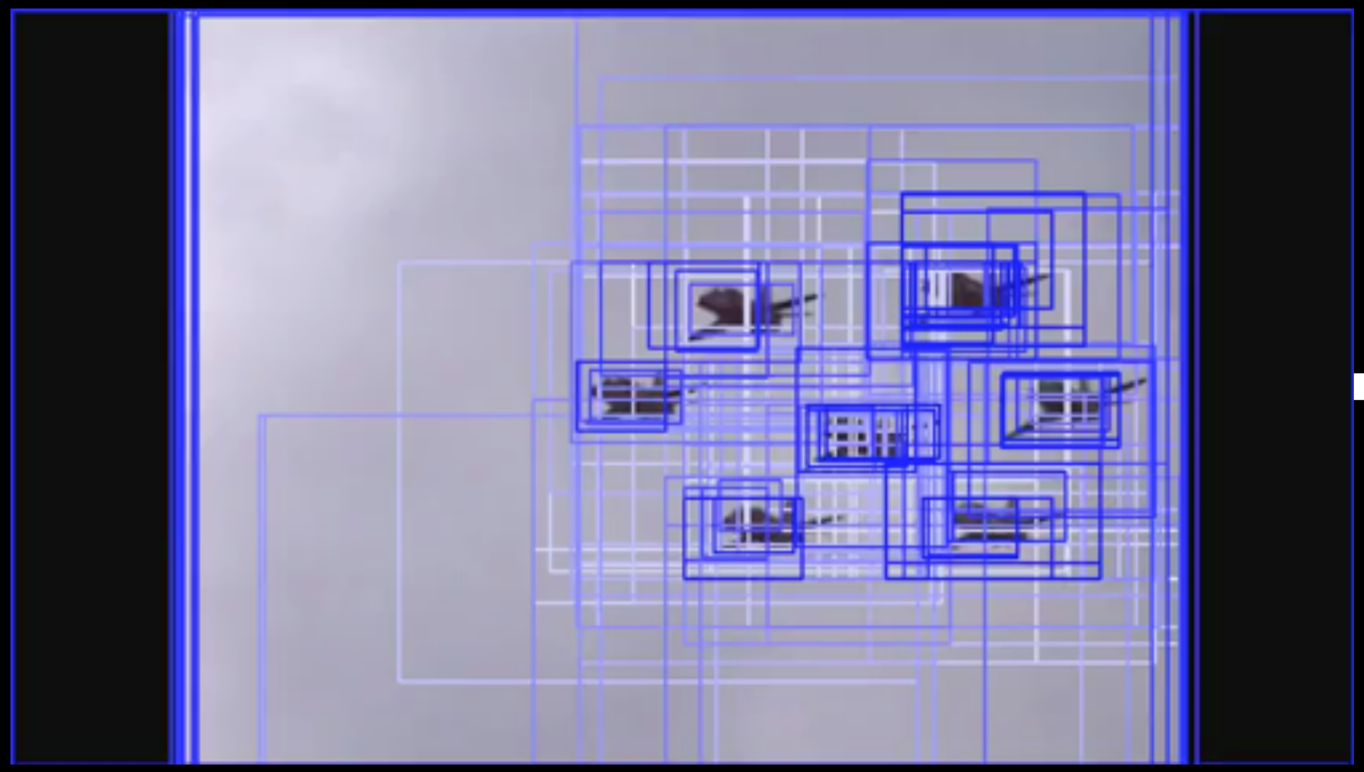
High-confidence Tracking

High-confidence Tracking



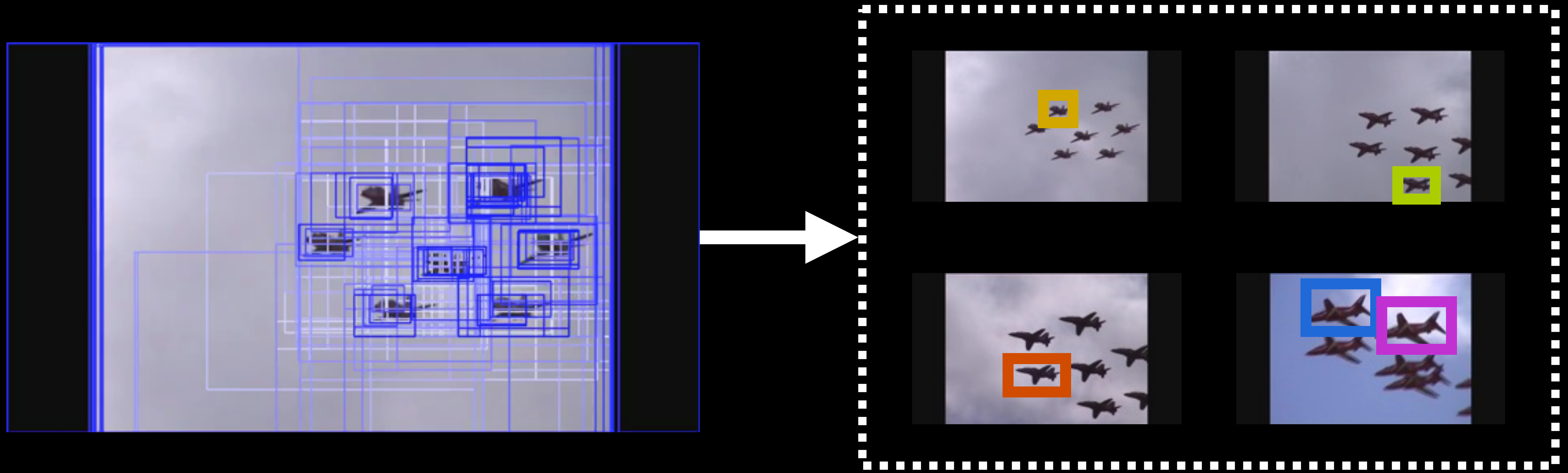
- Obtain detection results from still-image detectors

High-confidence Tracking



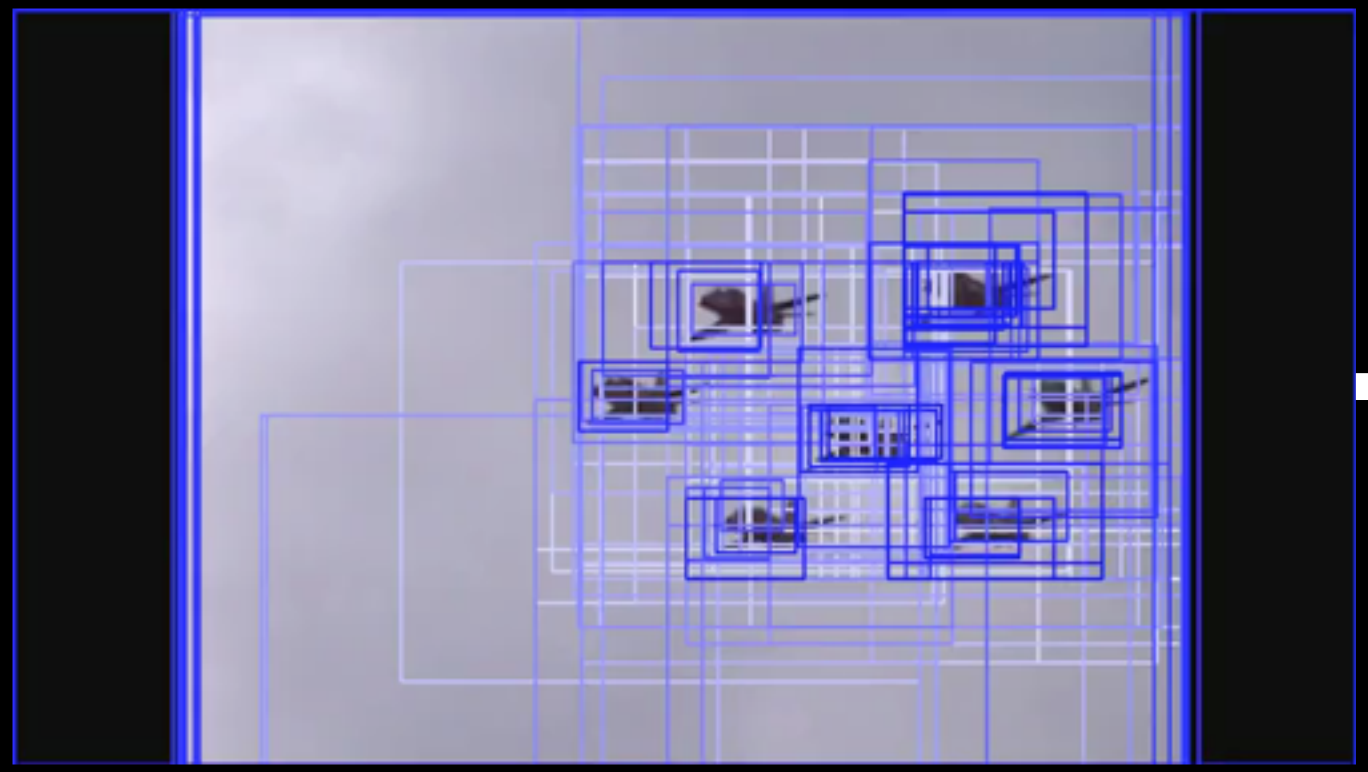
- Obtain detection results from still-image detectors

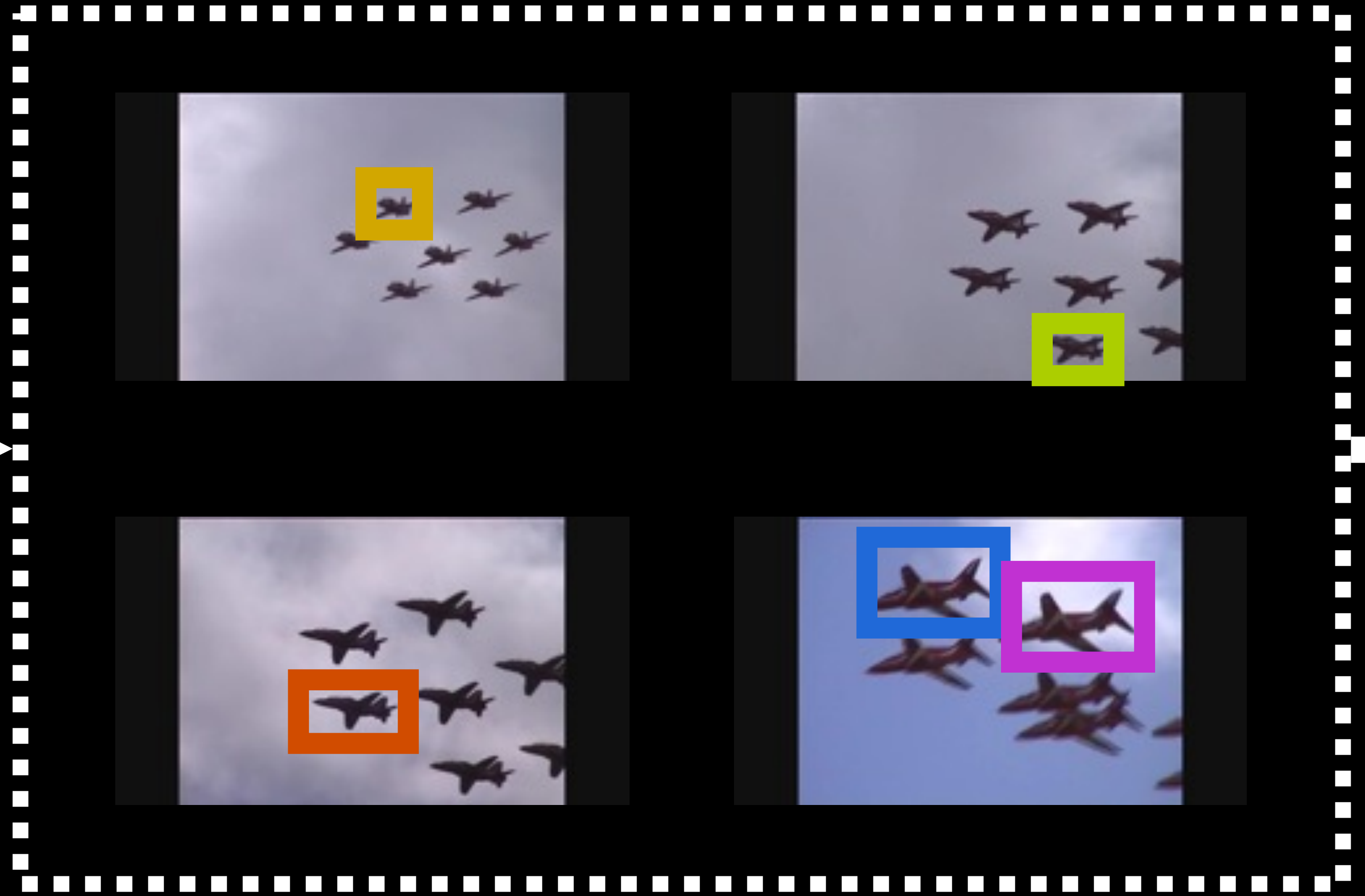
High-confidence Tracking



- Obtain detection results from still-image detectors
- Choose high-confidence detections as starting points (anchors) for tracking

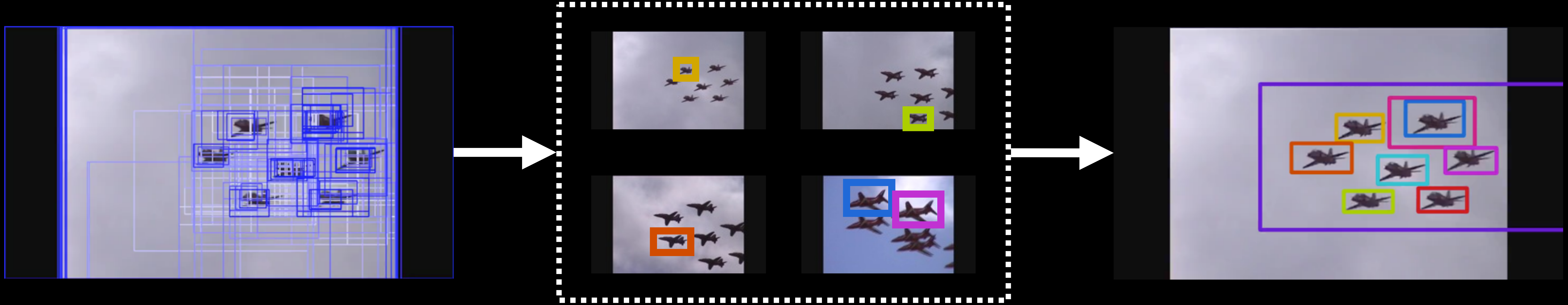
High-confidence Tracking





- Obtain detection results from still-image detectors
- Choose high-confidence detections as starting points (anchors) for tracking

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?

Spatial Max-pooling: Why?

- The detection scores on the tracked tubelets are not satisfactory

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**

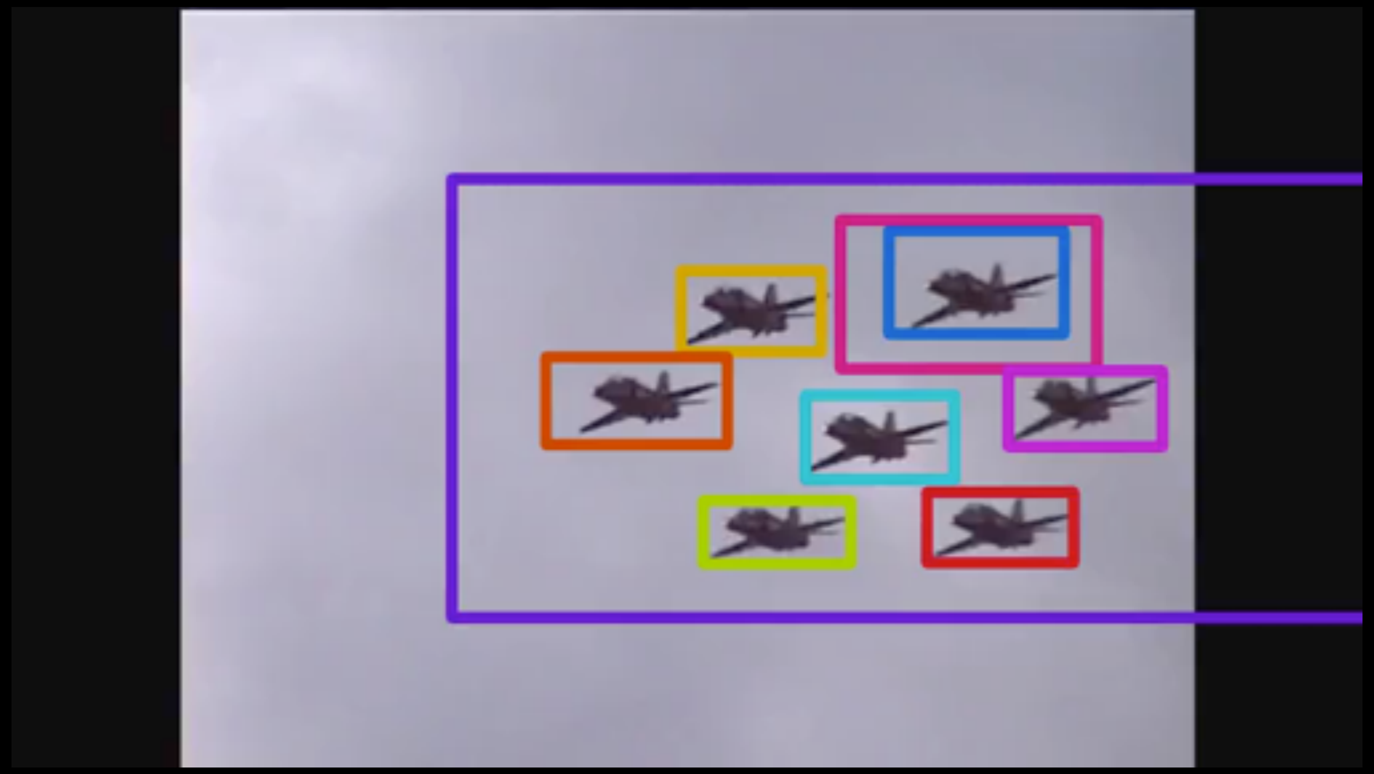
Spatial Max-pooling: Why?

- 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**

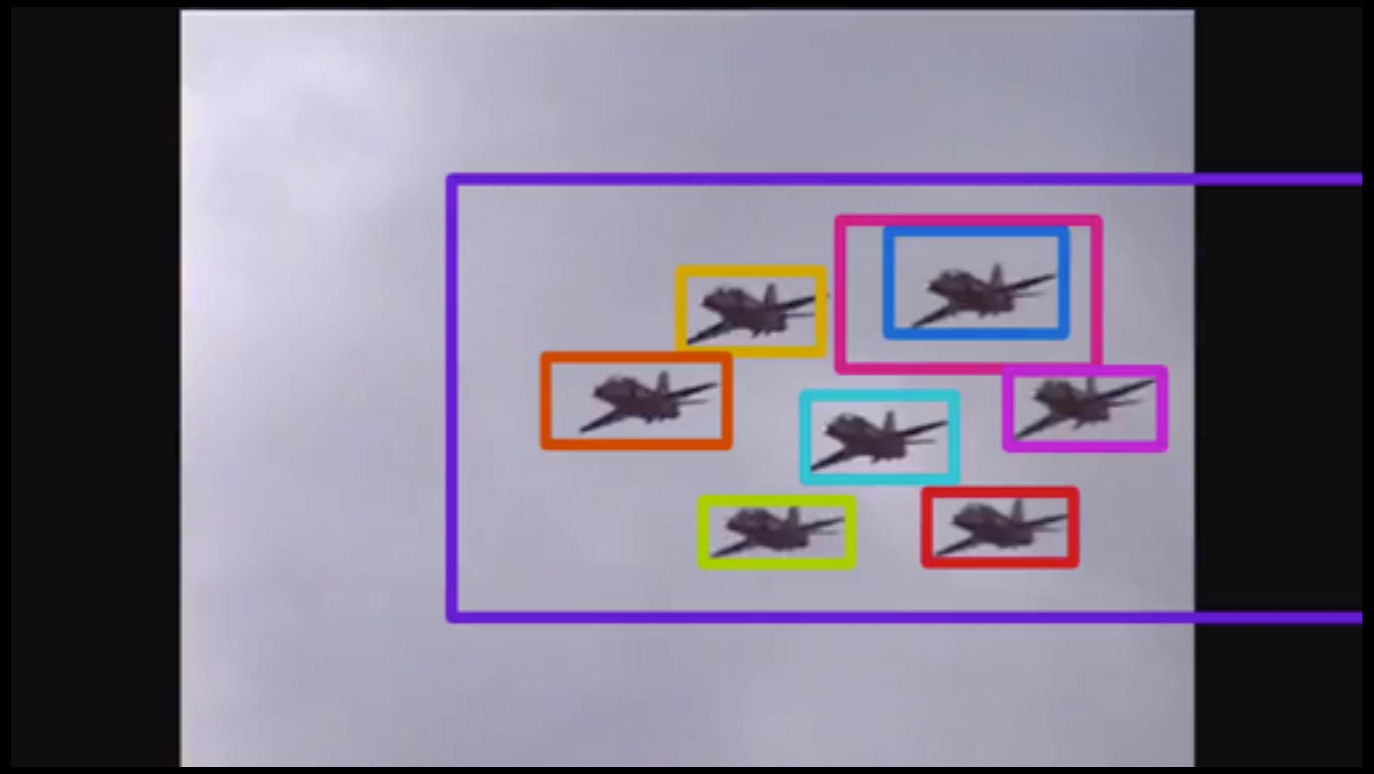
Spatial Max-pooling: Why?

- 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**

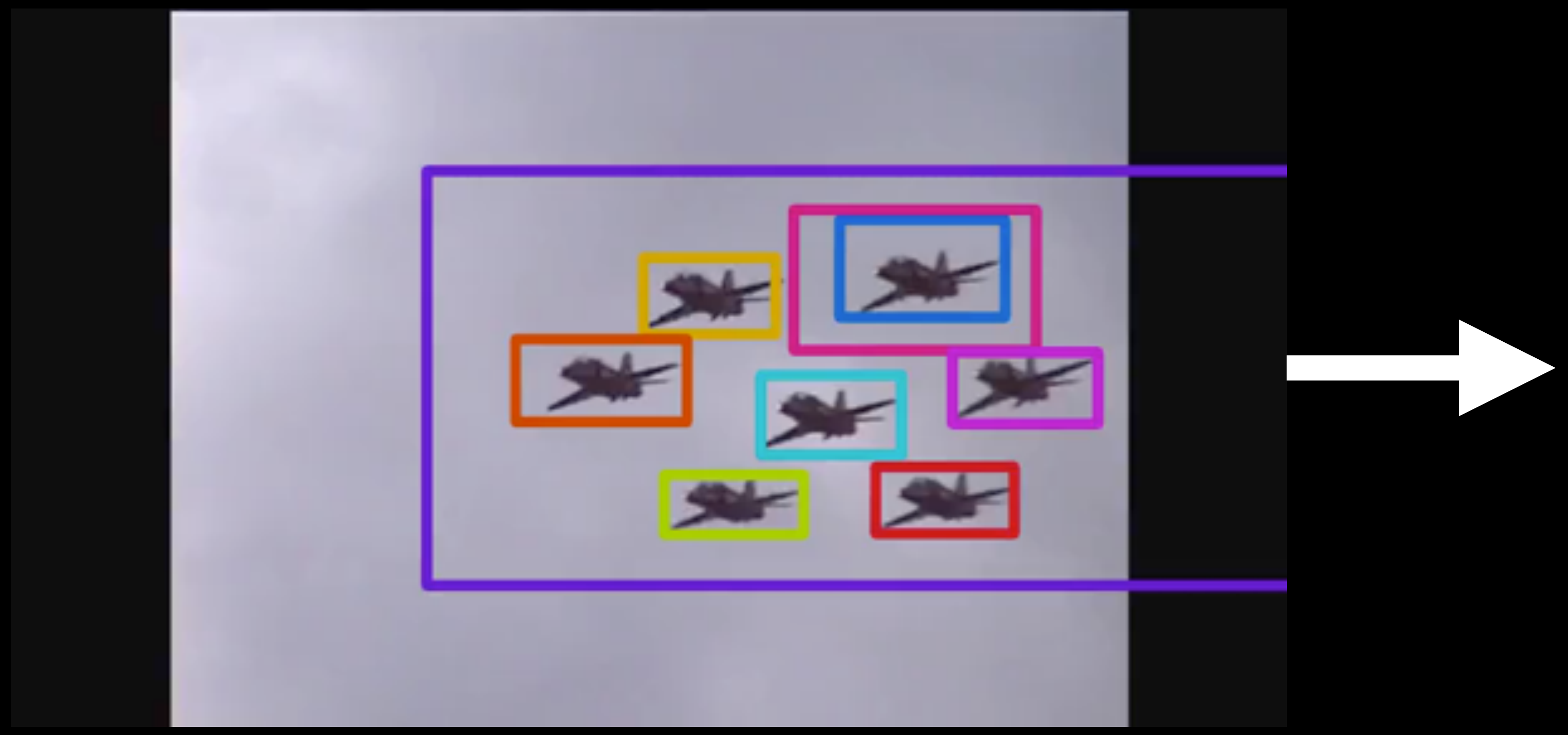
Spatial Max-pooling



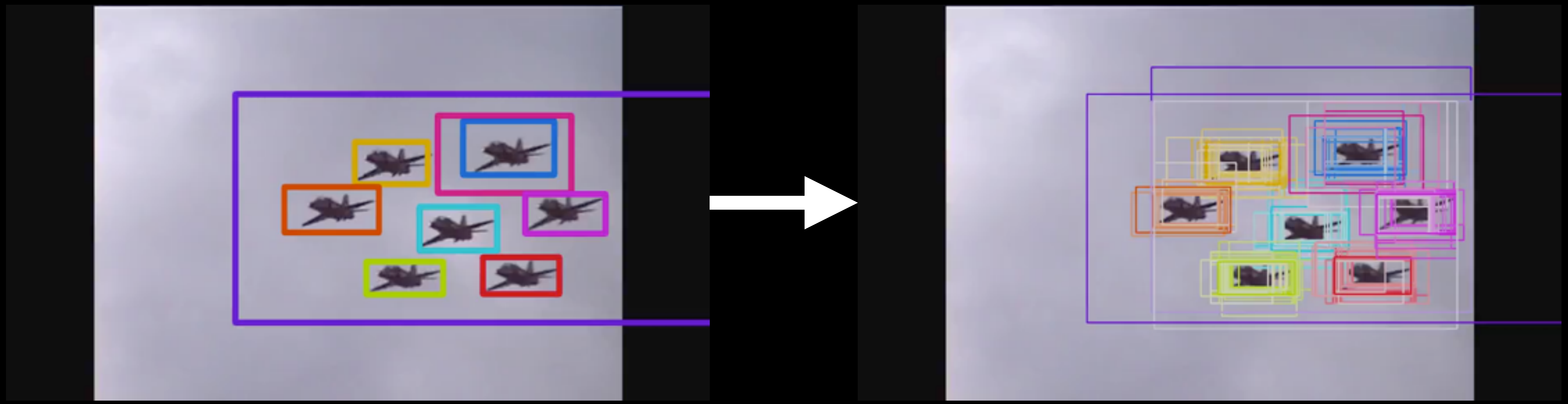
Spatial Max-pooling



Spatial Max-pooling



Spatial Max-pooling

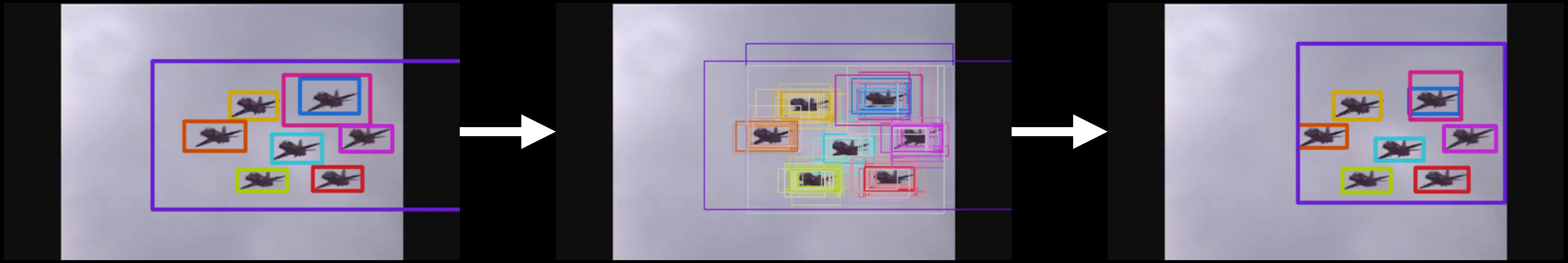


- Still-image detection results that have **large overlaps with tubelet boxes** are chosen for each tubelet

Spatial Max-pooling

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

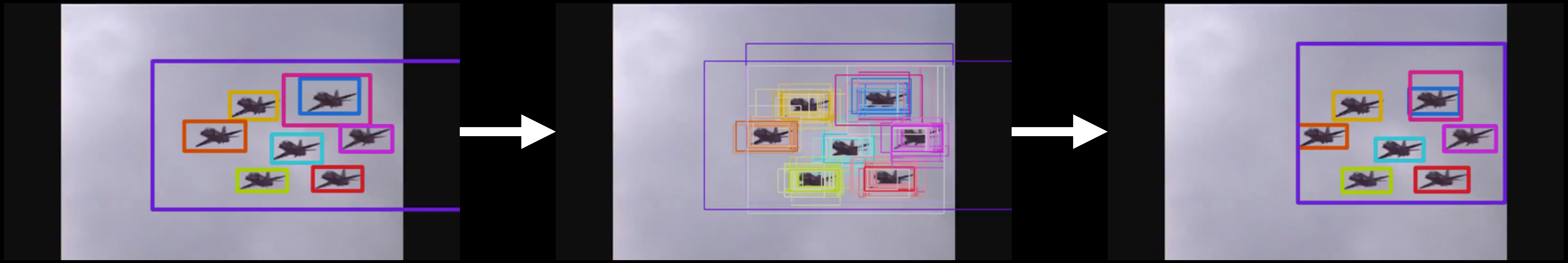
Spatial Max-pooling



- 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

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.

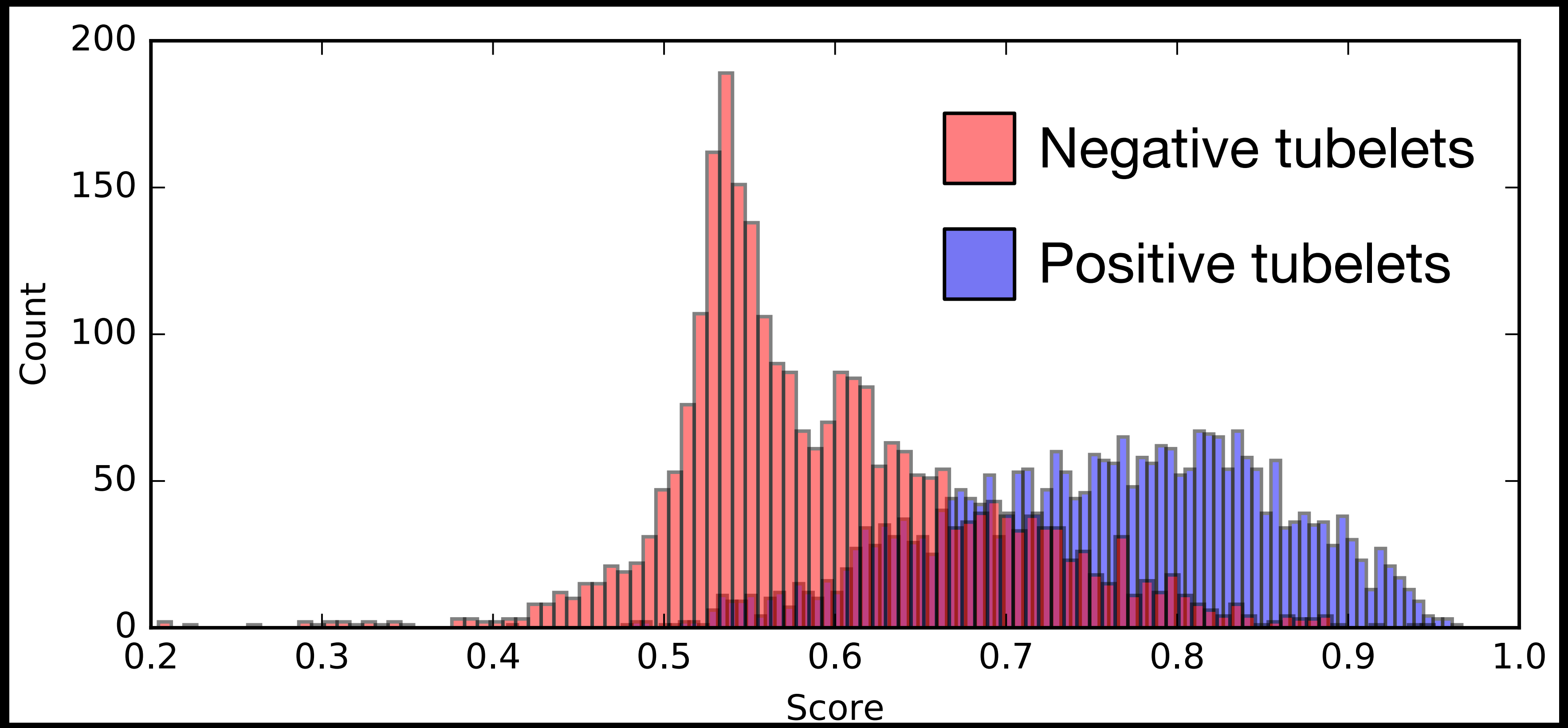
Temporal Re-scoring

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.

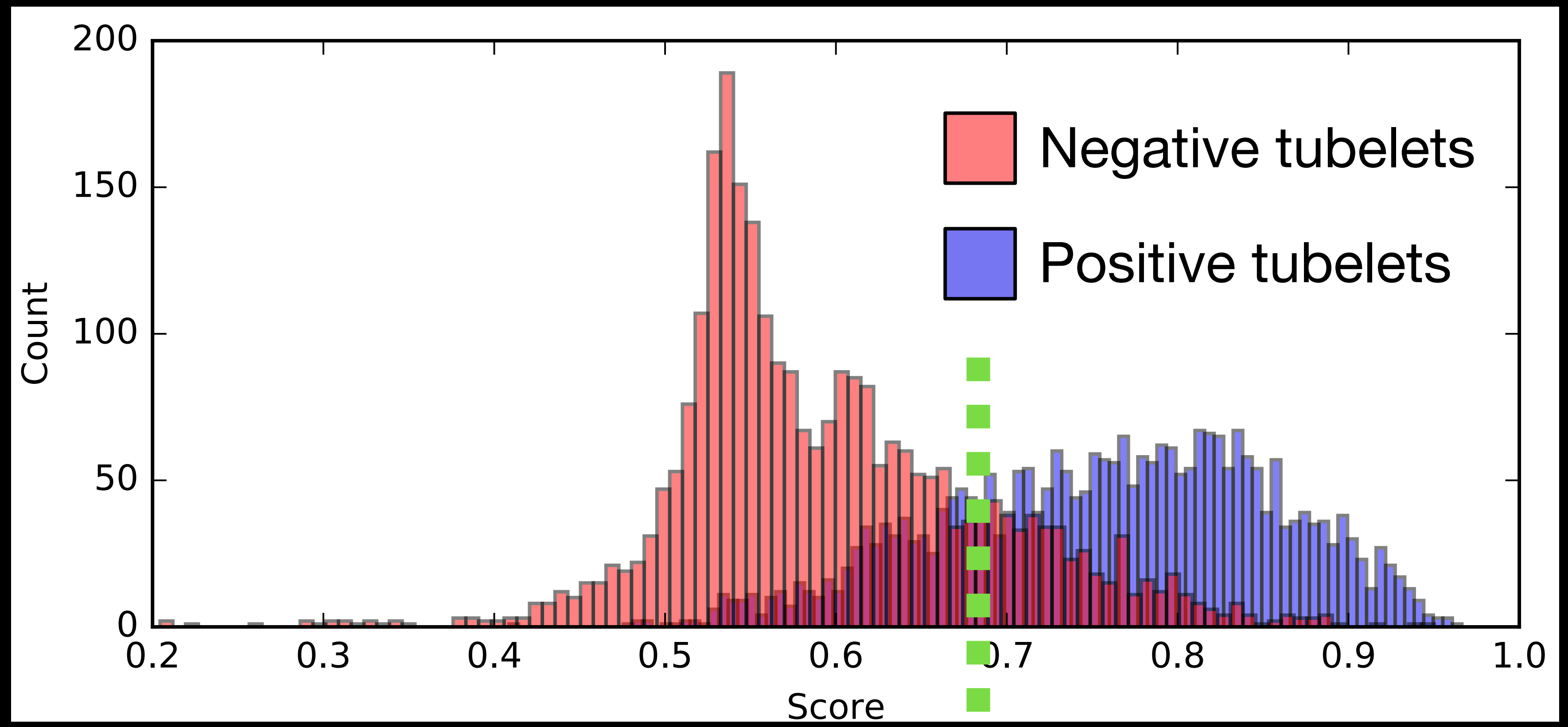
Temporal Re-scoring

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



Temporal Re-scoring

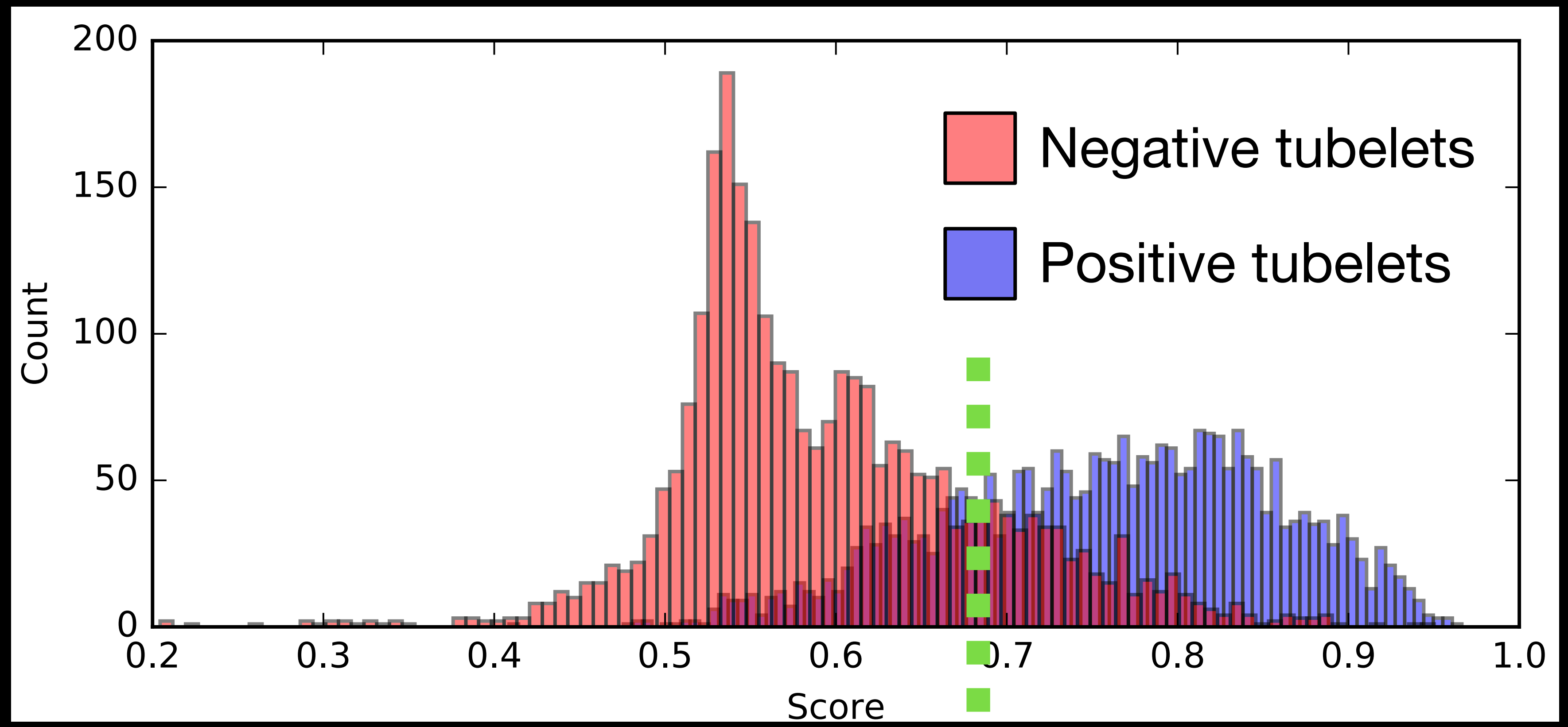
- **Tubelet Classification**. Classify tubelets based on statistics of detection scores (mean, median, top-k). A linear classifier is learnt based on the statistics.



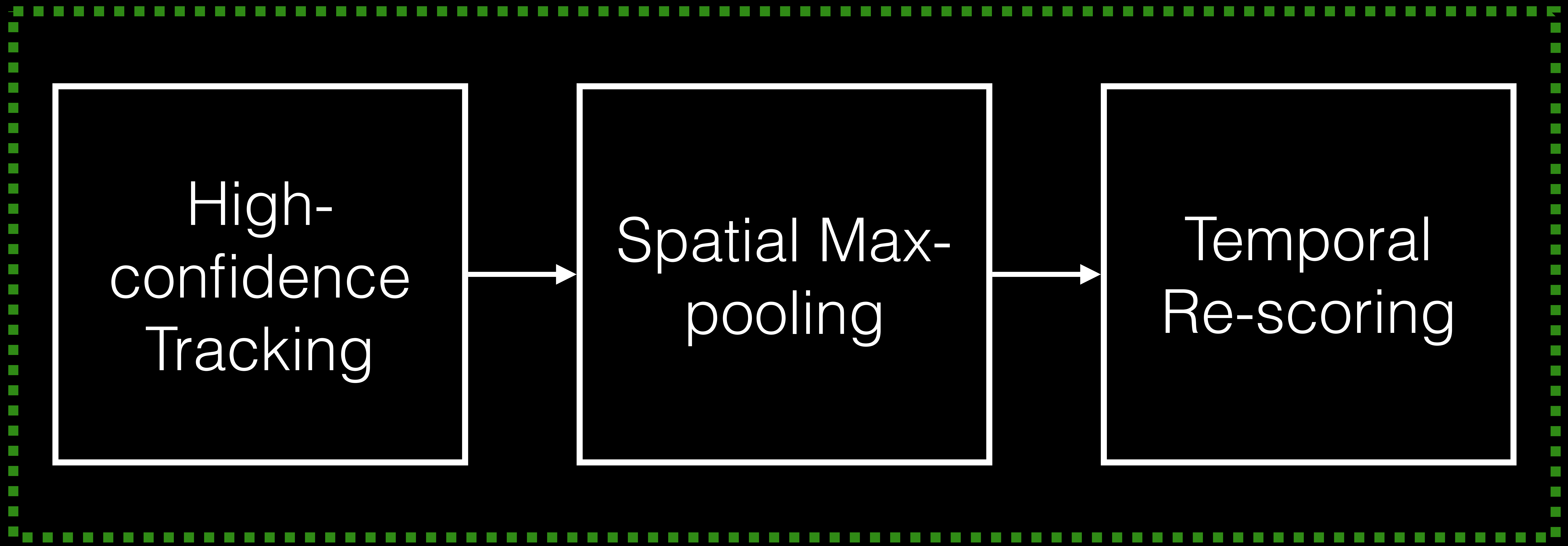
Temporal Re-scoring

- **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



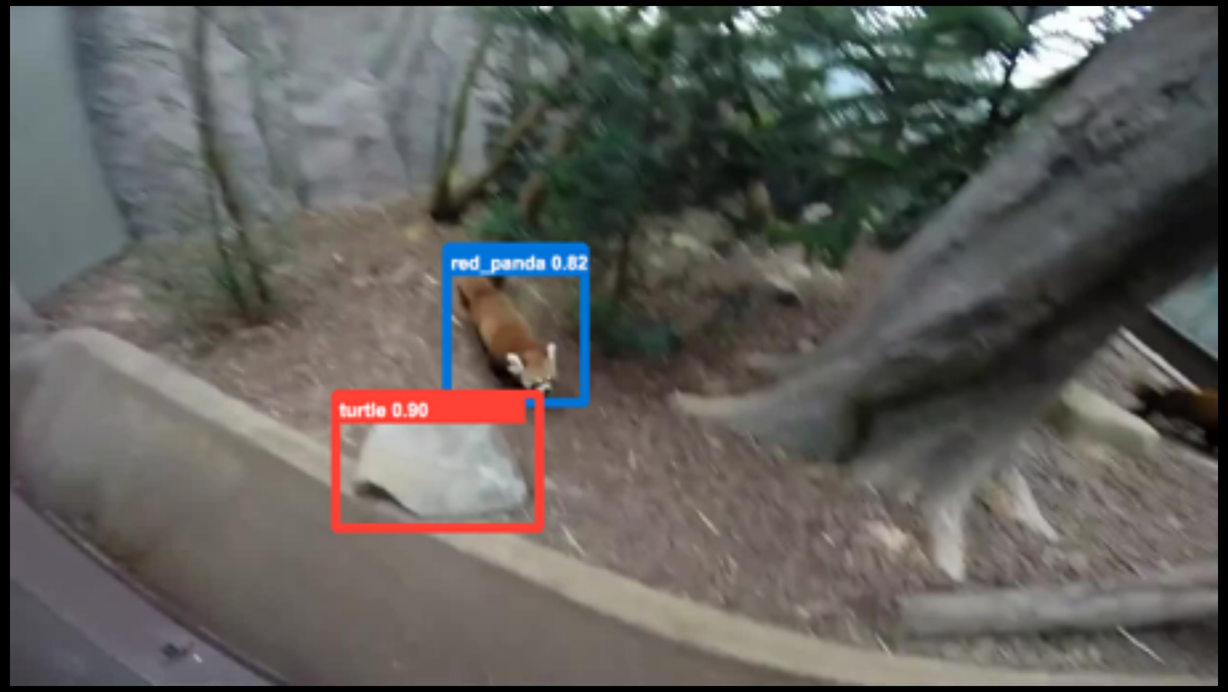
Still-image Detection: Limitation II

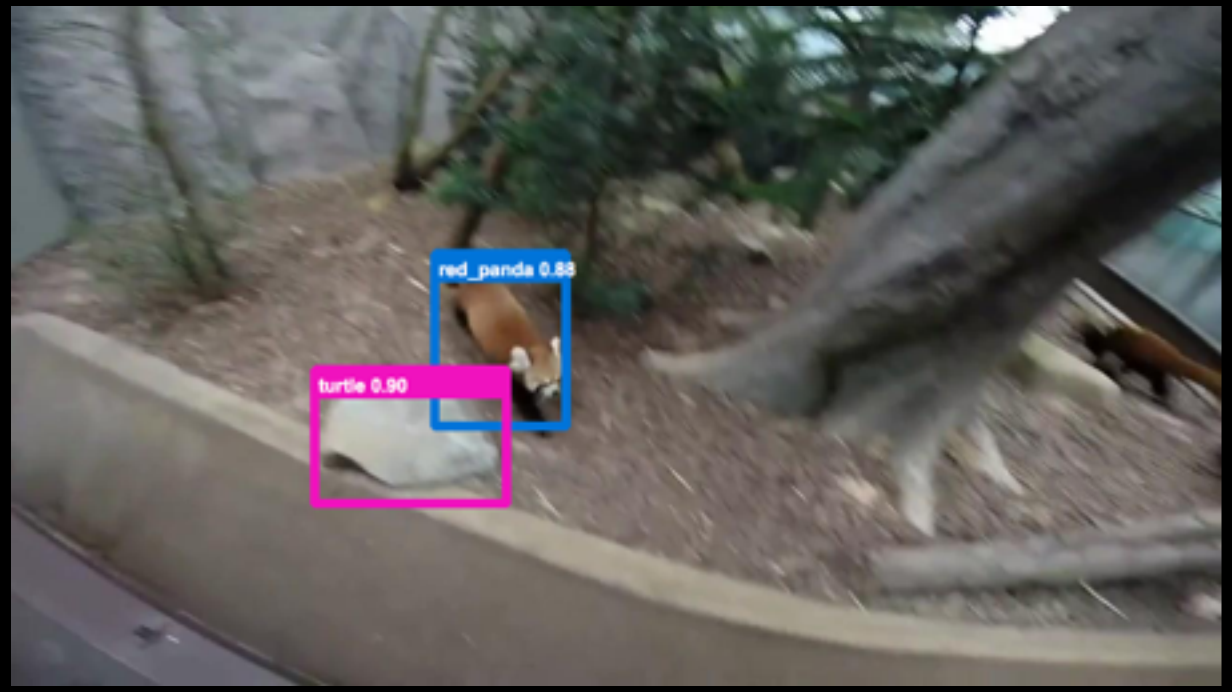
Still-image Detection: Limitation II

Ignored Context

Still-image Detection: Limitation II

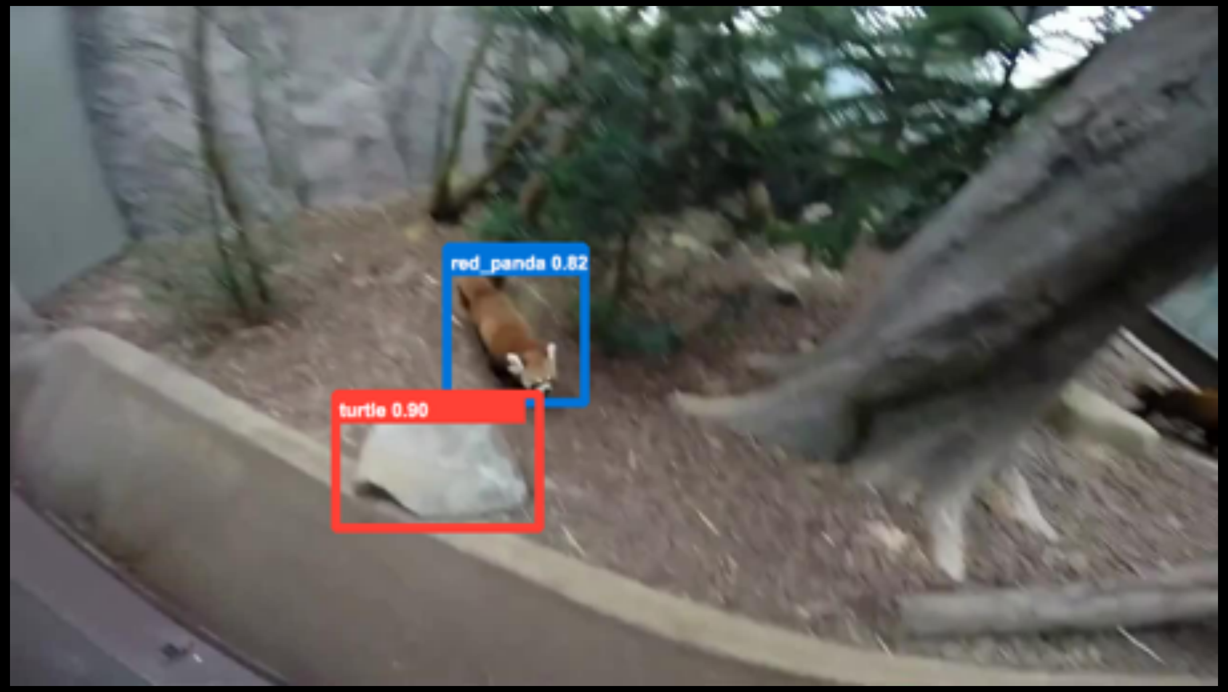
Ignored Context



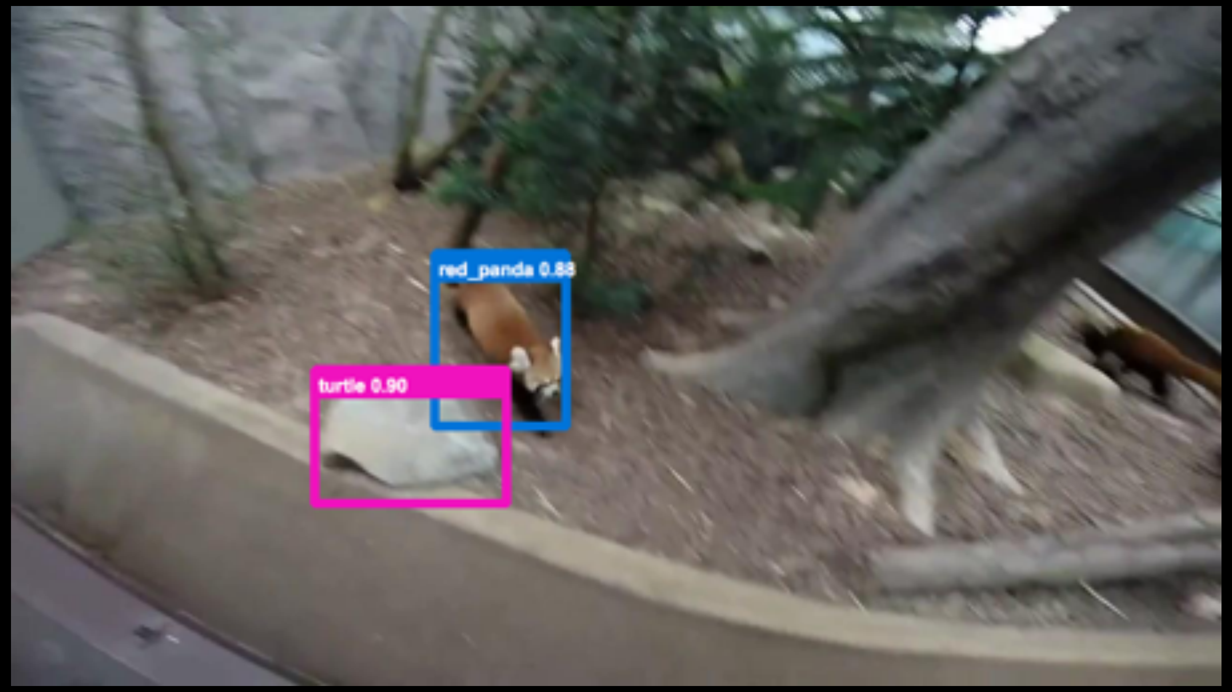


Still-image Detection: Limitation II

Ignored Context



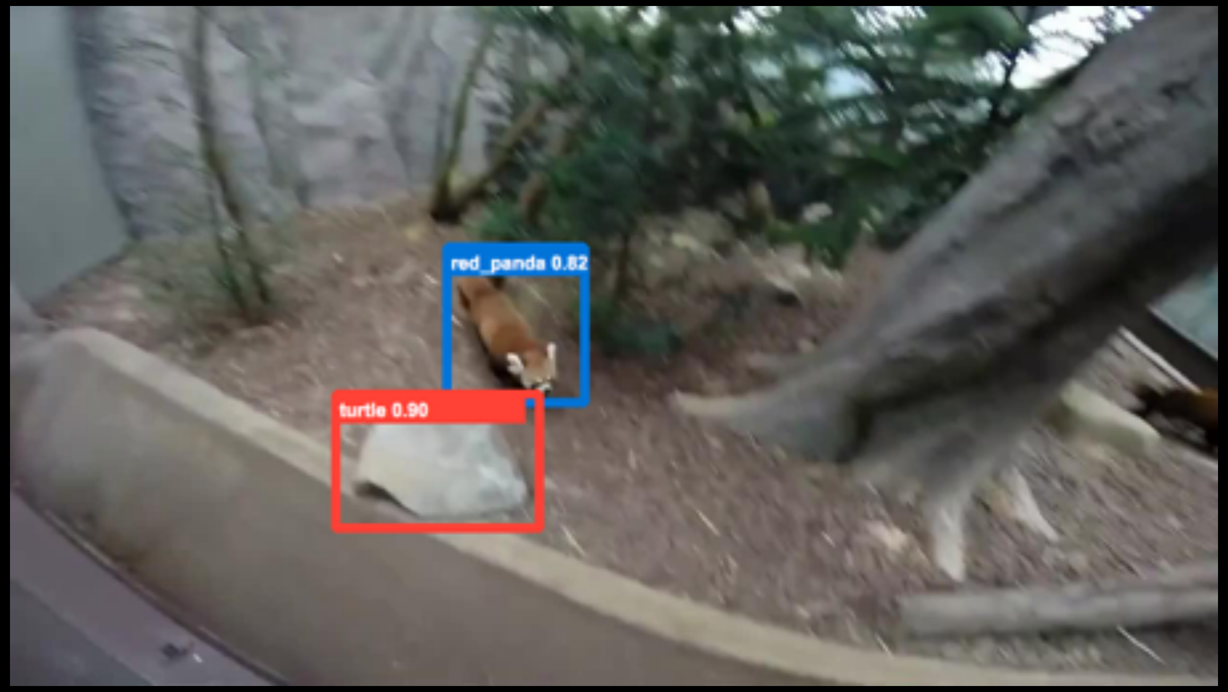
red panda turtle



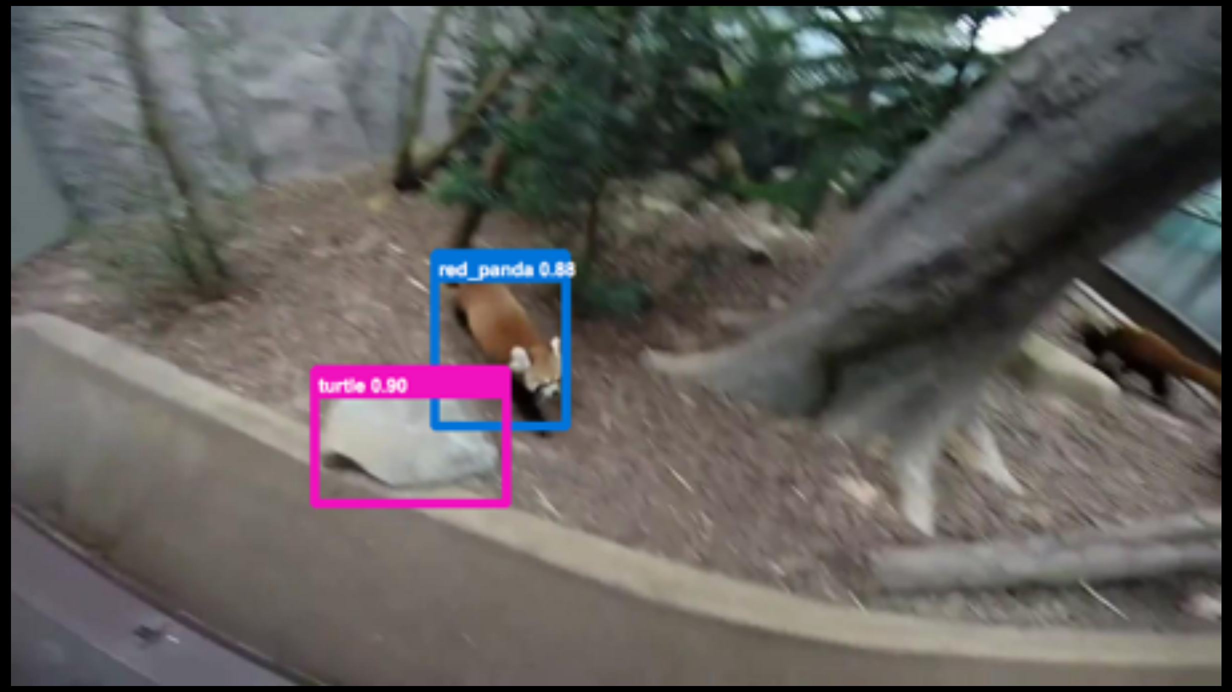
red panda turtle

Still-image Detection: Limitation II

Ignored Context



red panda turtle



red panda turtle

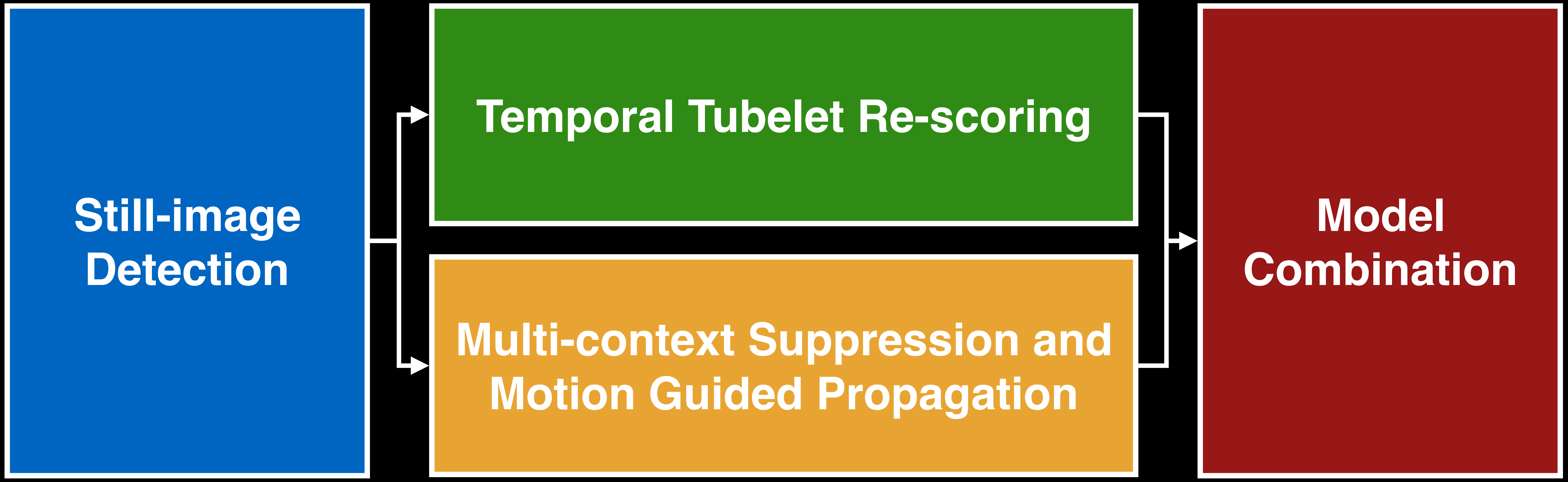
red panda

red panda

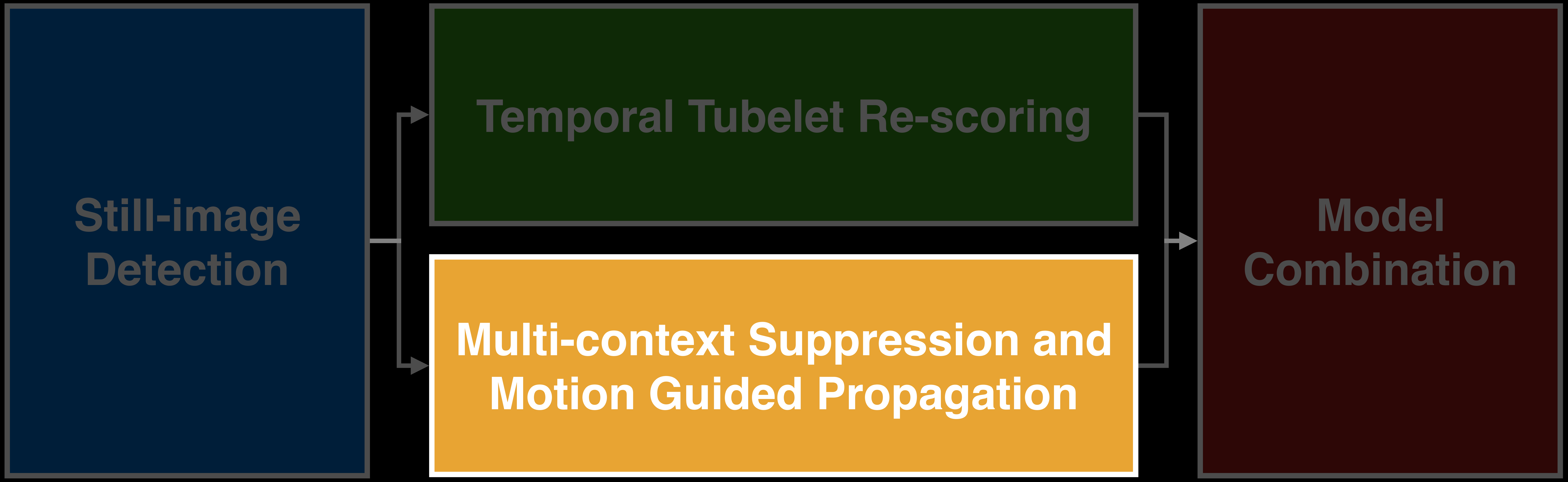
red panda

red panda

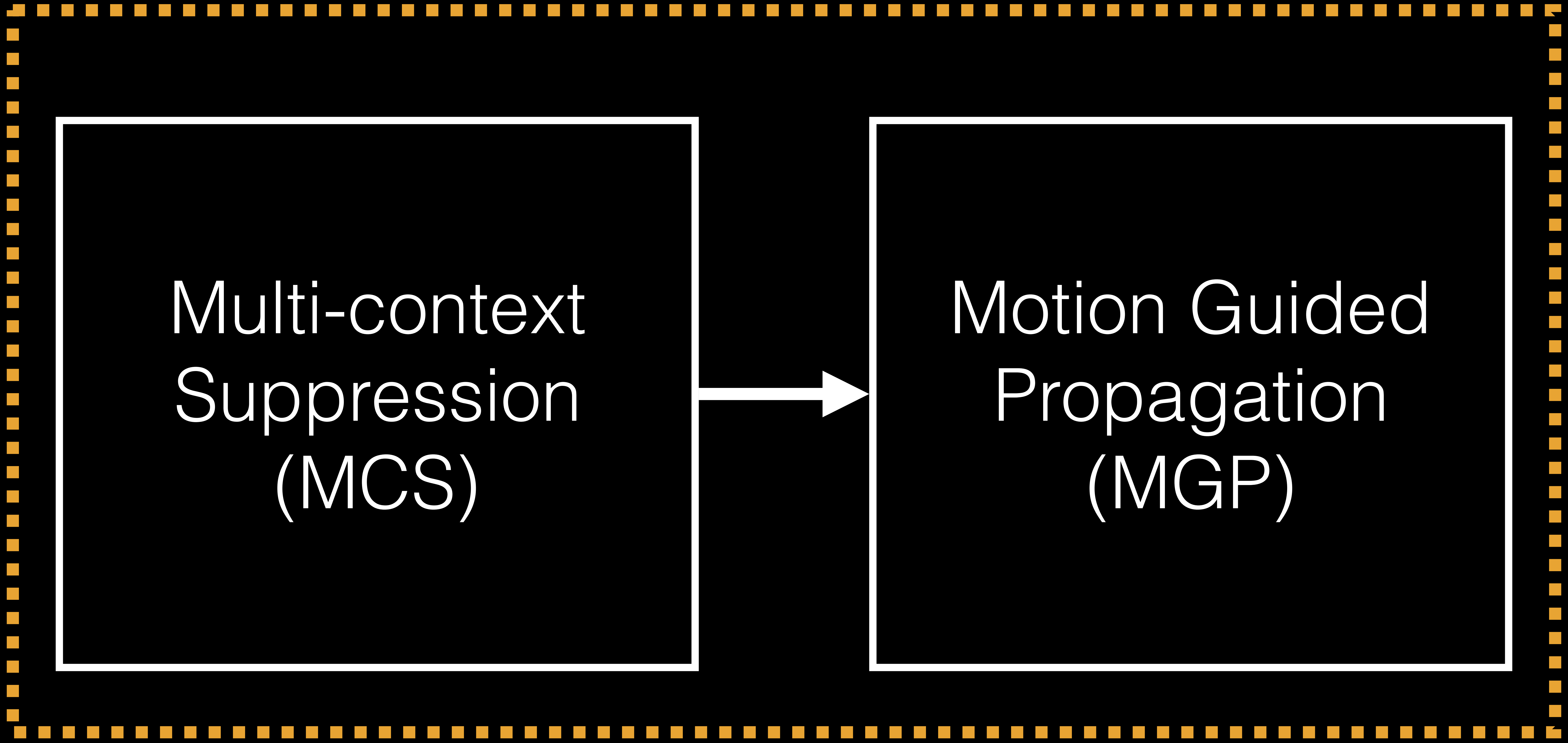
Proposed Framework



Proposed Framework



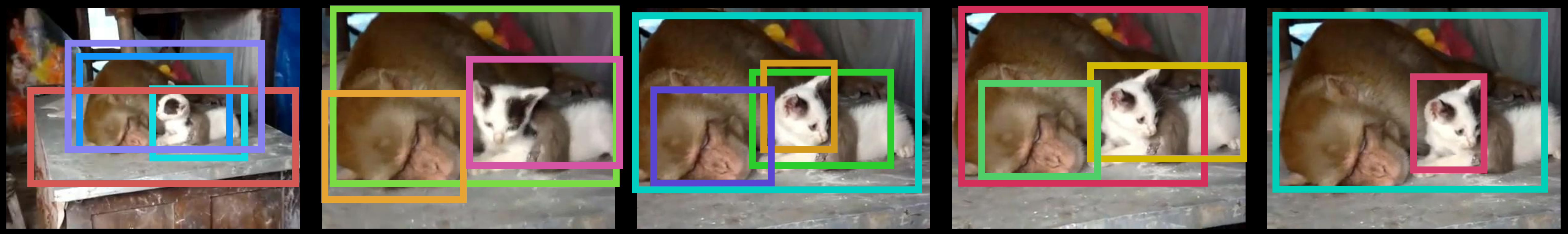
Multi-context Suppression and Motion Guided Propagation



Multi-context Suppression (MCS)

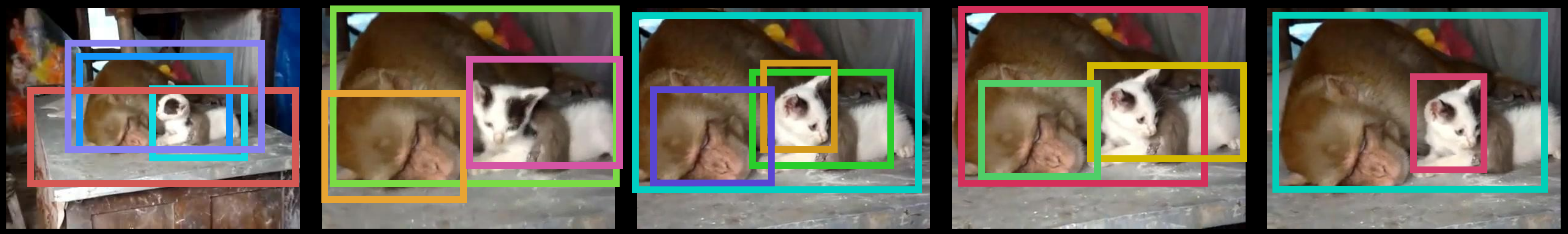
Multi-context Suppression (MCS)

Multi-context Suppression (MCS)



- **Sort** all detection scores of all proposals in a video in **descending order**

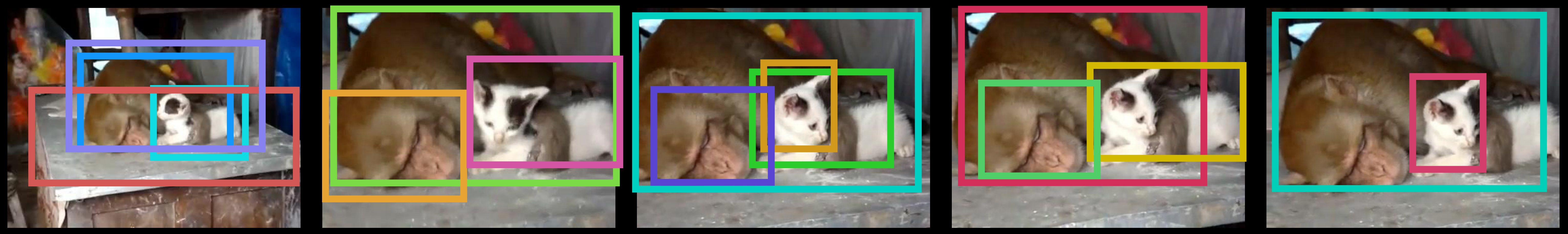
Multi-context Suppression (MCS)



monkey, cat

- **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

Multi-context Suppression (MCS)



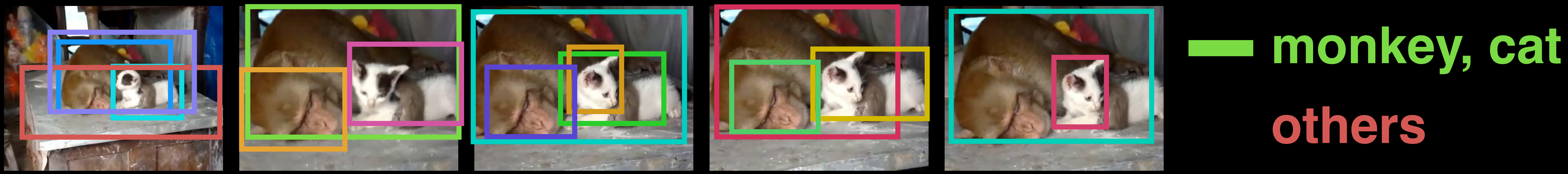
monkey, cat

others

- **Sort** all detection scores of all proposals in a video in **descending order**

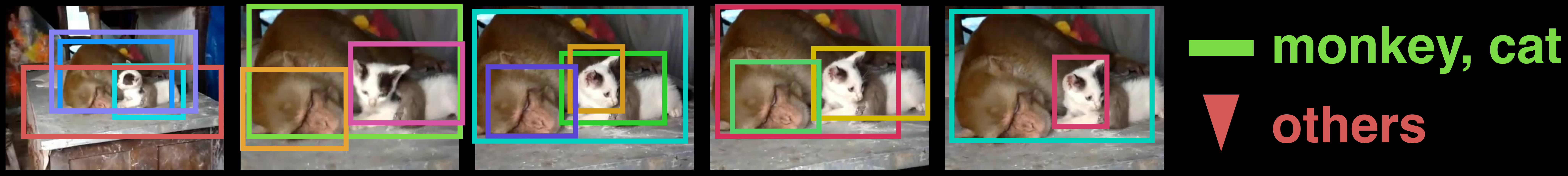
- The classes of the **high rankings** are denoted as the confident classes

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

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

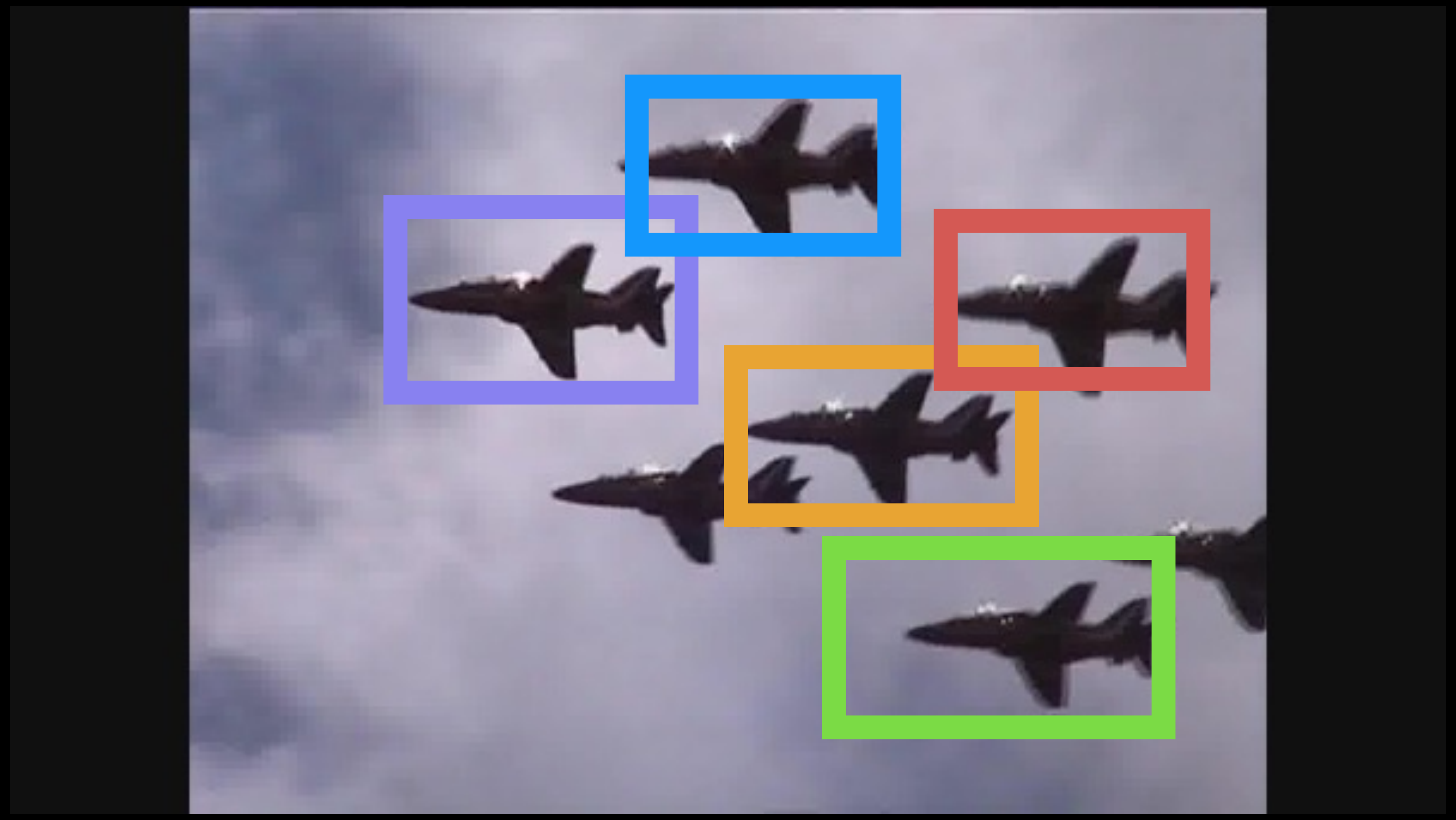
- The scores of **classes with low rankings** are suppressed, while the scores of confident classes remain unchanged

Motion Guided Propagation (MGP)

Motion Guided Propagation (MGP)

Frame t

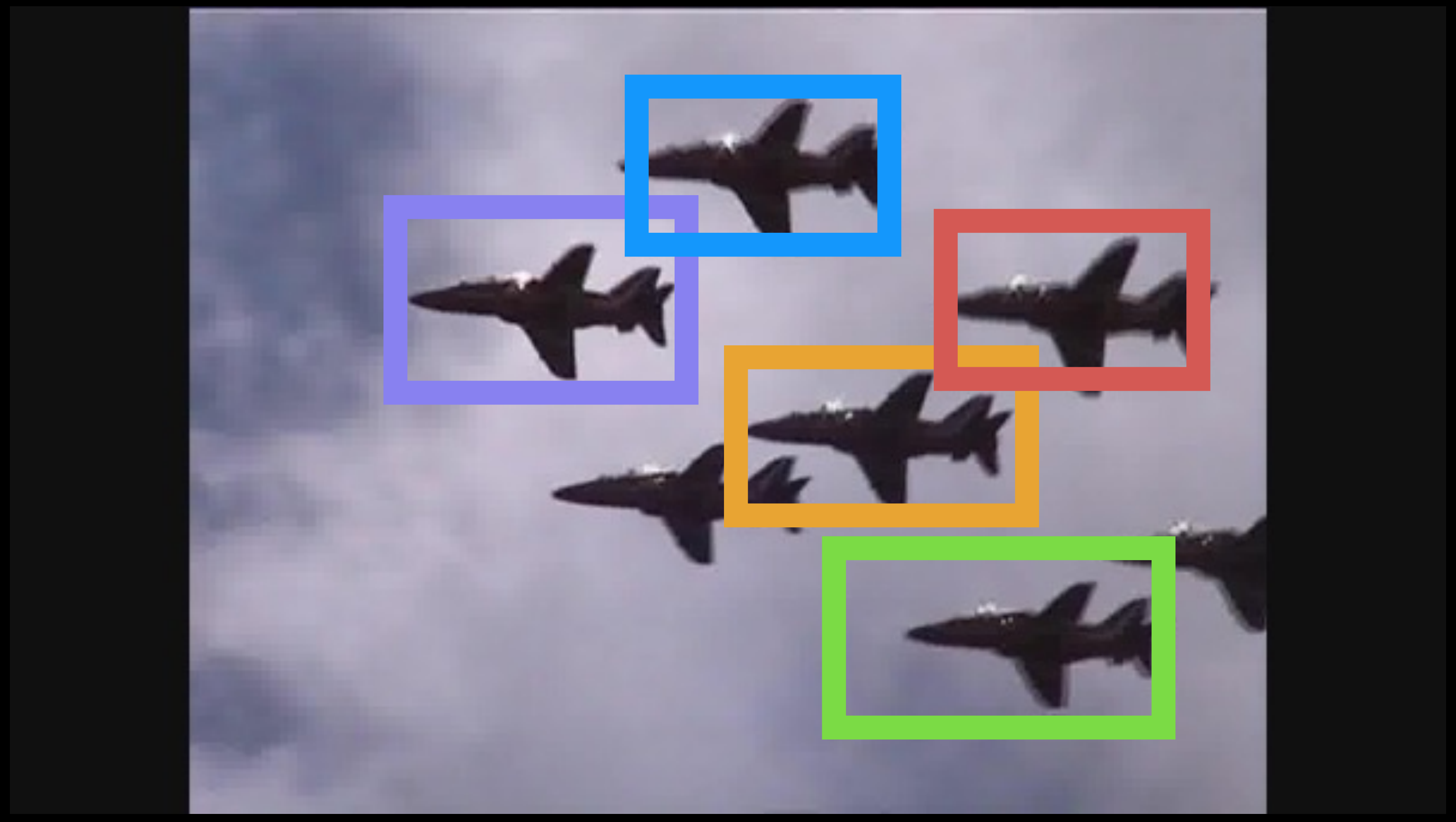
Motion Guided Propagation (MGP)



Frame t

Motion Guided Propagation (MGP)

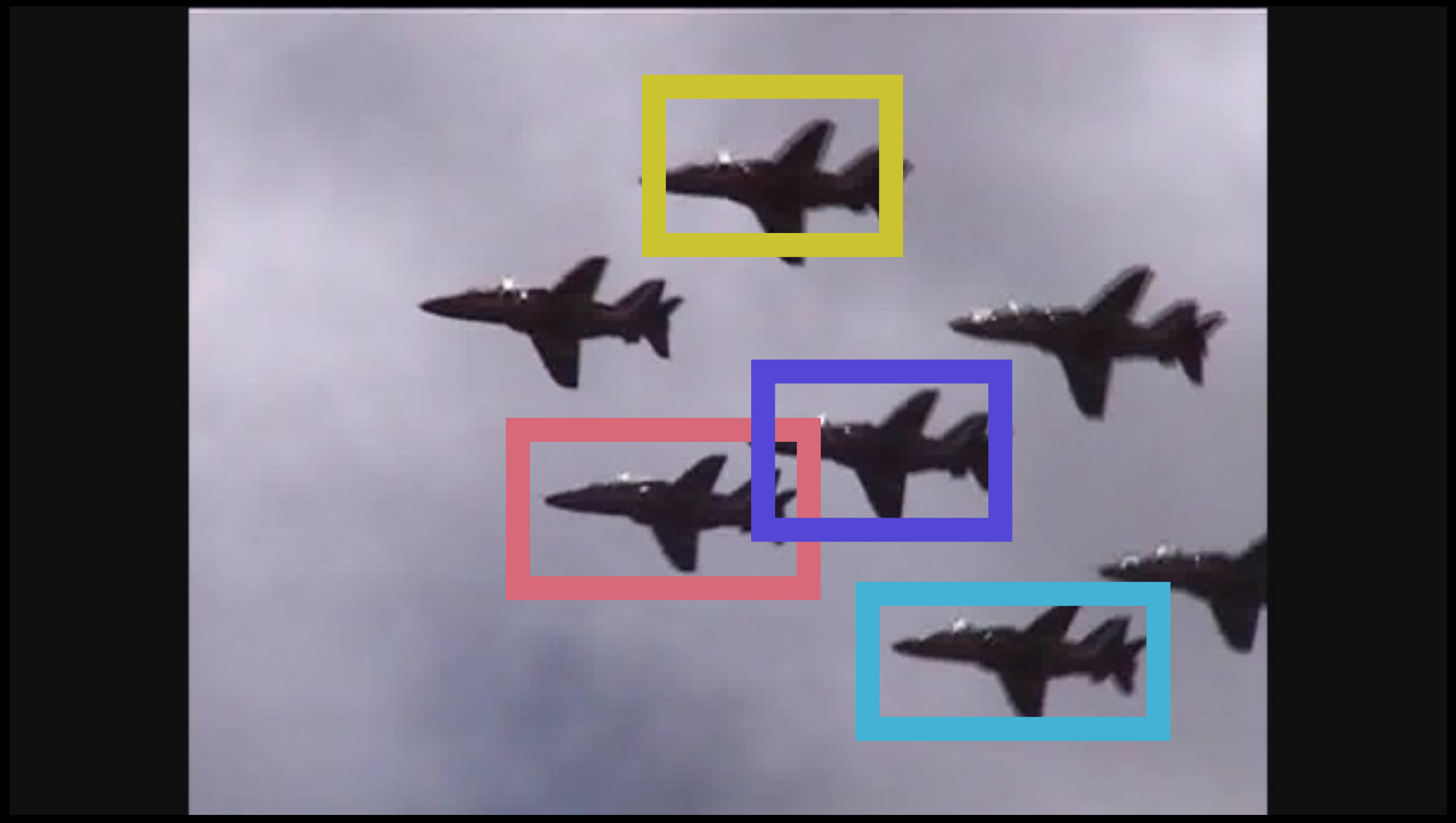
Frame t-1



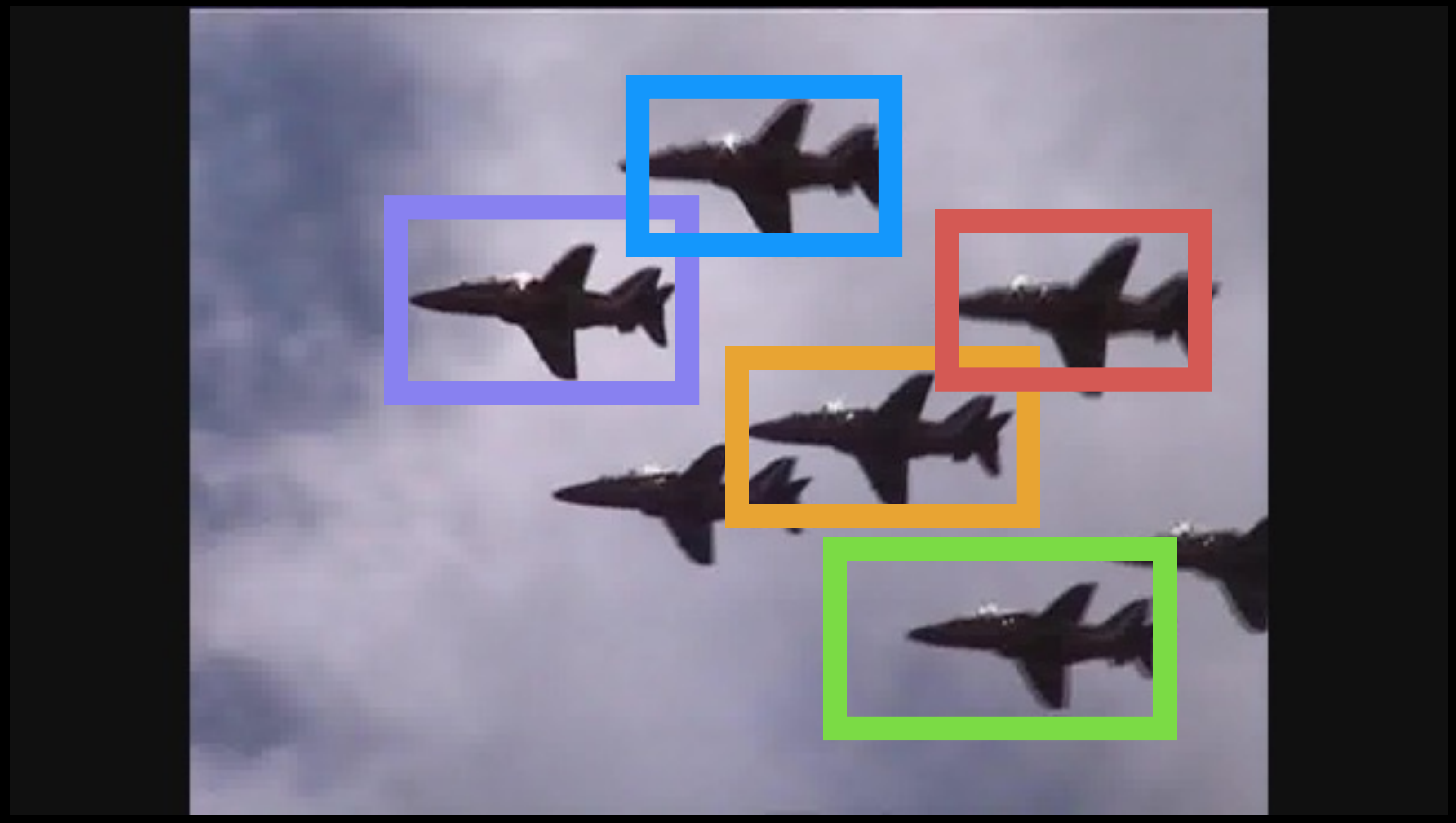
Frame t

Frame t+1

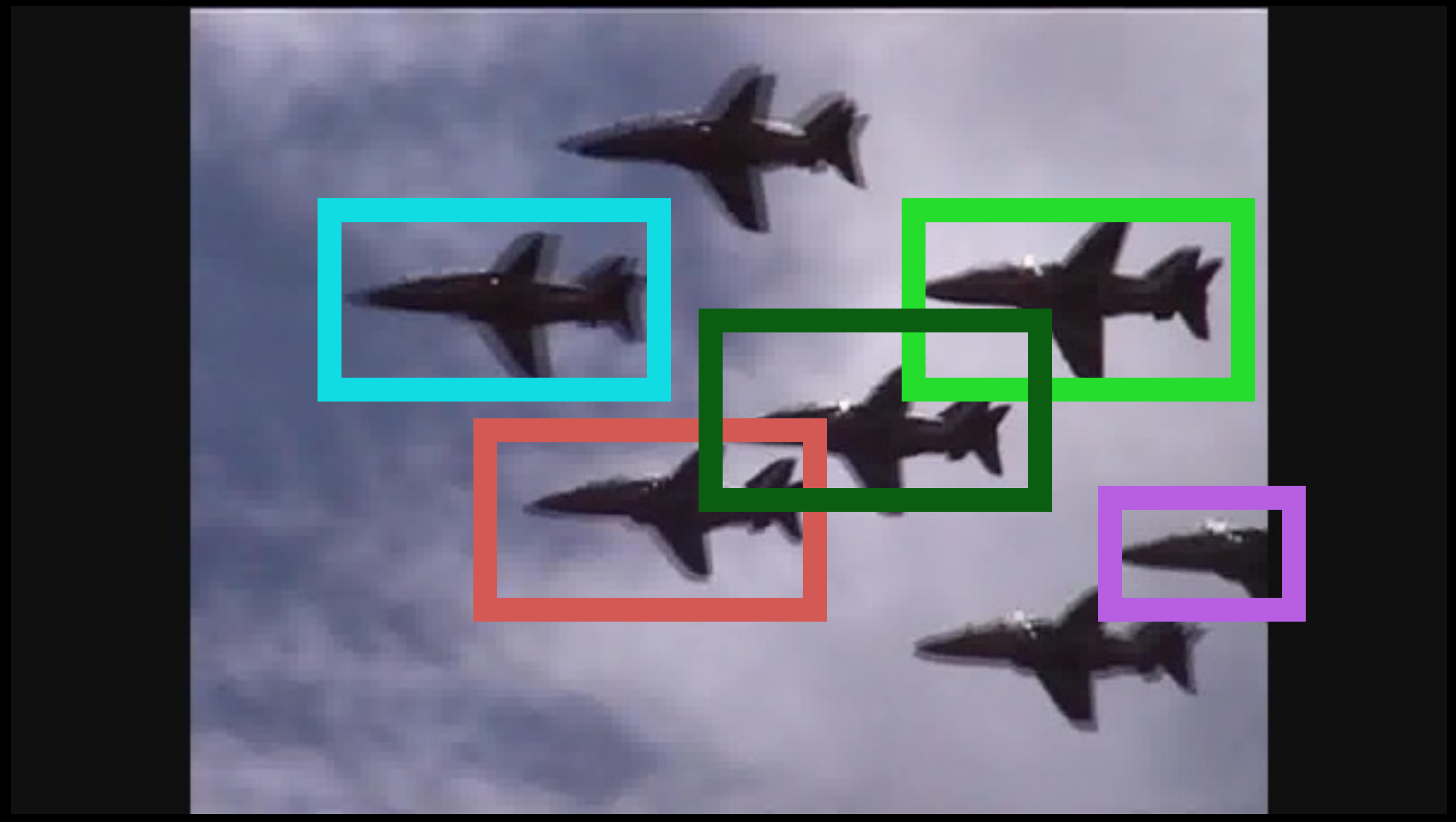
Motion Guided Propagation (MGP)



Frame t-1

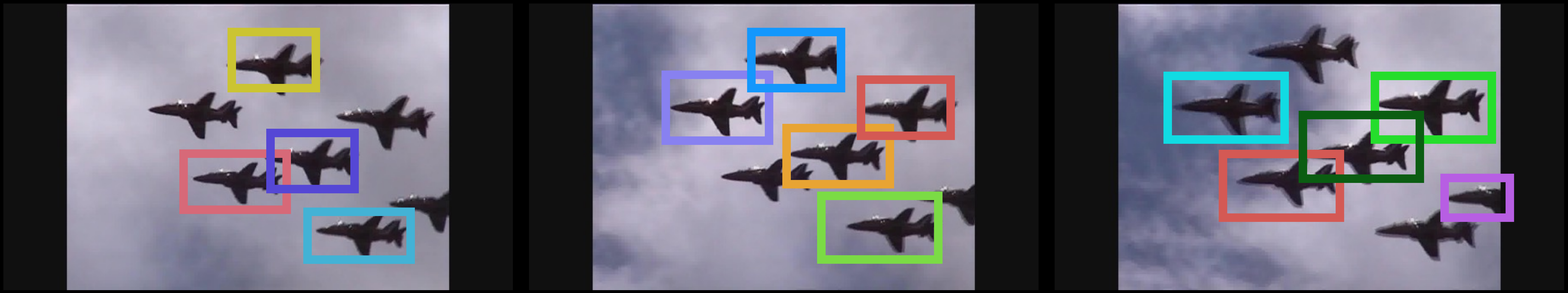


Frame t



Frame t+1

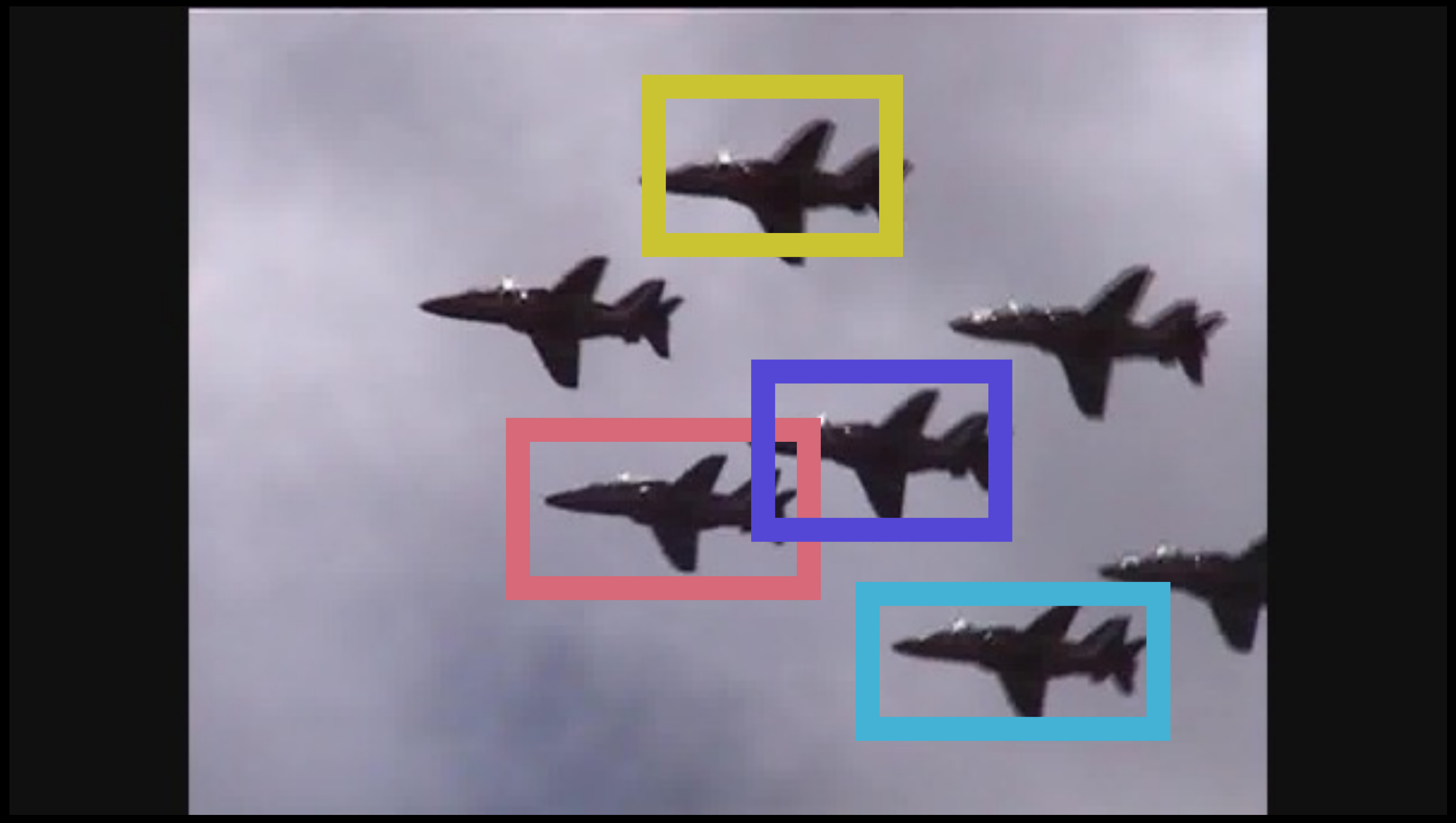
Motion Guided Propagation (MGP)



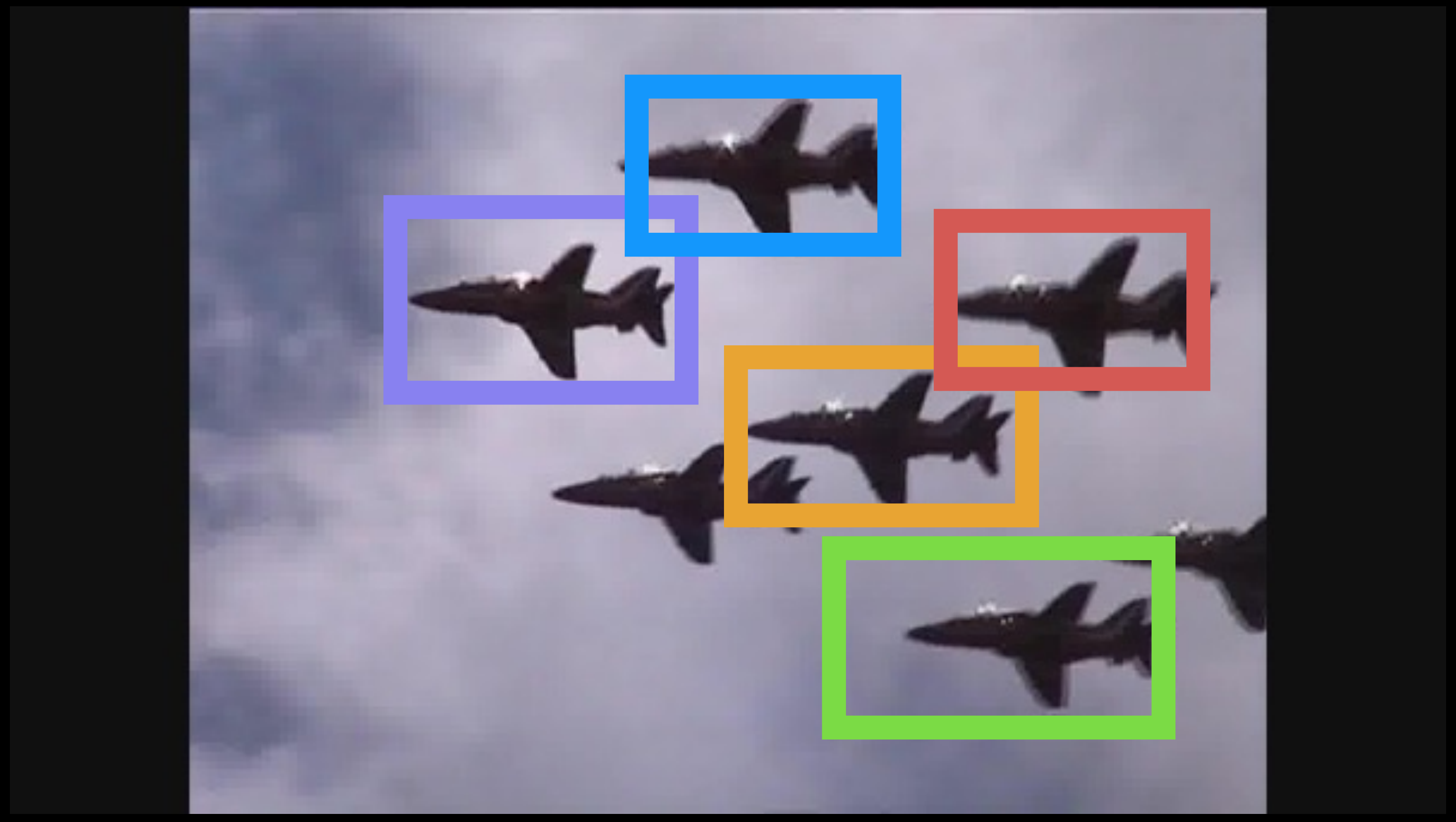
Frame t-1 | Frame t | Frame t+1

- In each frame, some objects are **not found by detector**. However, detections on adjacent frames are **complementary** to each other.

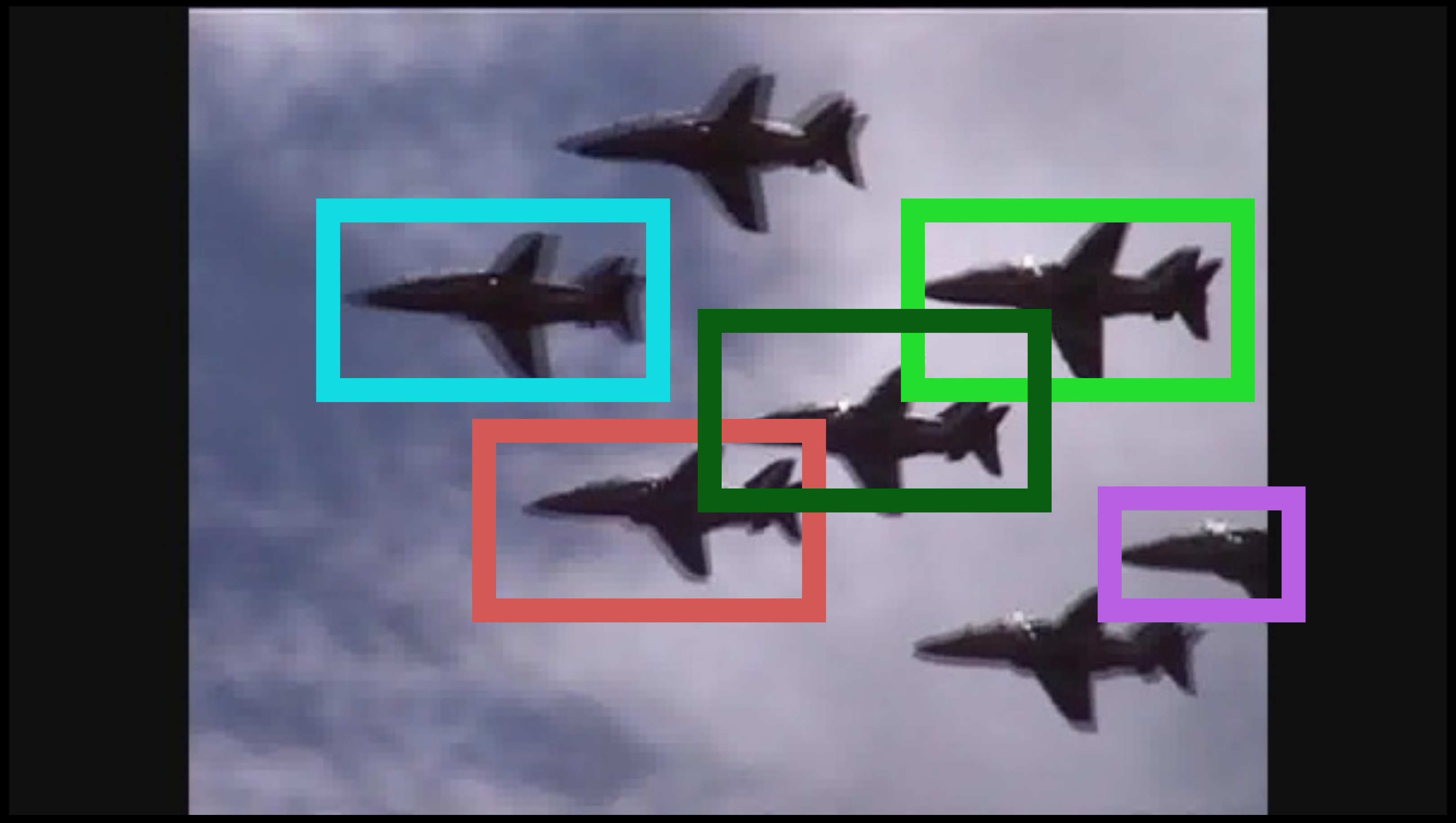
Motion Guided Propagation (MGP)



Frame t-1



Frame t

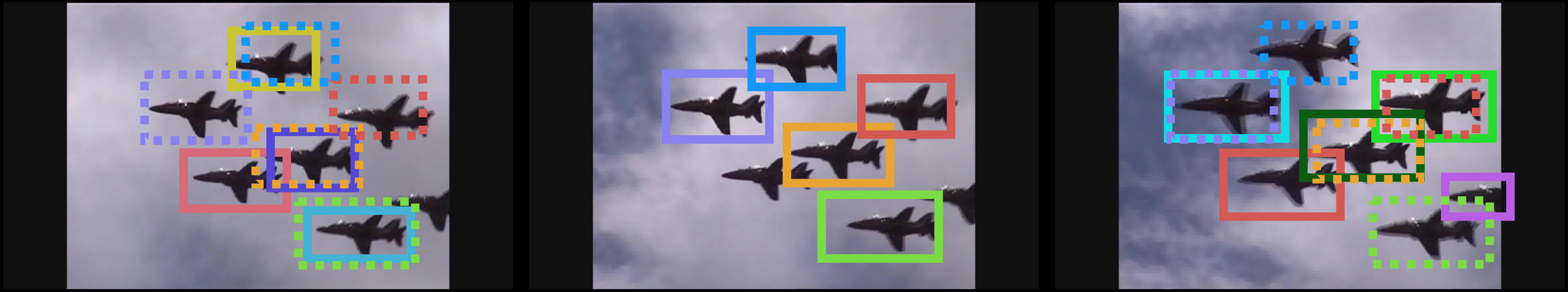


Frame t+1

- 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.

Motion Guided Propagation (MGP)

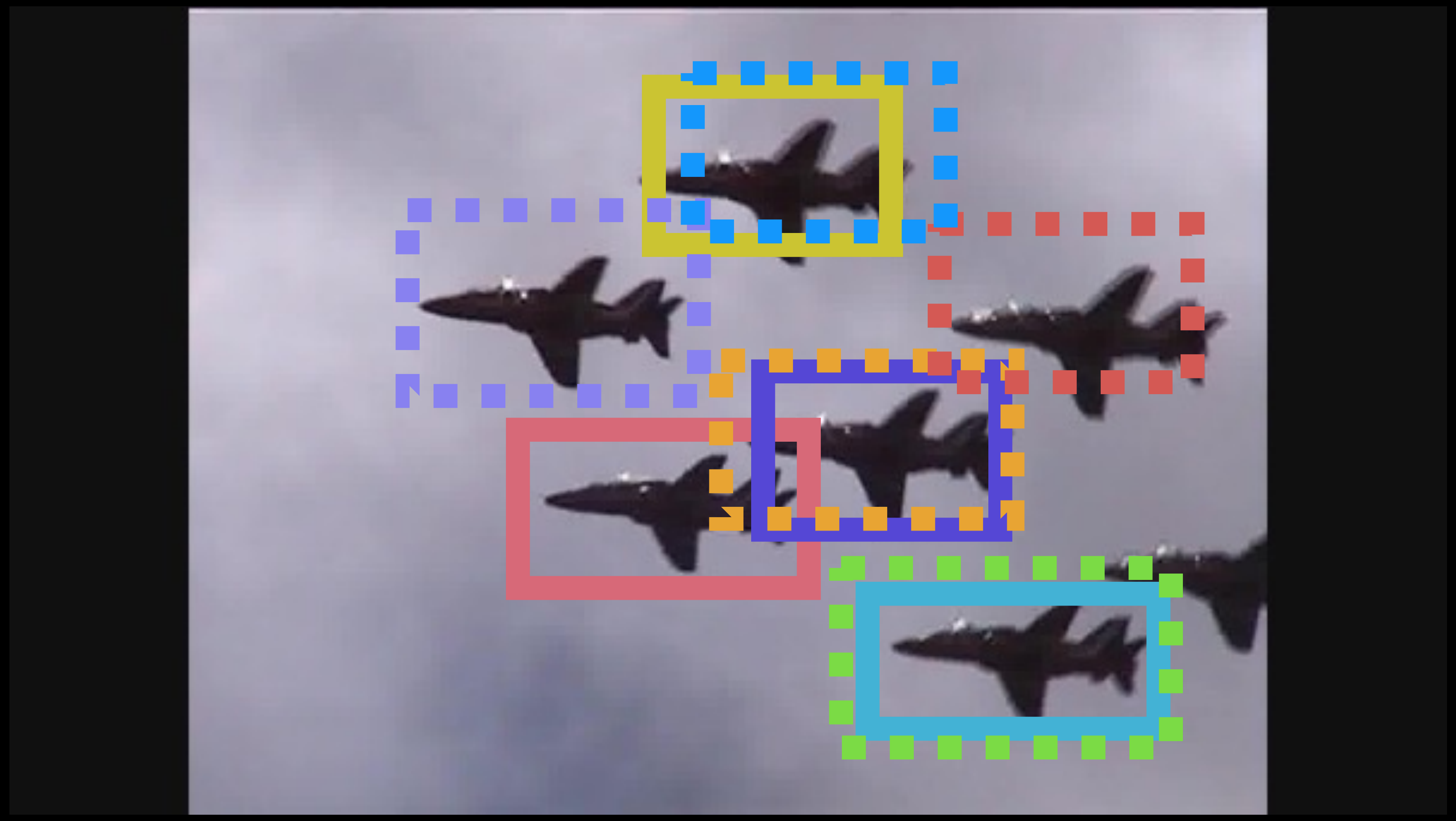


Frame t-1 Frame t Frame t+1

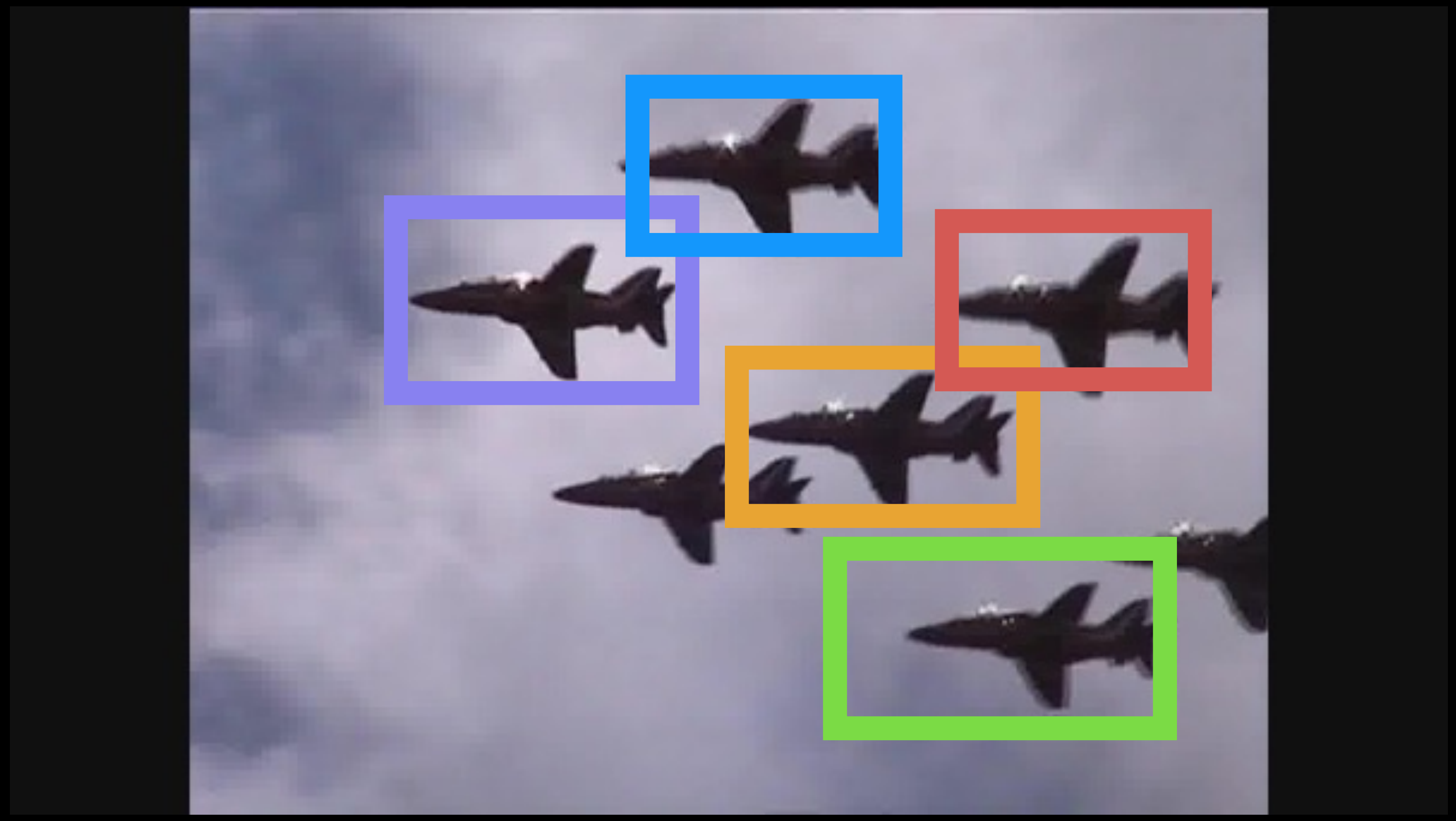
- In each frame, some objects are **not found by detector**. However, detections on adjacent frames are **complementary** to each other.

- Detections are **propagated to adjacent** frames. Optical flow is used for guiding the propagation.

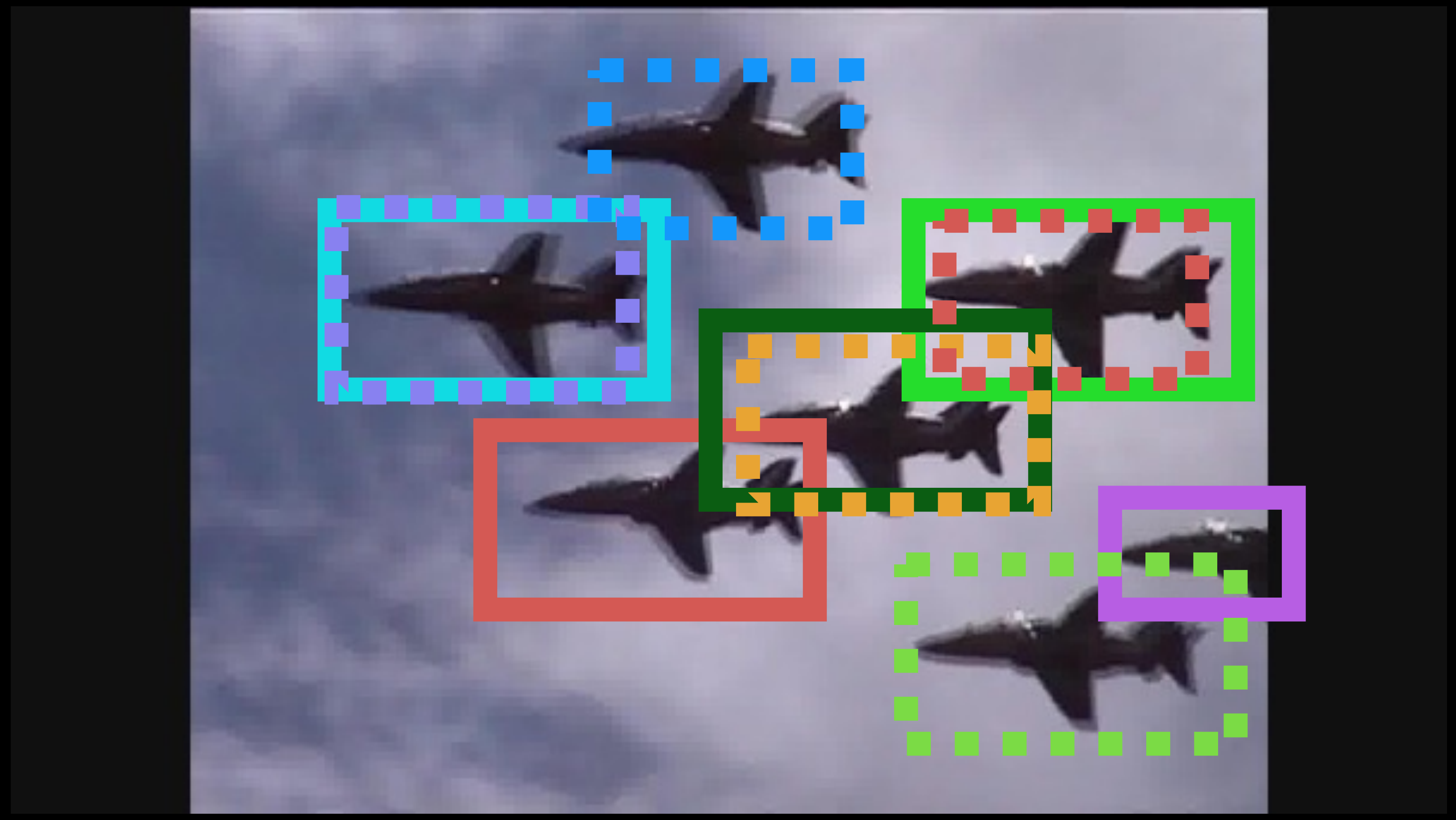
Motion Guided Propagation (MGP)



Frame t-1



Frame t



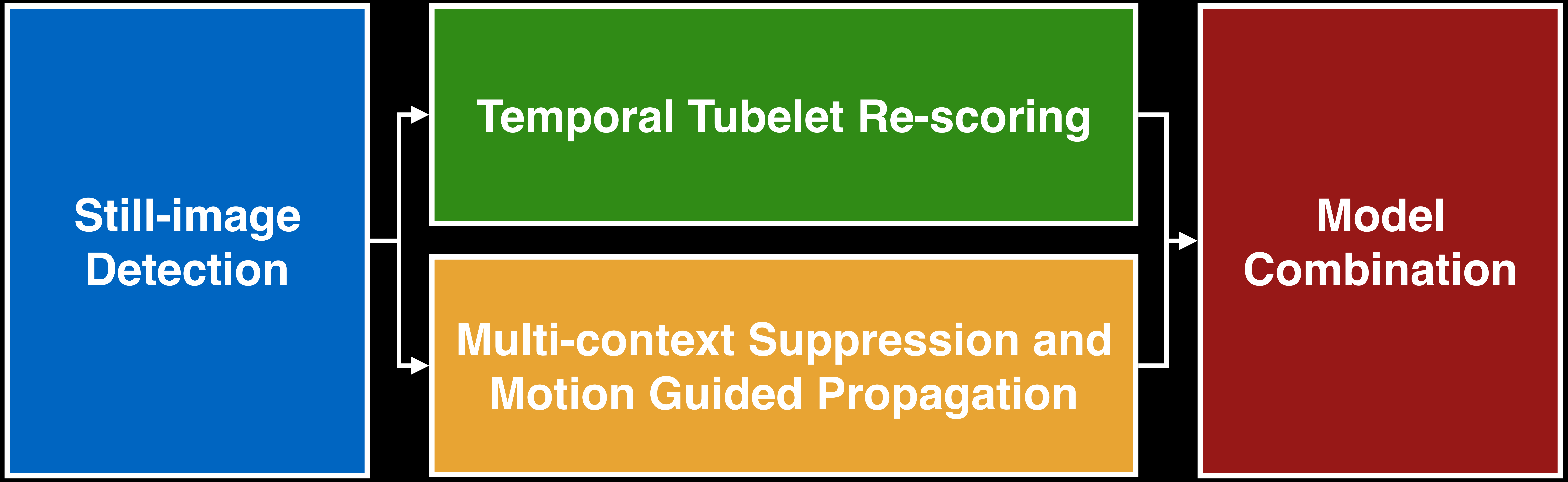
Frame t+1

- In each frame, some objects are **not found by detector**. However, detections on adjacent frames are **complementary** to each other.

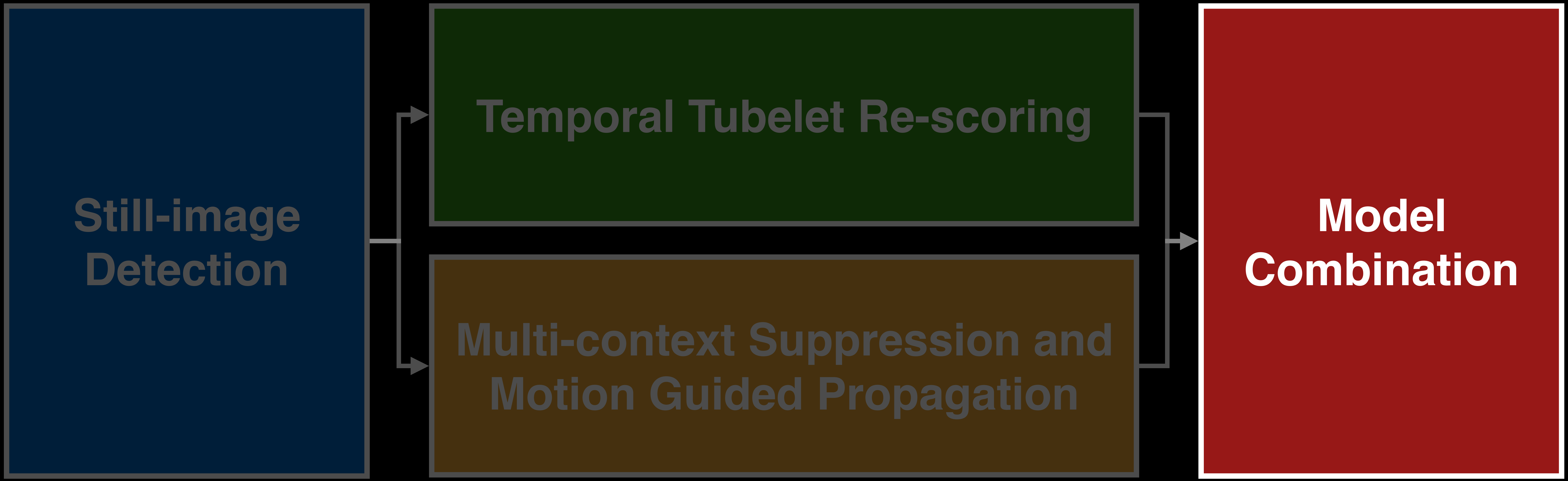
- Detections are **propagated to adjacent** frames. Optical flow is used for guiding the propagation.

- Propagation results in redundant boxes, which can be **easily handled** by non-maximum suppression (NMS)

Proposed Framework



Proposed Framework



Model Combination

Model Combination

Score Average

- Two groups of proposals: 1) Region Proposal Networks (RPN), 2) Selective Search + EdgeBox. Given a group of proposals, their detection scores can be obtained by averaging several models.

Model Combination

- Two groups of proposals: 1) Region Proposal Networks (RPN), 2) Selective Search + EdgeBox. Given a group of proposals, their detection scores can be obtained by averaging several models.

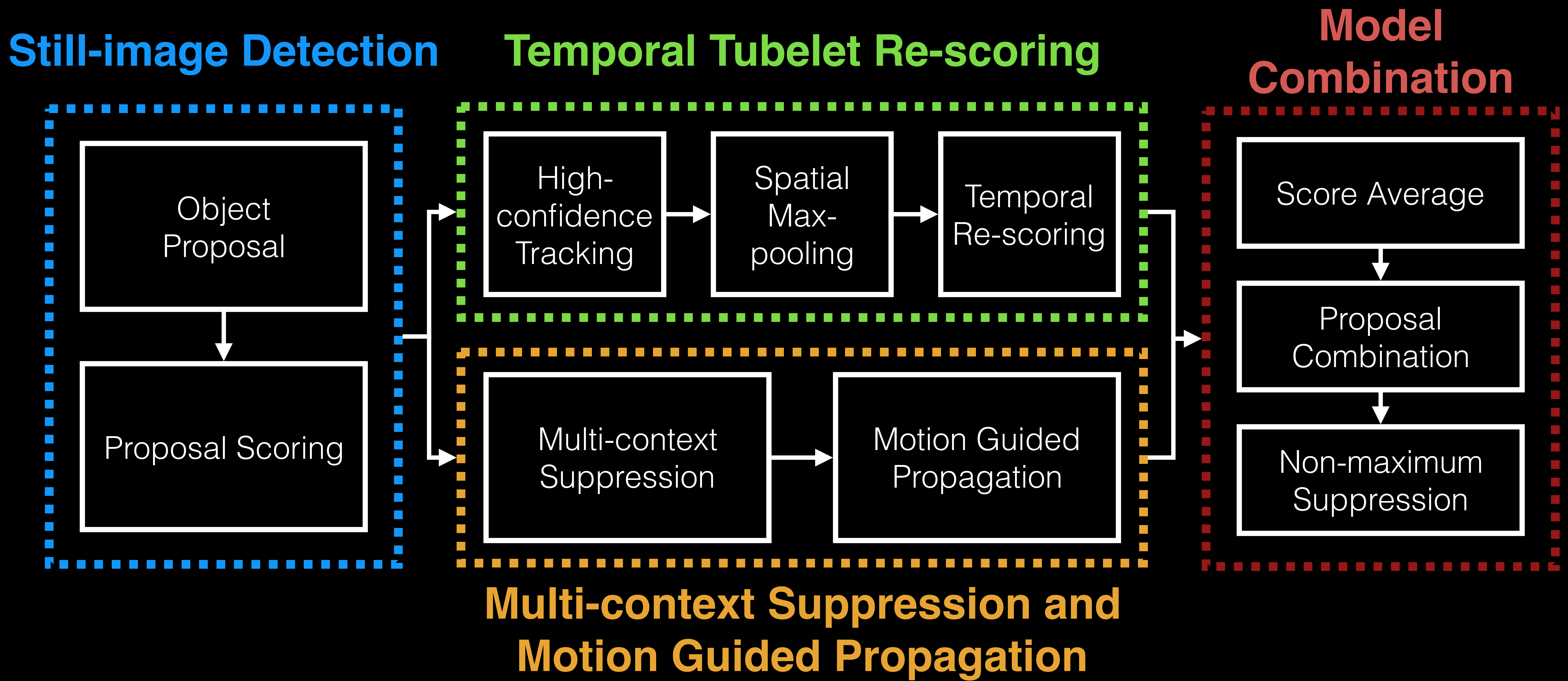
- NMS is used for combining multiple groups of proposals

Model Combination

- Two groups of proposals: 1) Region Proposal Networks (RPN), 2) Selective Search + EdgeBox. Given a group of proposals, their detection scores can be obtained by averaging several models.

- NMS is used for combining multiple groups of proposals

Proposed Framework



Component Analysis

Training Data Configuration

CNN Training Data

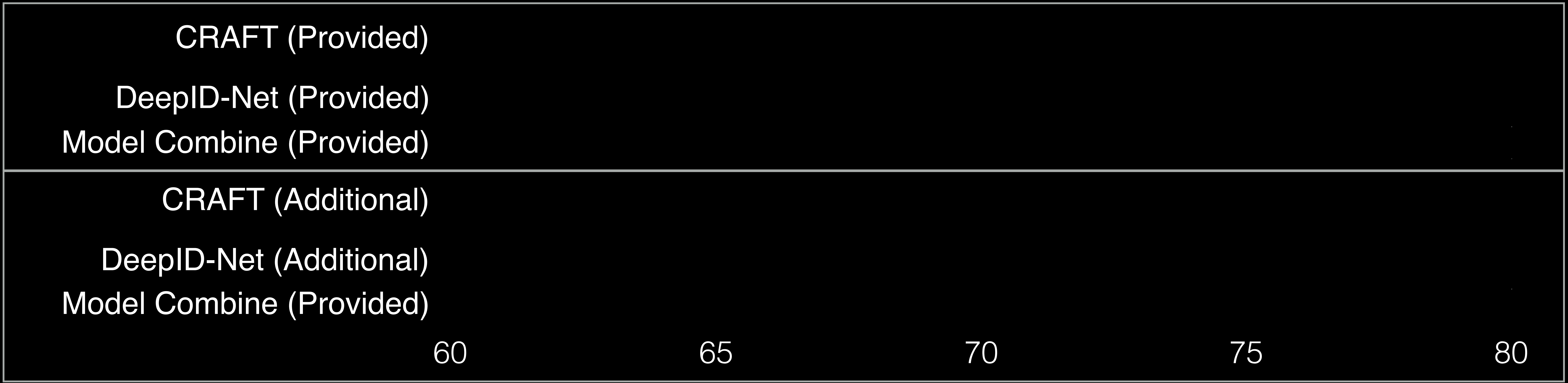
DET:VID Ratio	1:0	3:1	2:1	1:1	1:3
MeanAP / %	49.8	56.9	**58.2**	57.6	57.1

SVM Training Data

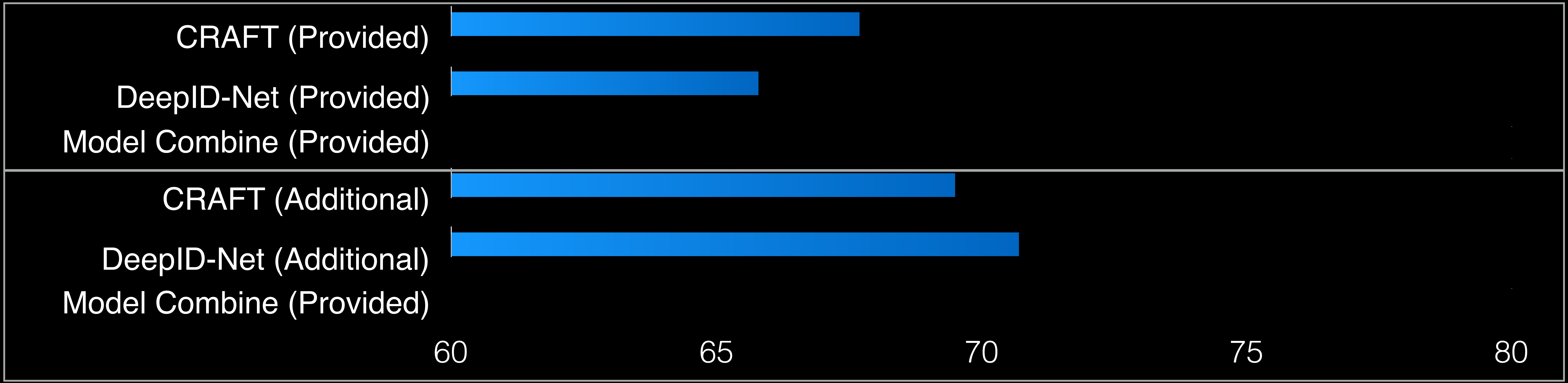
DET Positive	✅	✅	❌	❌	❌	✅
VID Positive	❌	✅	✅	✅	✅	✅
DET Negative	✅	✅	✅	✅	❌	✅
VID Negative	❌	❌	❌	✅	✅	✅
MeanAP / %	49.8	47.1	35.8	51.6	52.3	**53.7**

Framework Components

Framework Components

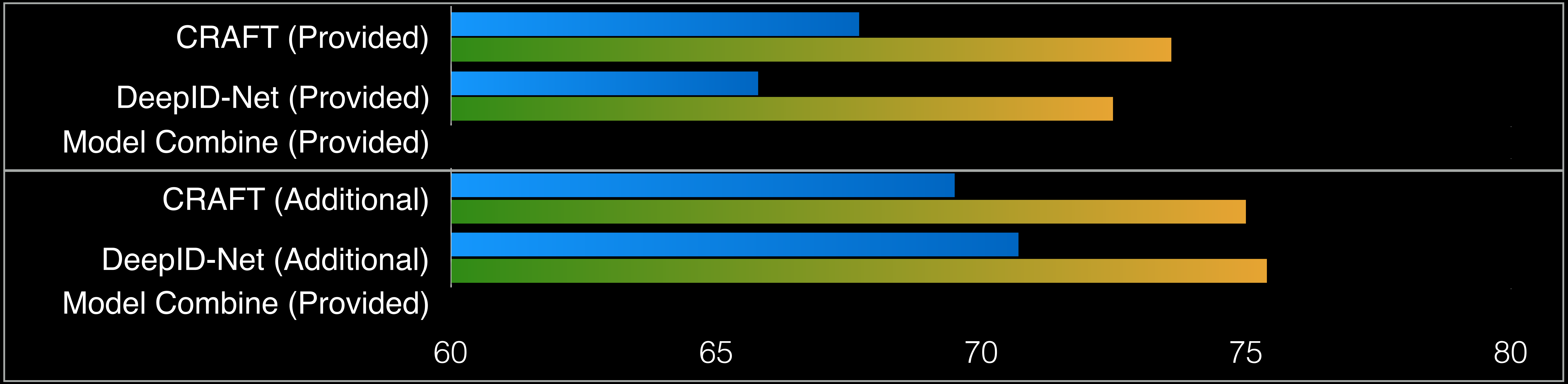


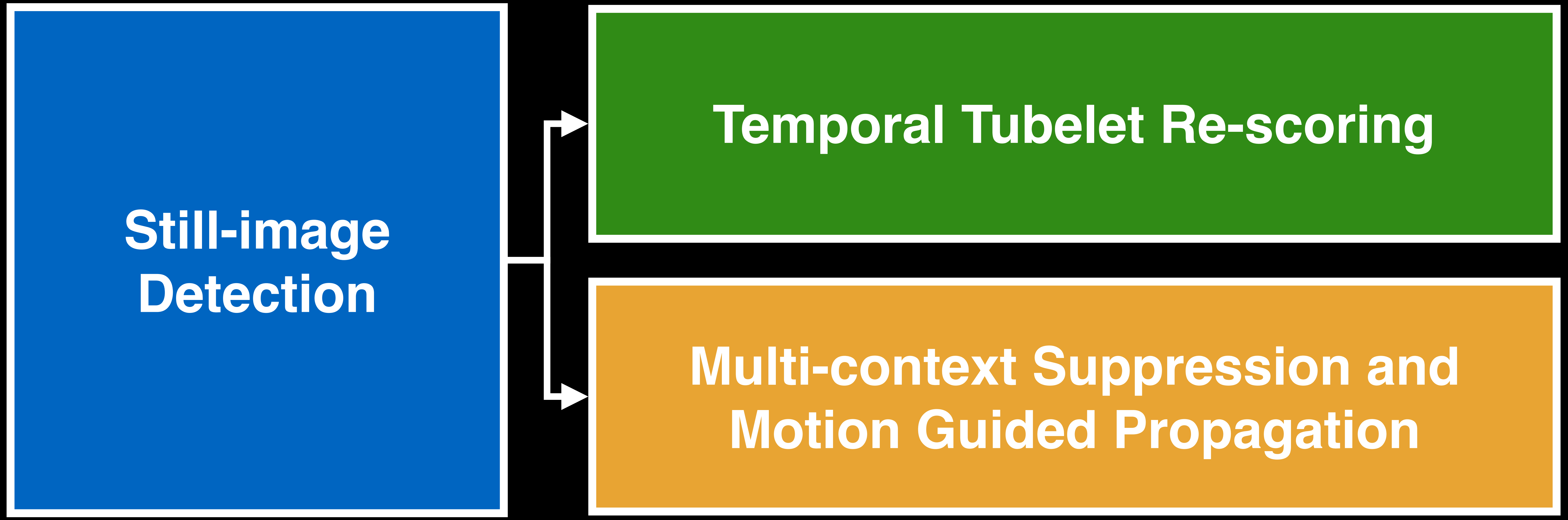
Framework Components



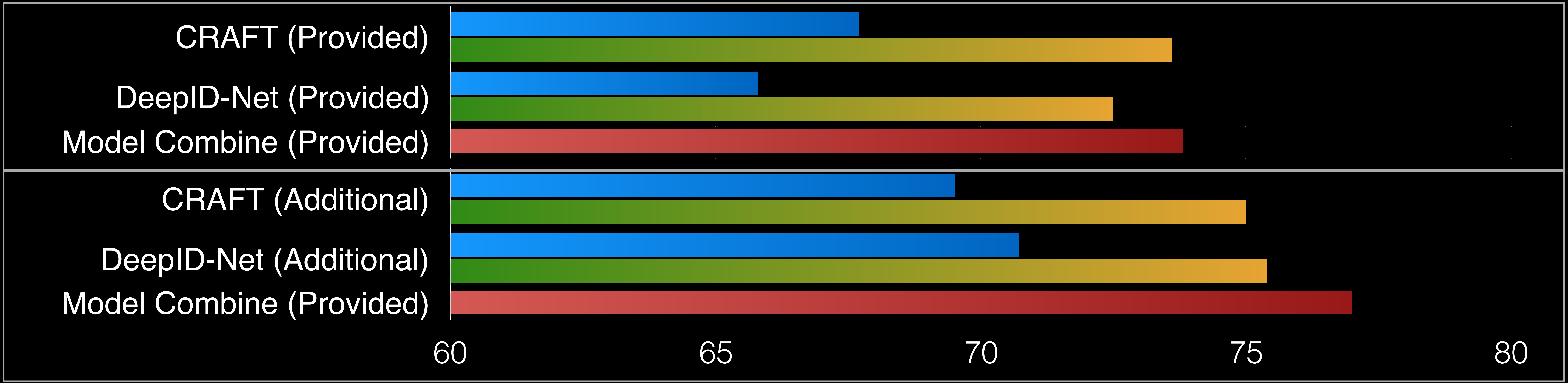
Still-image Detection

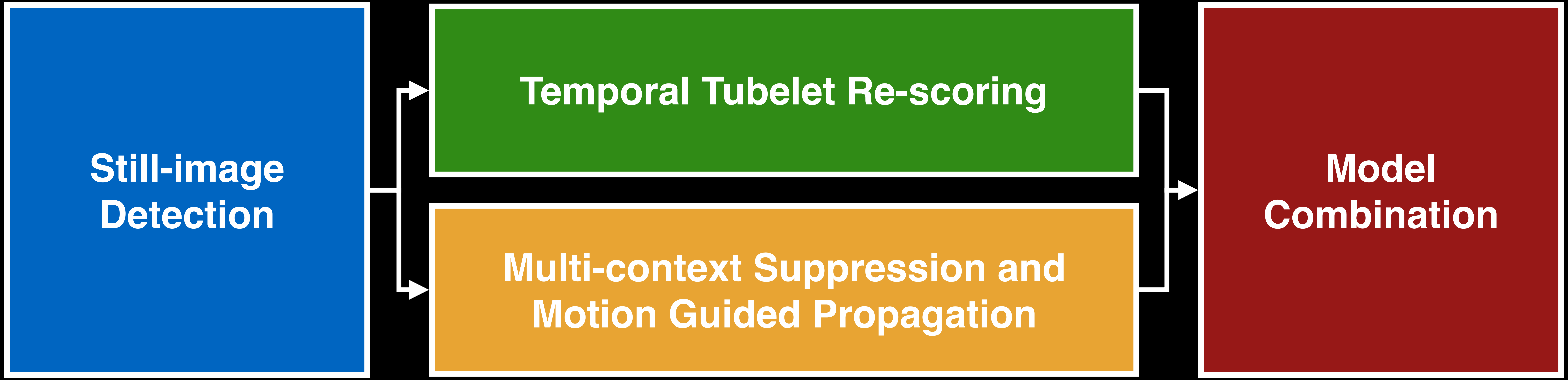
Framework Components





Framework Components





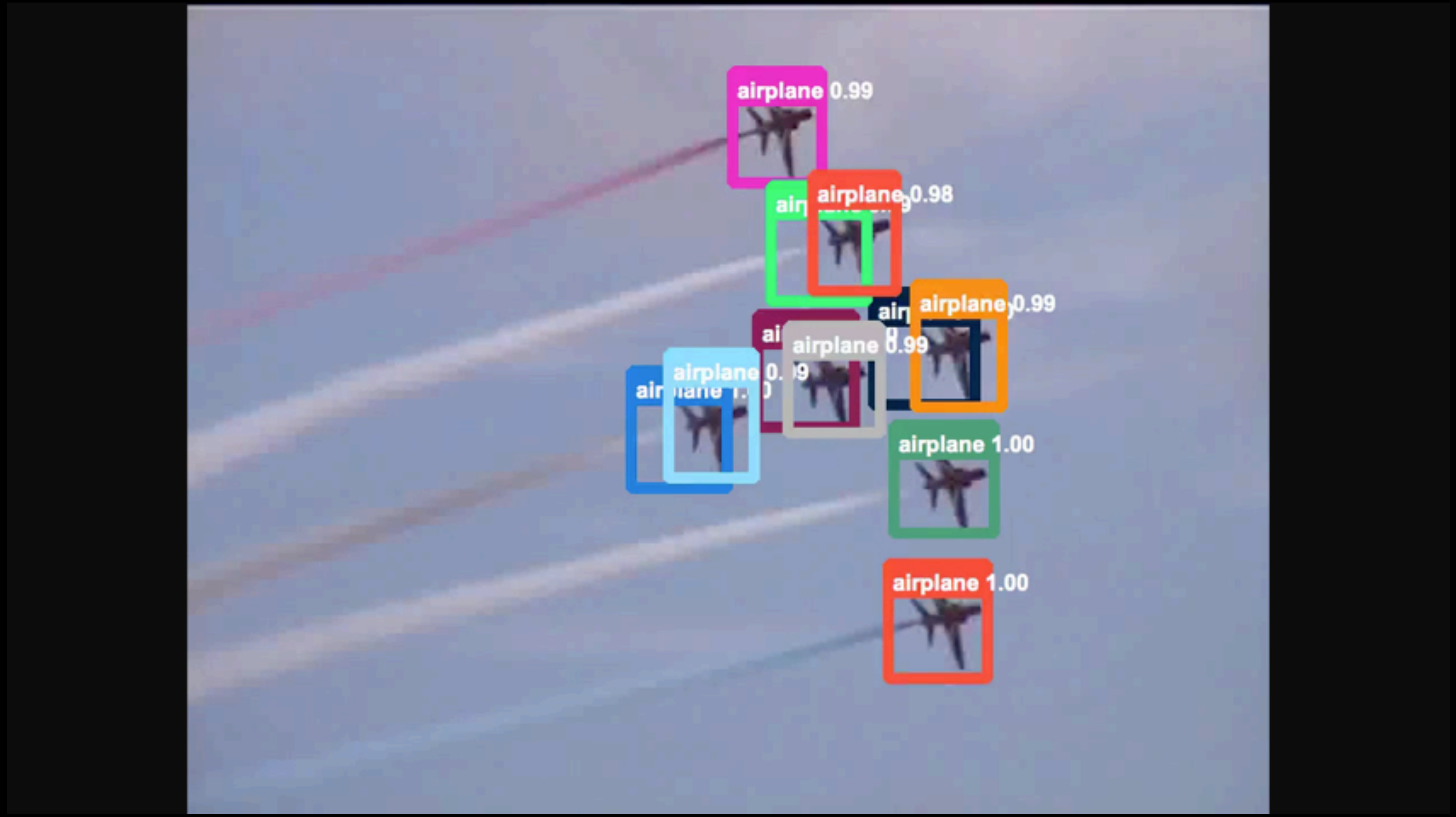
Results

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

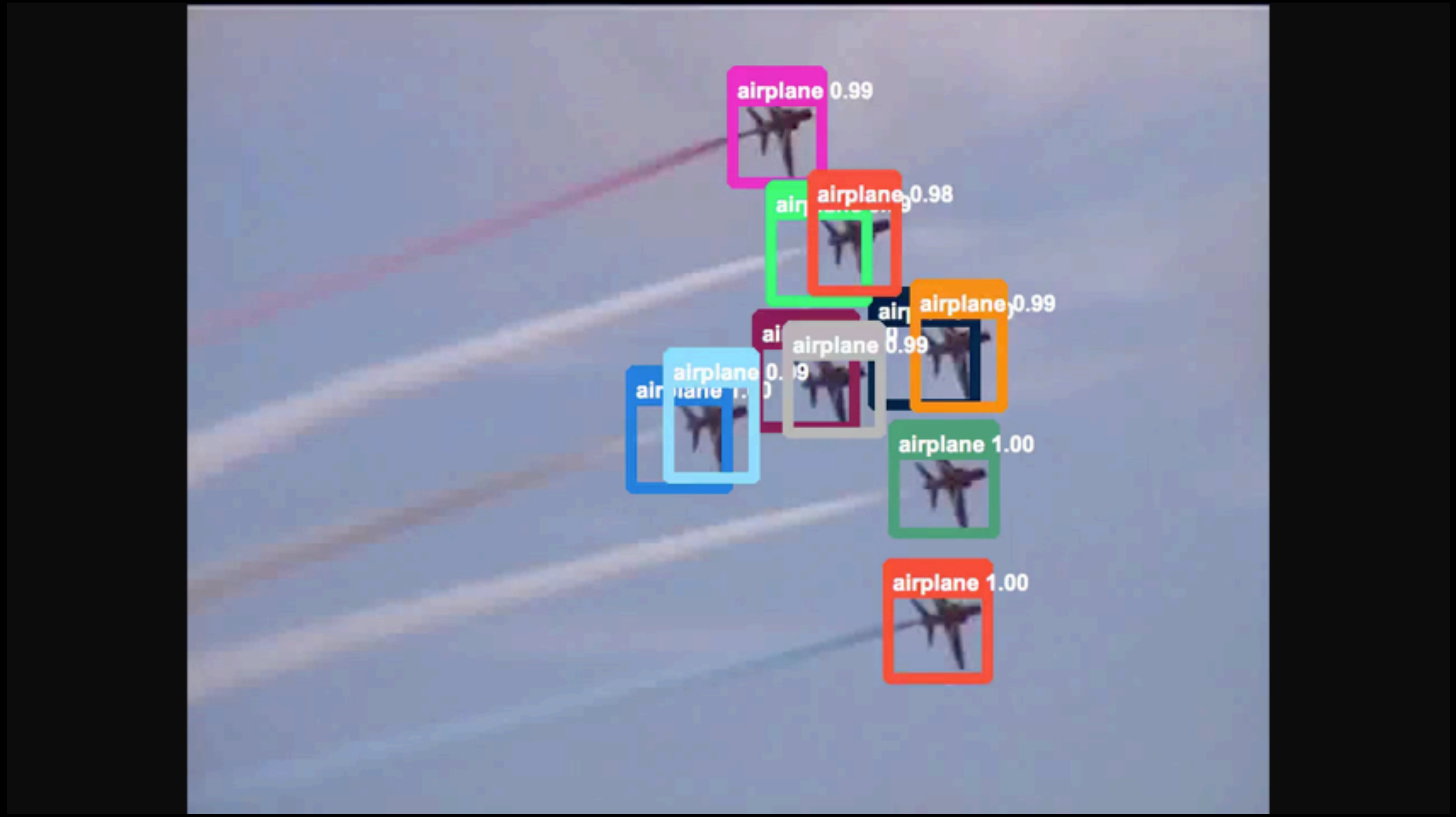
[1] J. Yan, et al. CRAFT Objects from Images, axiv preprint.
[2] W. Ouyang, et al. Deepid-net: Deformable deep convolutional neural networks for object detection. CVPR, 2015.
[3] X. Zeng, et al. Window-Object Relationship Guided Representation Learning for Generic Object Detections , axiv preprint.
[4]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

Results


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

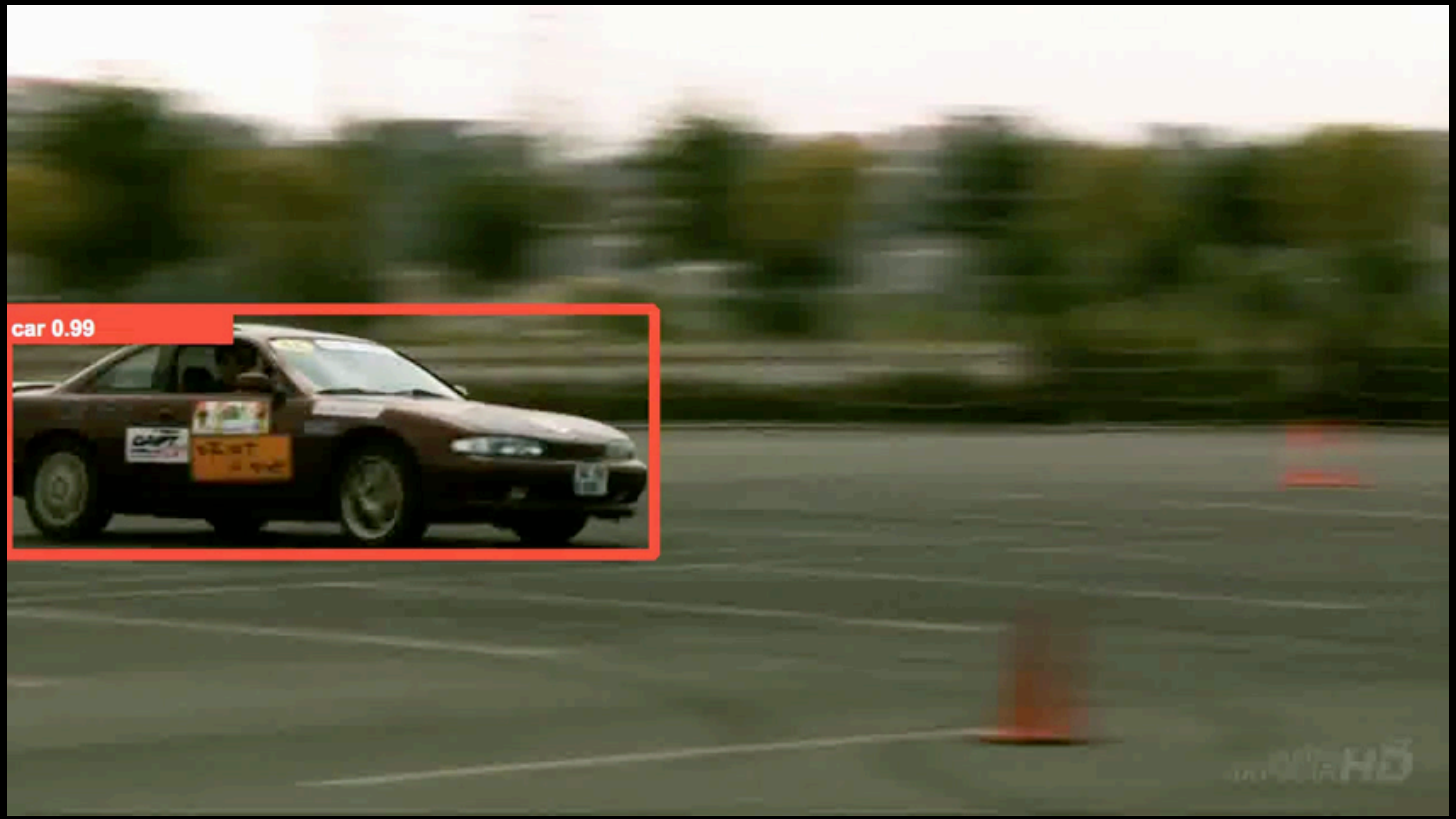

airplane 0.78
airplane 0.85
airplane 0.88
airplane 0.66
airplane 0.72

Results

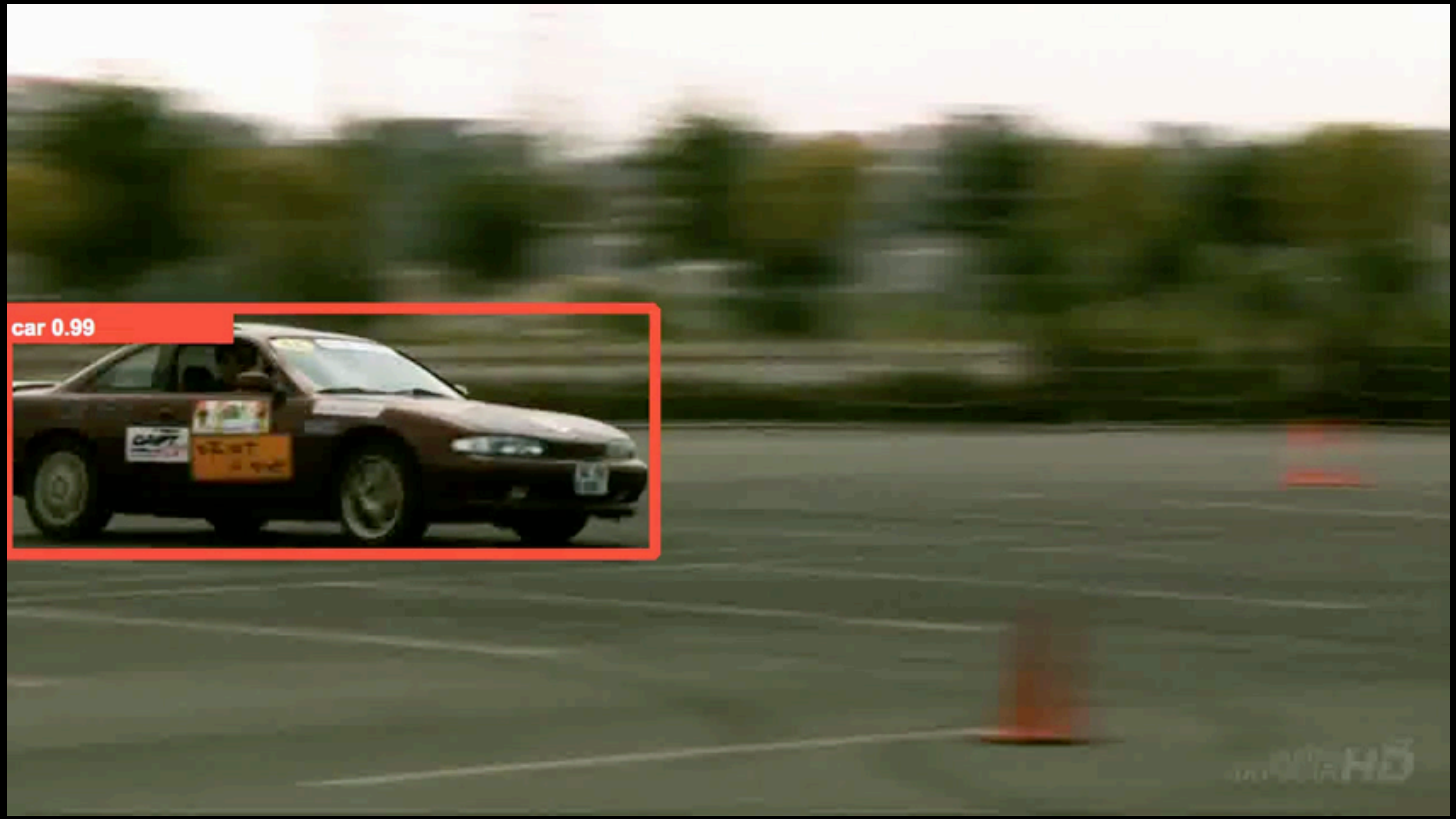

airplane 0.99
airplane 0.98
airplane 0.99
airplane 0.99
airplane 1.00
airplane 1.00


airplane 0.78
airplane 0.85
airplane 0.88
airplane 0.66
airplane 0.72

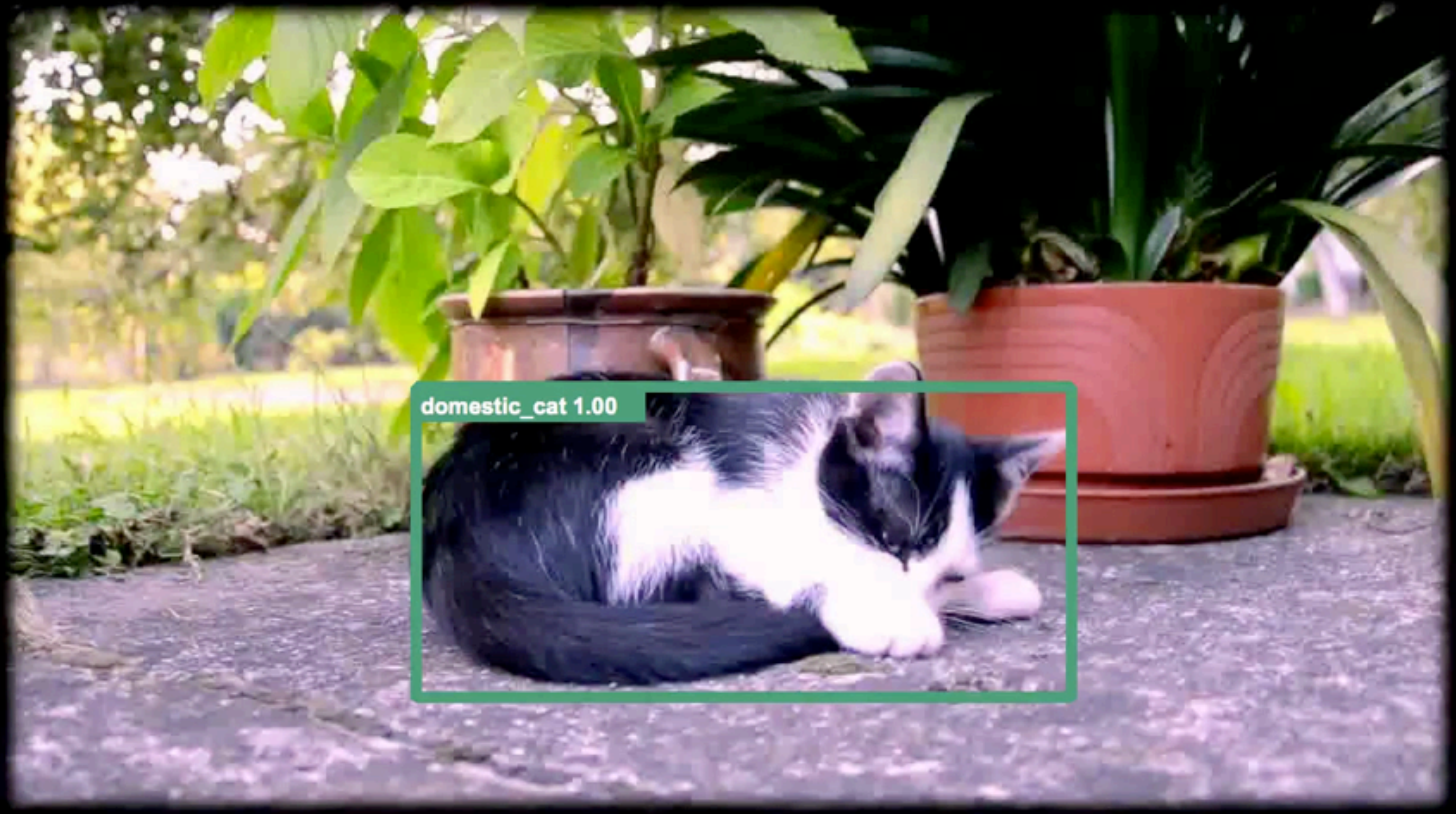
Results

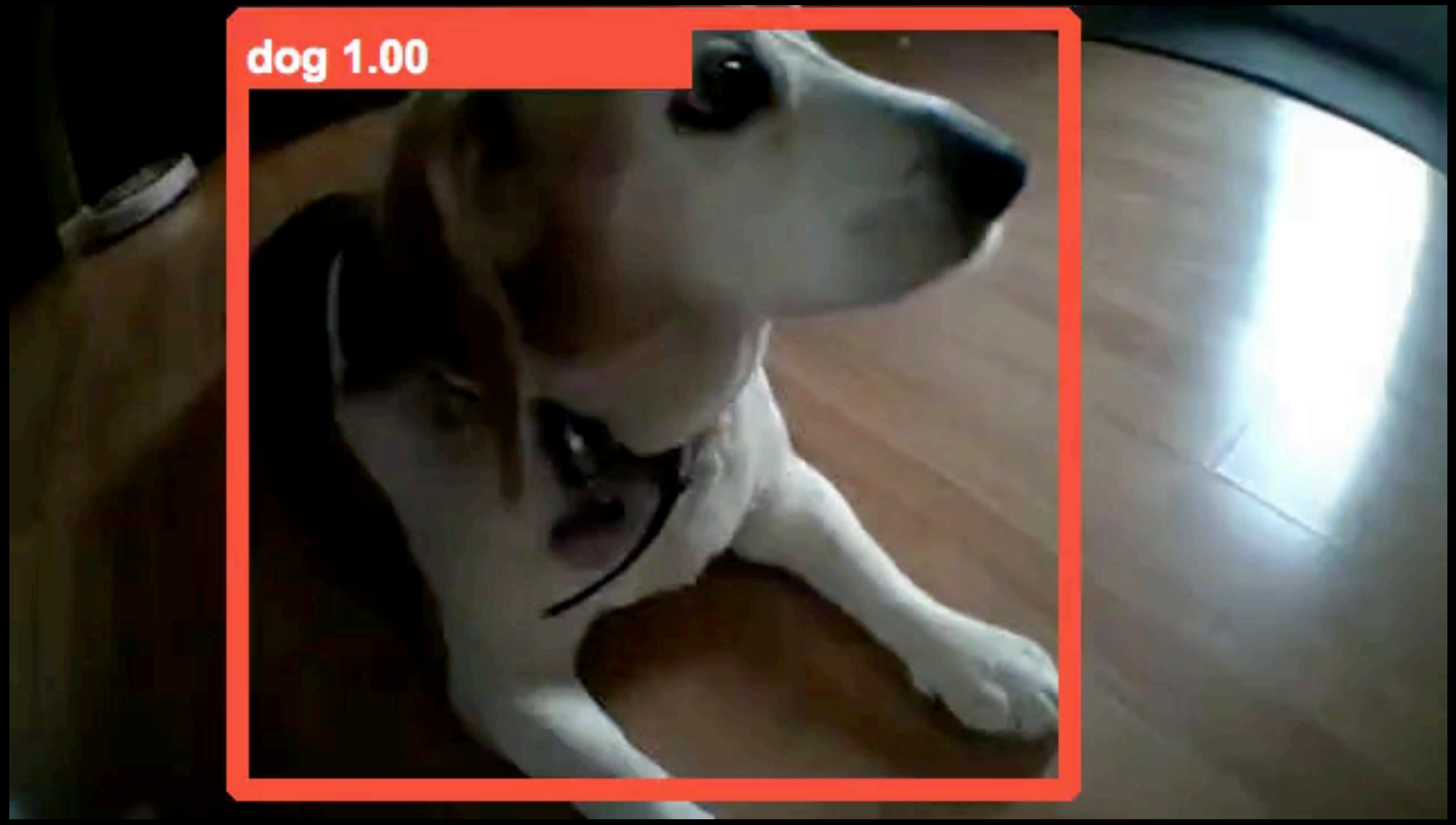


Results

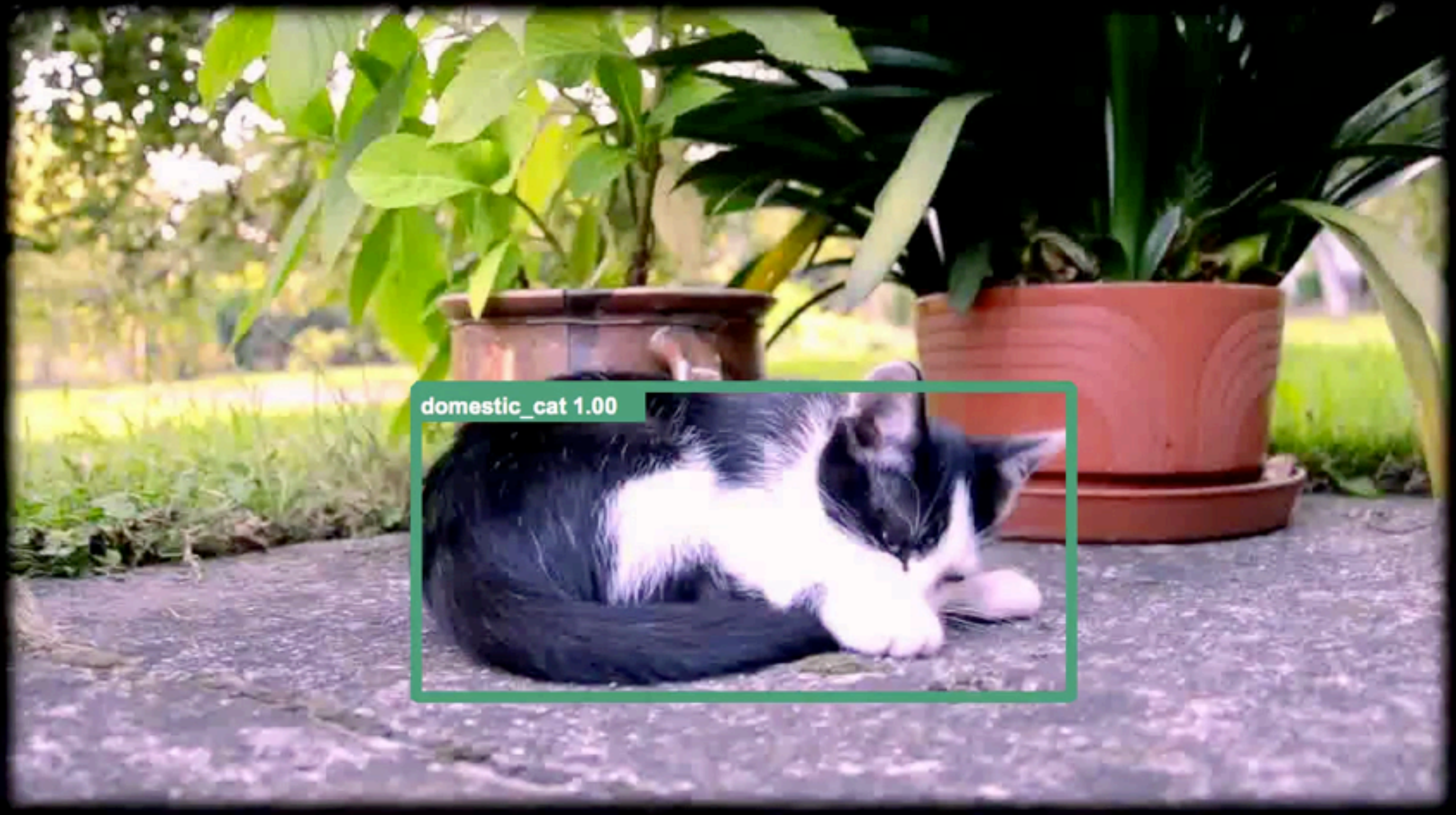


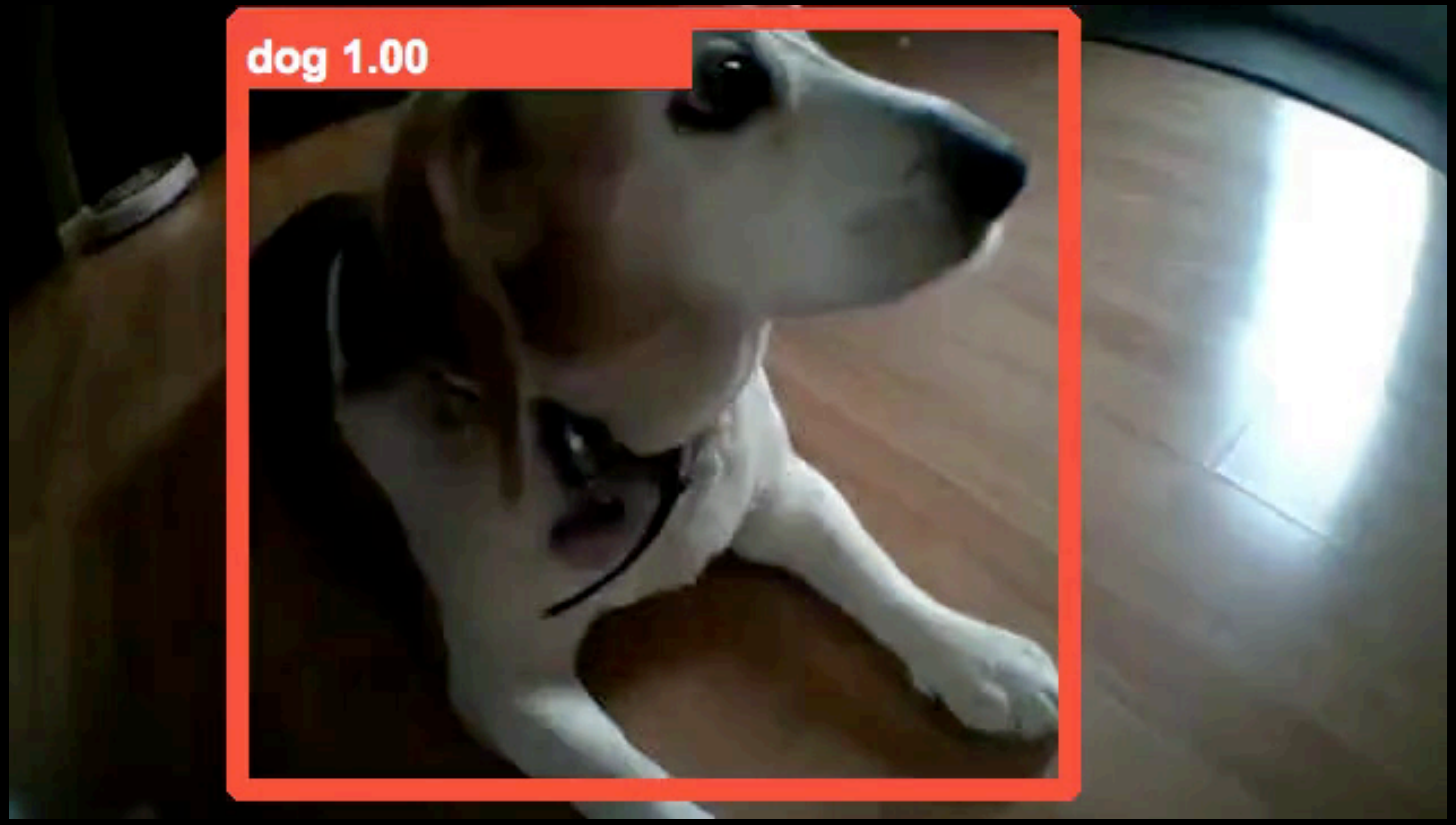
Results



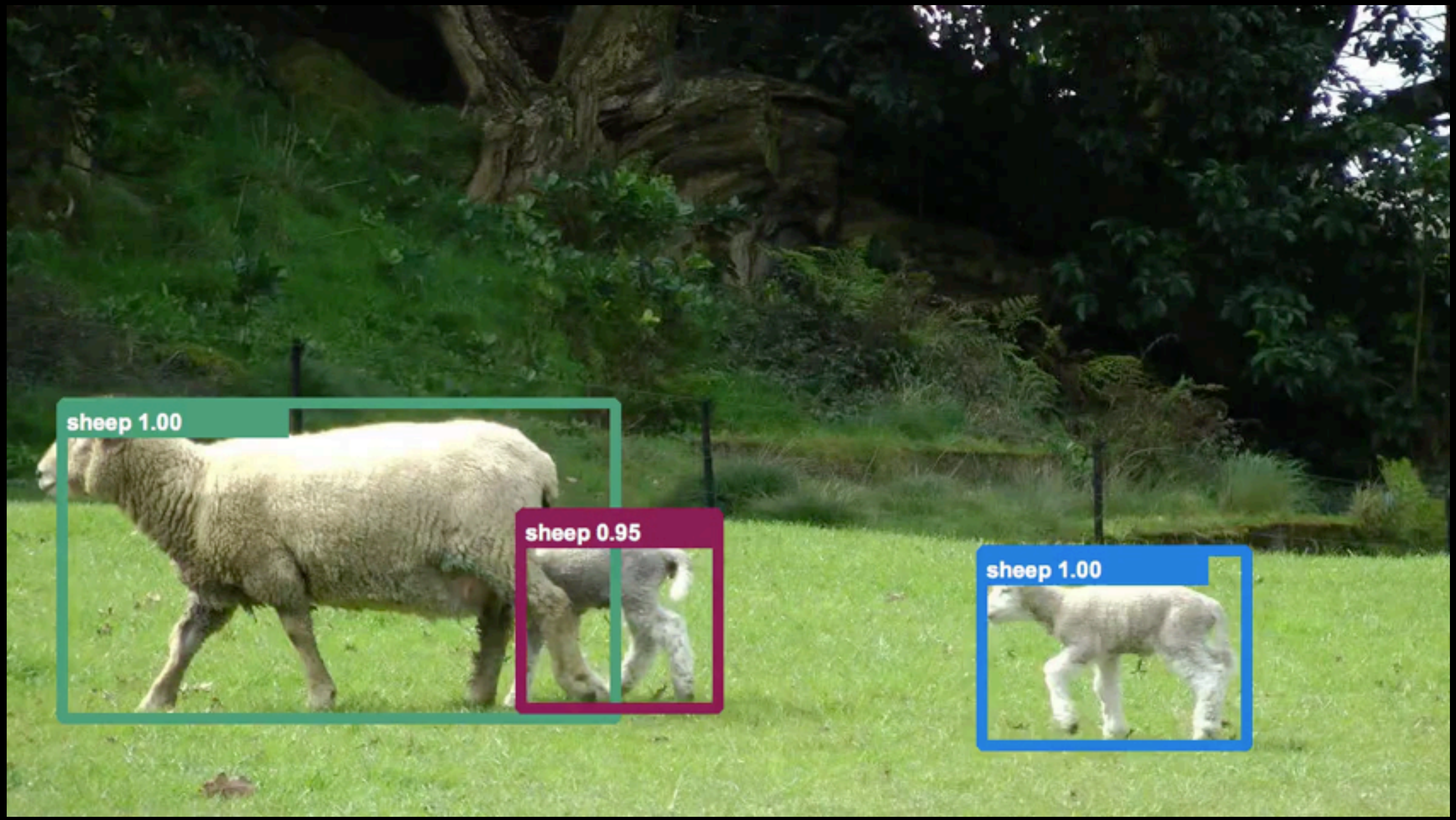


Results

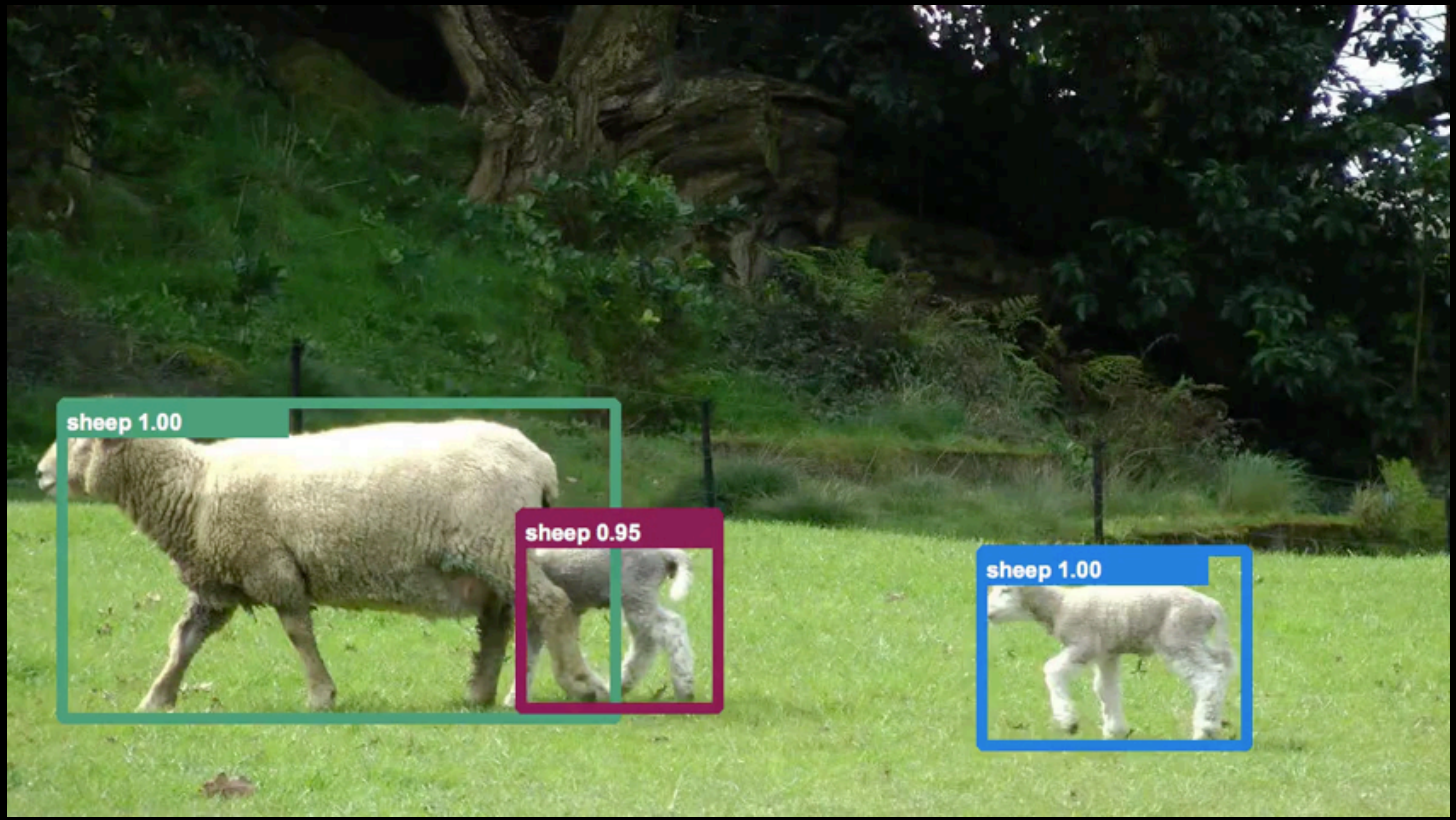




Results



Results



Results

Results

Thank You!

Questions?