MCG-ICT-CAS: Multimedia Computing Group

中国科学院计算技术研究所
Institute of Computing Technology, Chinese Academy of Sciences

IMAGENET ECCV'16 EUROPEAN CONFERENCE ON COMPUTER VISION October 8 – 16, 2016 | Amsterdam | the Netherlands

2nd ImageNet and COCO Visual Recognition Challenges Joint Workshop

MCG-ICT-CAS Object Detection at ILSVRC 2016

Tang Sheng, Li Yu, Wang Bin, Xiao Junbin, Zhang Rui, Zhang Yongdong, Li Jintao

Corresponding Email: ts@ict.ac.cn

Institute of Computing Technology, Chinese Academy of Sciences

October 9th, 2016

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

Team Members


Tang Sheng


Li Yu


Wang Bin


Xiao Junbin

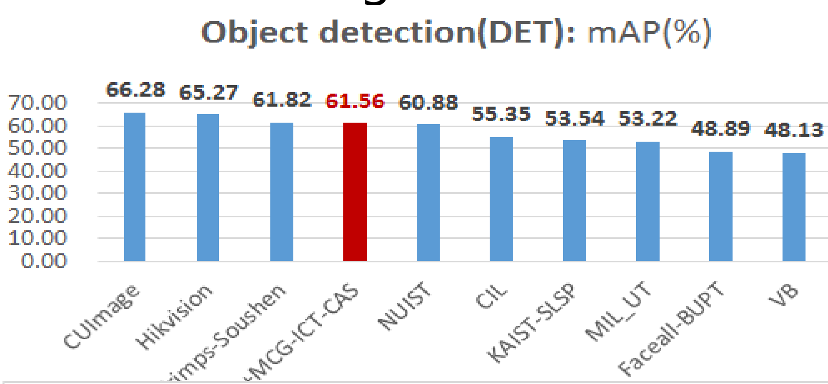

Zhang Rui

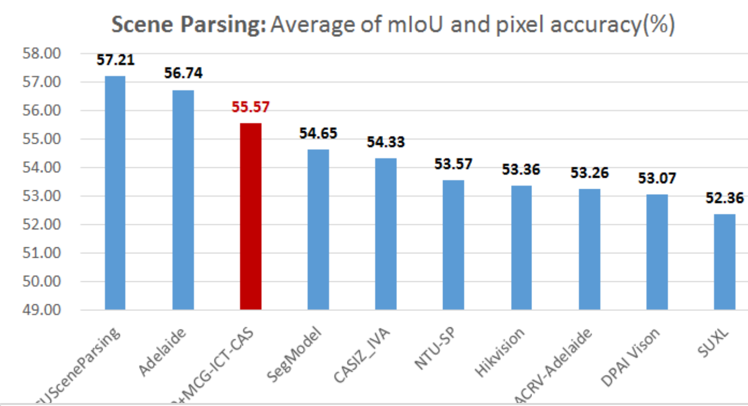

Zhang Yongdong

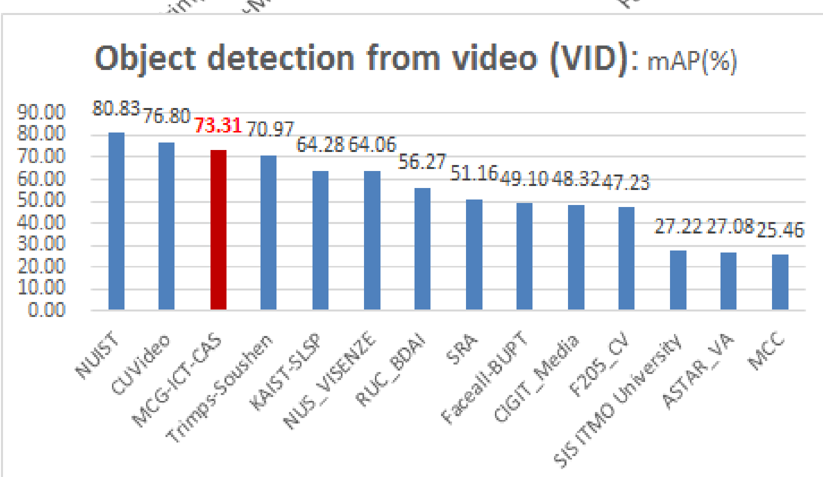

Li Jintao

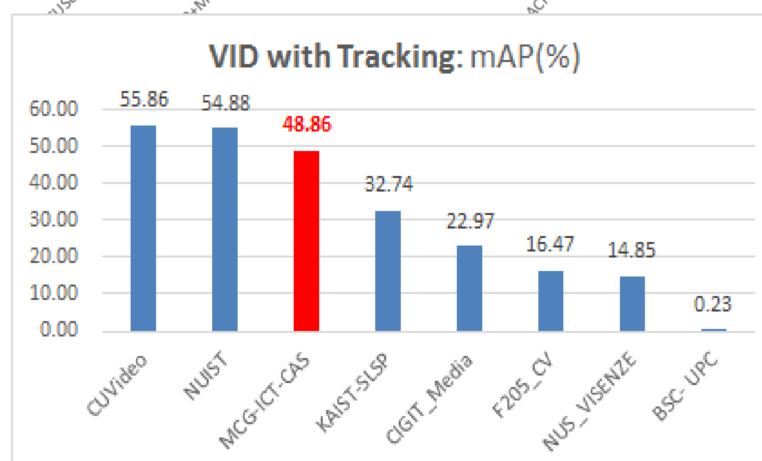
Results of our 3 tasks

- Three tasks with provided data:
  - Object detection (DET): 4th
  - Object detection from video (VID):3rd
  - Scene Parsing: 3rd









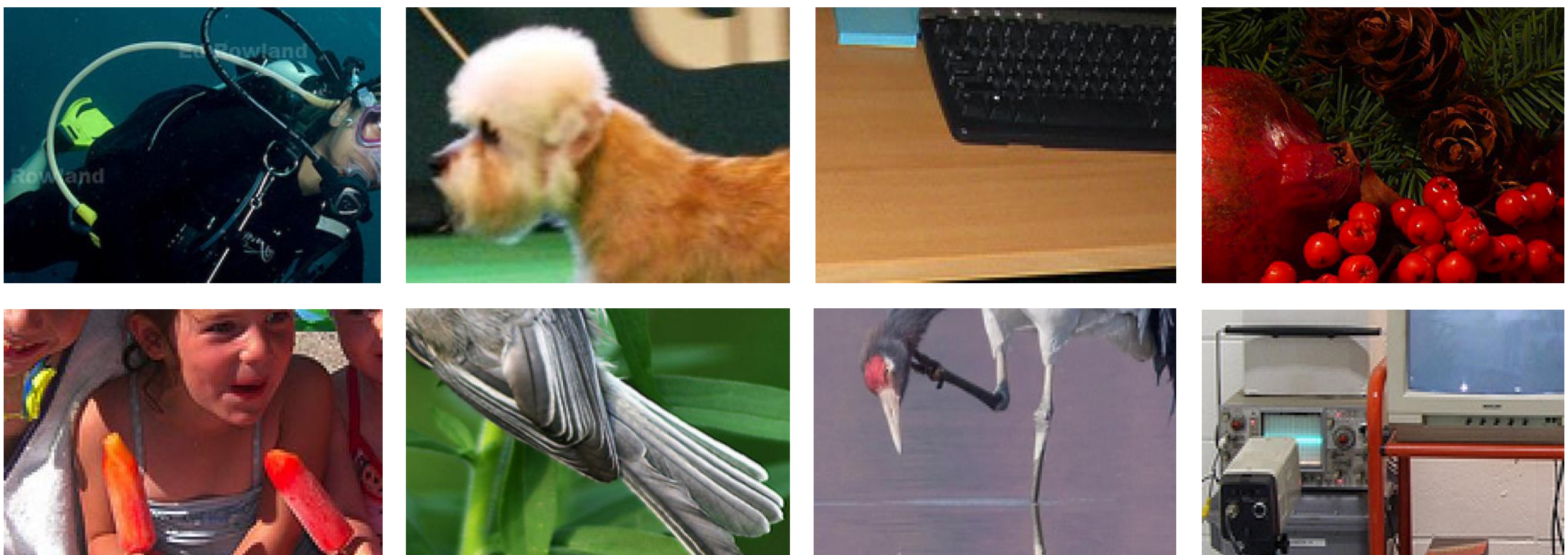
Object Detection (DET)

DET: Overview

- **Improvements of loss function**
  - Implicit sub-categories of background class
  - Sink class when necessary
- Other training and testing tricks
  - Segmentation feature
  - Dilation as context
  - Multi-scale testing
  - Box refinement && box voting

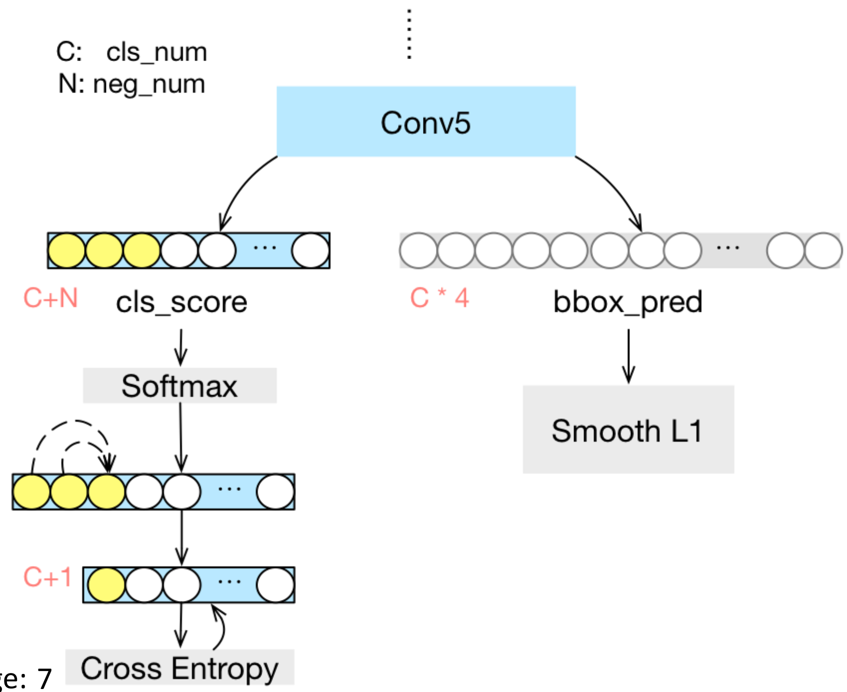
Implicit subcategories of BG

- Background(BG): **Indiscrimination**
  – As ONE class equally as other object classes
  – But: varies greatly
  – Unreasonable to describe as one pattern



Implicit subcategories of BG

- Add N output nodes in last FC layer
  - Represent N subcategories of BG, Cross entropy loss
  - Allocate more parameters to many BG class by adding latent BG subclasses. Improve identification capability.
  - Voc 2007 with Resnet50: ↑ **1%**

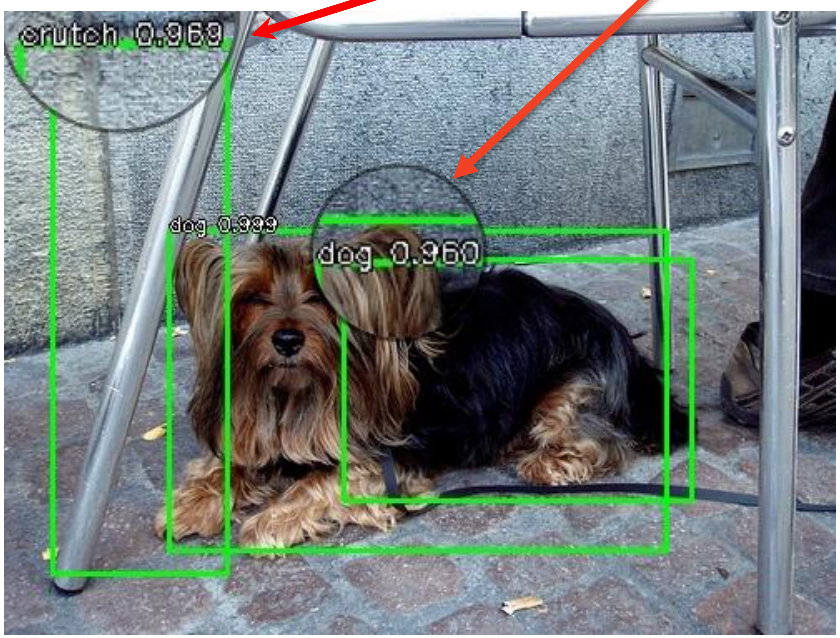


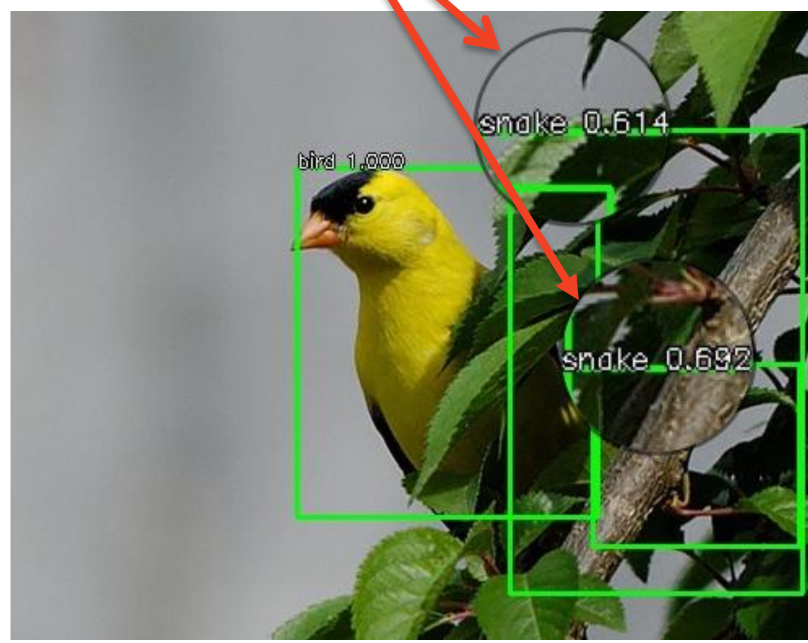
Model	mAP on VOC07
Res50 baseline	77.5%
Res50+Implicit Sub categories-5 nodes	78.5% ↑**1%**

Sink class when necessary

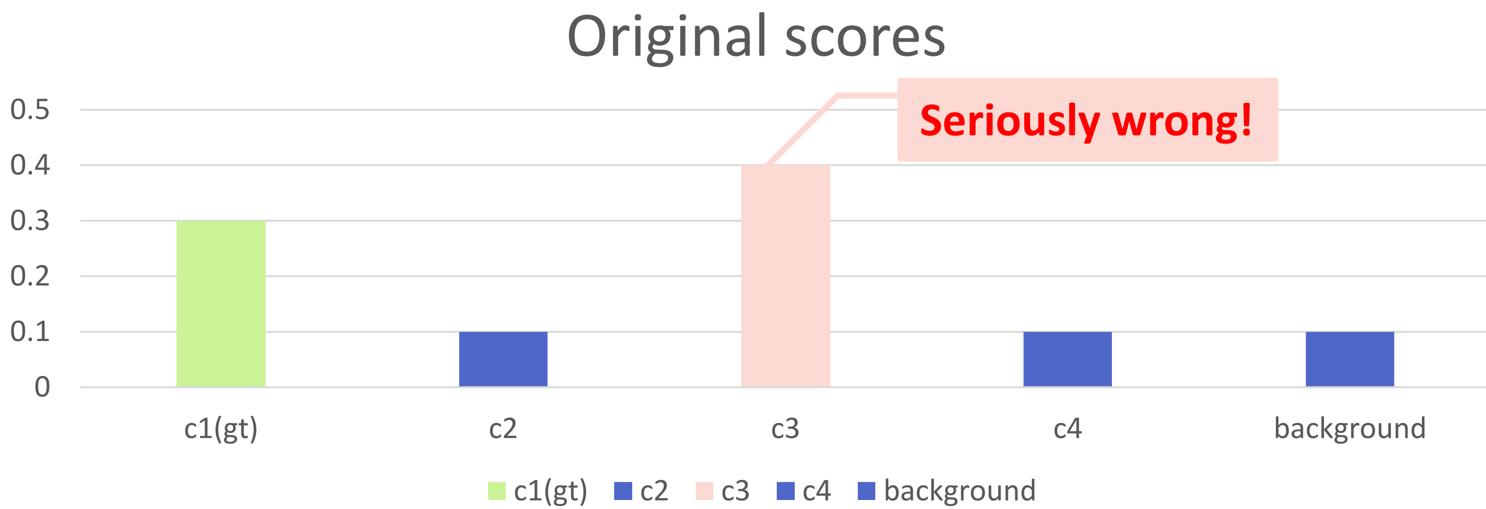
- **Flow diversion** of score to wrong classes
  - Scores of true classes become relatively low

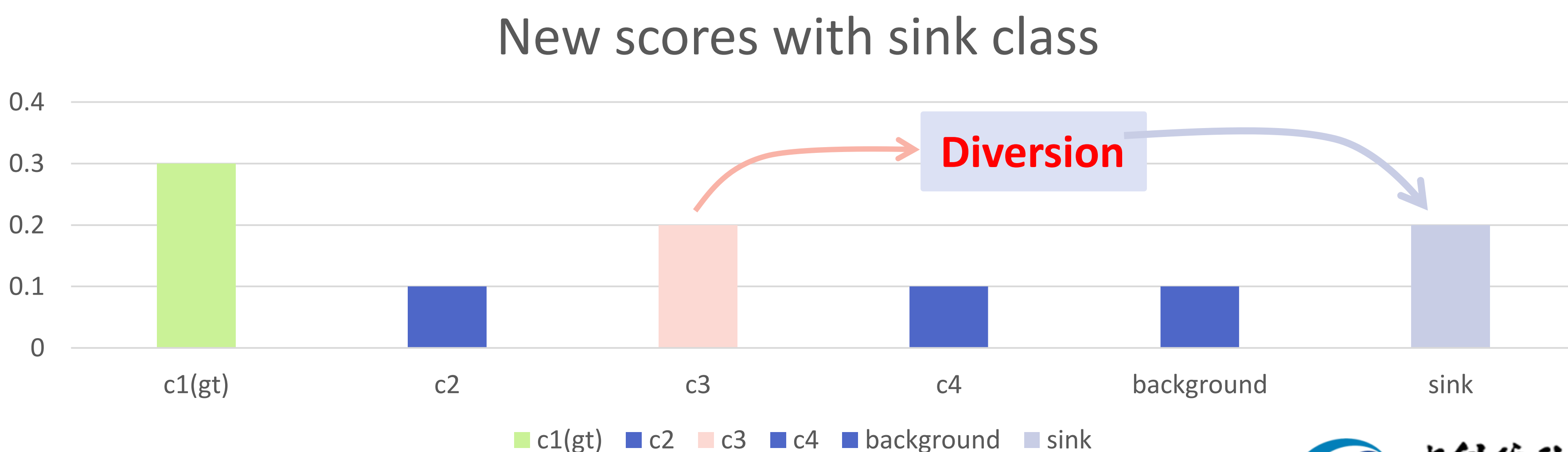
Score: High, wrong!





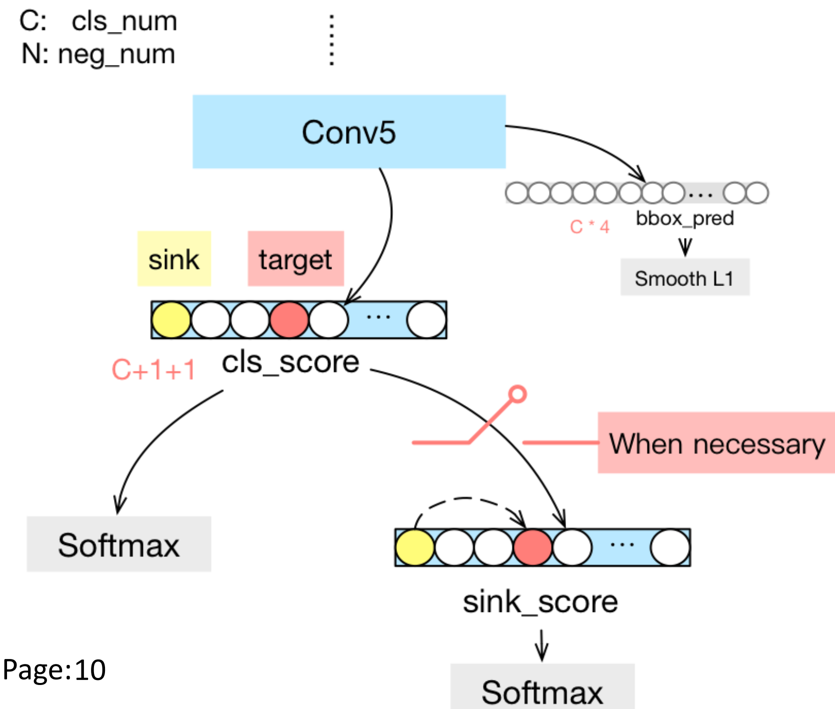
Sink class when necessary





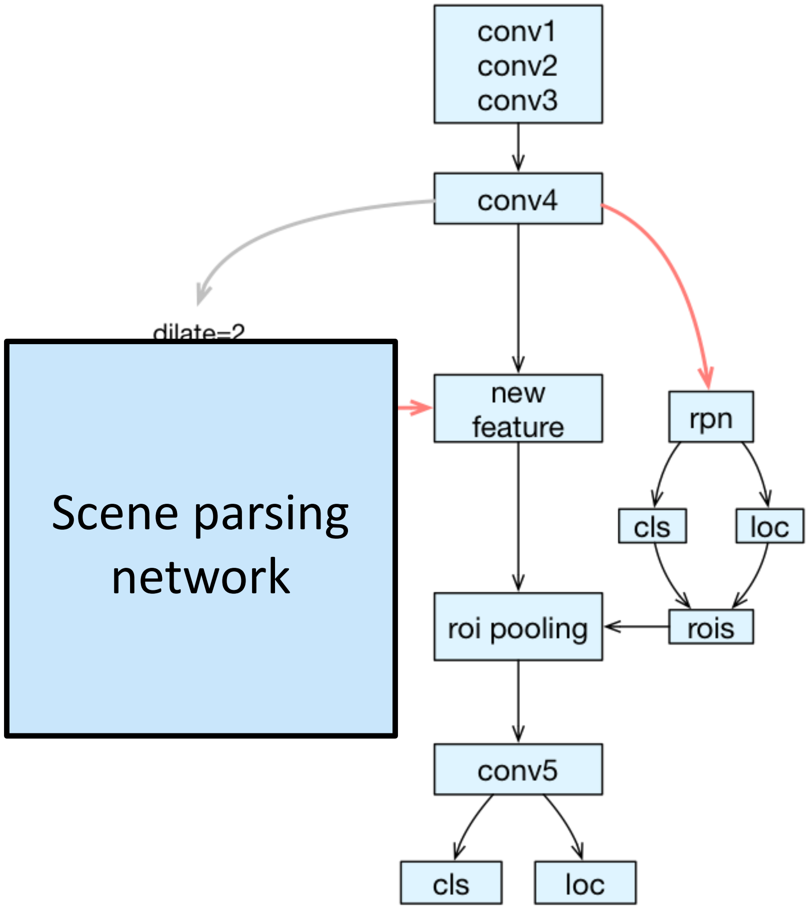
Sink class when necessary

- Add a Sink class
  - Optimize: Minimize((loss (target) && loss (target+sink)), only if all the Top-*K* results are wrong during training
  - Flow diversion of high wrong scores during testing
  - Give true class with low scores more chances to win
  - Voc 2007 with Resnet50: ↑ **0.7%**

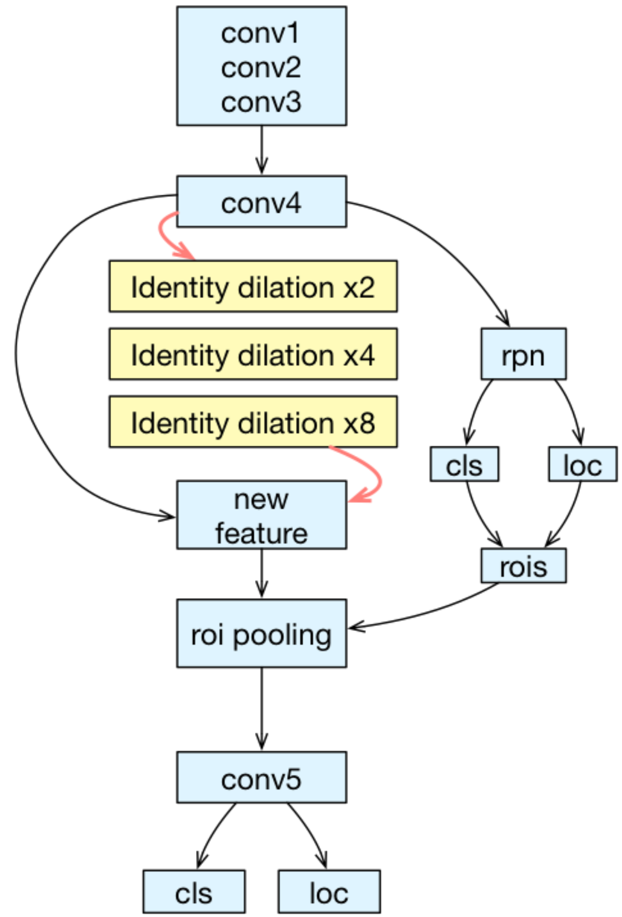


Model	mAP on VOC07
Res50 baseline	77.5%
Res50+Sink-top5	78.2% ↑**0.7%**

Tricks: Segmentation+Dilation



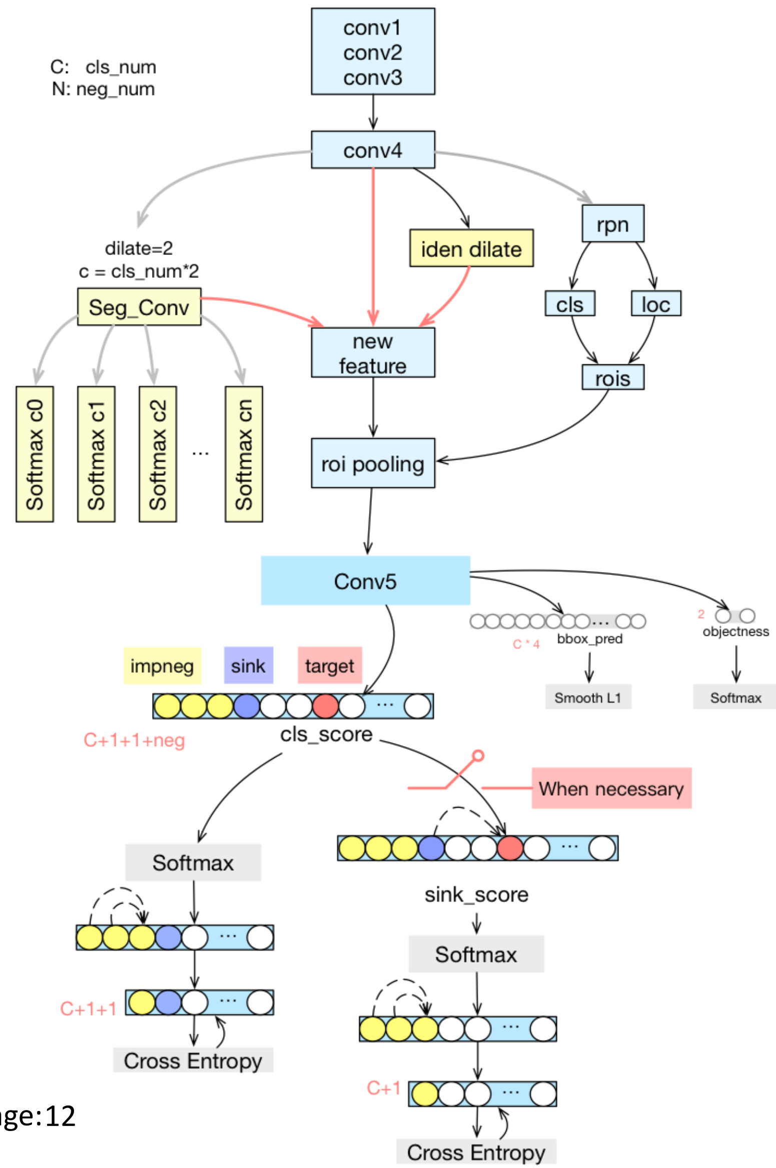
Segmentation feature[2]



Dilation as context[3]

VOC 2007: Segmentation ↑ 0.8%, Dilation ↑ 0.8%

All together



Model	mAP on VOC07
Res50 baseline	77.5%
+dilation as context +sink class when necessary +implicit sub-categoris +segmentation feature +testing tricks	80.62% +3.1%

Results on ILSVRC 2016 DET

Model	mAP on val	mAP on test
Res200 baseline+tt(testing tricks)	62.9	59.1
Res200+all+tt	62.3	57.7 T_T
Res101	58.0	
Res152	57.8	
Res200	60.5	
Res101+all	57.8	
Res152+all	58.0	61.6(with tt)
Res200+all	60.3	
Res152+scene parsing feature	55.6	
Res200+all in half	60.2	
Res200+logistic	57.9	

Results on ILSVRC 2016 DET

Model	mAP on val2	mAP on test
Res200 baseline+tt(testing tricks)	62.9 ↑ 2.4%	59.1 ↑ 0.3%
Res200+all+tt	62.3	57.7
Res101+Res152+Res200	64.0 ↑ 0.4%	60.6 ↓1.5%
Res101	58.0	
Res101+all	57.8	
Res152	57.8	
Res152+all	58.0	
Res200	60.5	61.6(with tt)
Res200+all	60.3	
Res152+scene parsing feature	55.6	
Res200+all in half	60.2	
Res200+logistic	57.9	

Results on ILSVRC 2016 DET

Model	mAP on val2	mAP on test
Res200 baseline+tt(testing tricks)	62.9	59.1
Res200+all+tt	62.3	57.7
Res101+Res152+Res200	63.9	60.6
Res101	58.0	61.6(with tt)
Res101+all	57.8	
Res152	57.8	
Res152+all	58.0	
Res200	60.5	
Res200+all	60.3	
Res152+scene parsing feature	55.6	
Res200+all in half	60.2	
Res200+logistic	57.9	

Results on ILSVRC 2016 DET

Model	mAP on val2	mAP on test
Res200 baseline+tt(testing tricks)	62.9	59.1
Res200+all+tt	62.3	57.7
Res101+Res152+Res200	63.9	60.6
Res101	58.0	61.6(with tt)
Res101+all	57.8	
Res152	57.8	
Res152+all	58.0	
Res200	60.5	
Res200+all	60.3	
Res152+scene parsing feature	55.6	
Res200+all in half	60.2	
Res200+logistic	57.9	

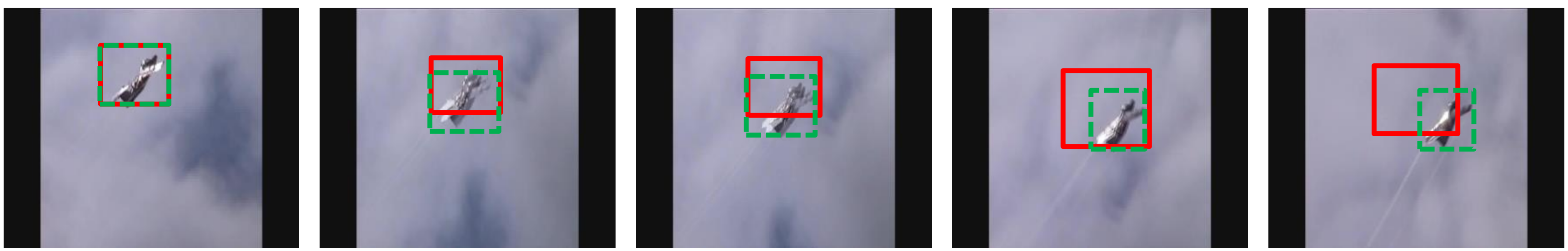
Results on ILSVRC 2016 DET

Model	mAP on val2	mAP on test
Res200 baseline+tt(testing tricks)	62.9	59.1
Res200+all+tt	62.3	57.7
Res101+Res152+Res200	63.9	60.6
Res101	58.0	
Res101+all	57.8	
Res152	57.8	
Res152+all	58.0	
Res200	60.5	61.6(with tt)
Res200+all	60.3	
Res152+scene parsing feature	55.6	
Res200+all in half	60.2	
Res200+logistic	57.9	

Object Detection from Video (VID)

Motivation: Tracking-based

- VID: Challenging task
  - Frame detection performance & adjacent motion information
  - Tracking-based tubelet generation is an effective solution



Ground Truth

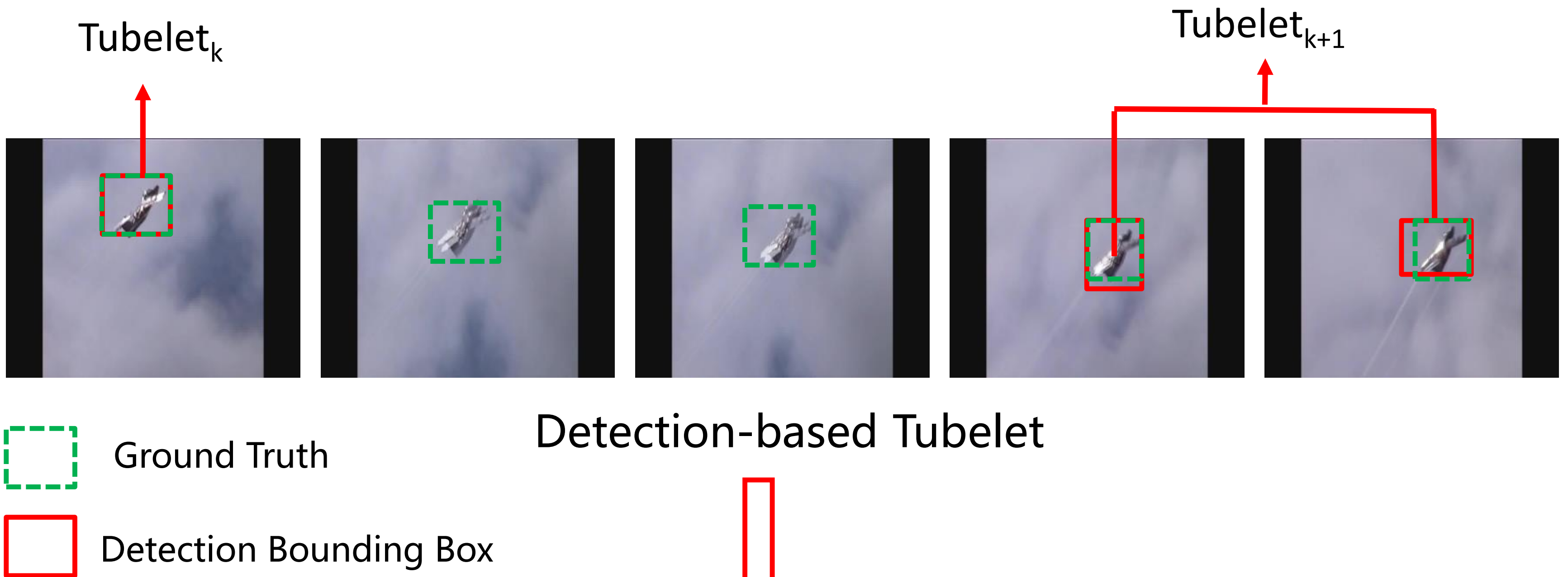
Tracking Bounding Box

Tracking-based Tubelet

Drifting location!

Our Detection-based

- Detection Box Sequentialization
- Adjacent Checking with optical-flow

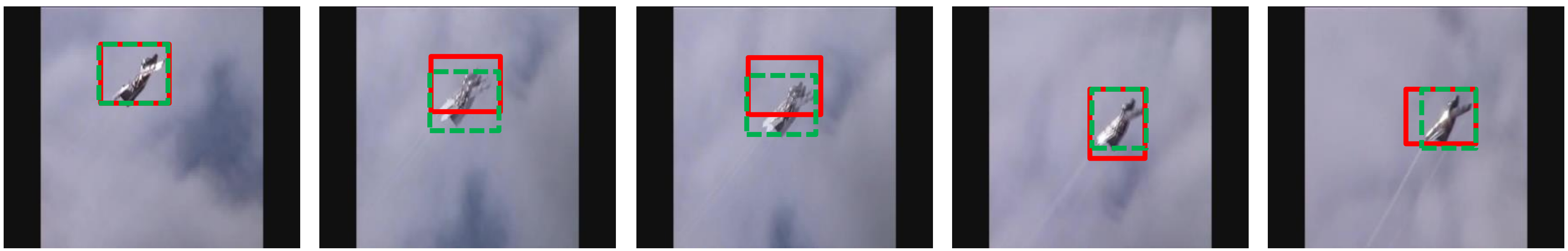


Detection-based Tubelet

Target missing & discrete trajectory!

Our DAT Framework

- **DAT** tubelet generation & fusion framework
  - Tubelet generation: complementary Detection And Tracking (DAT)
  - Focused on precision & recall respectively
  - Followed by a novel tubelet merging method



Ground Truth

Detection Bounding Box

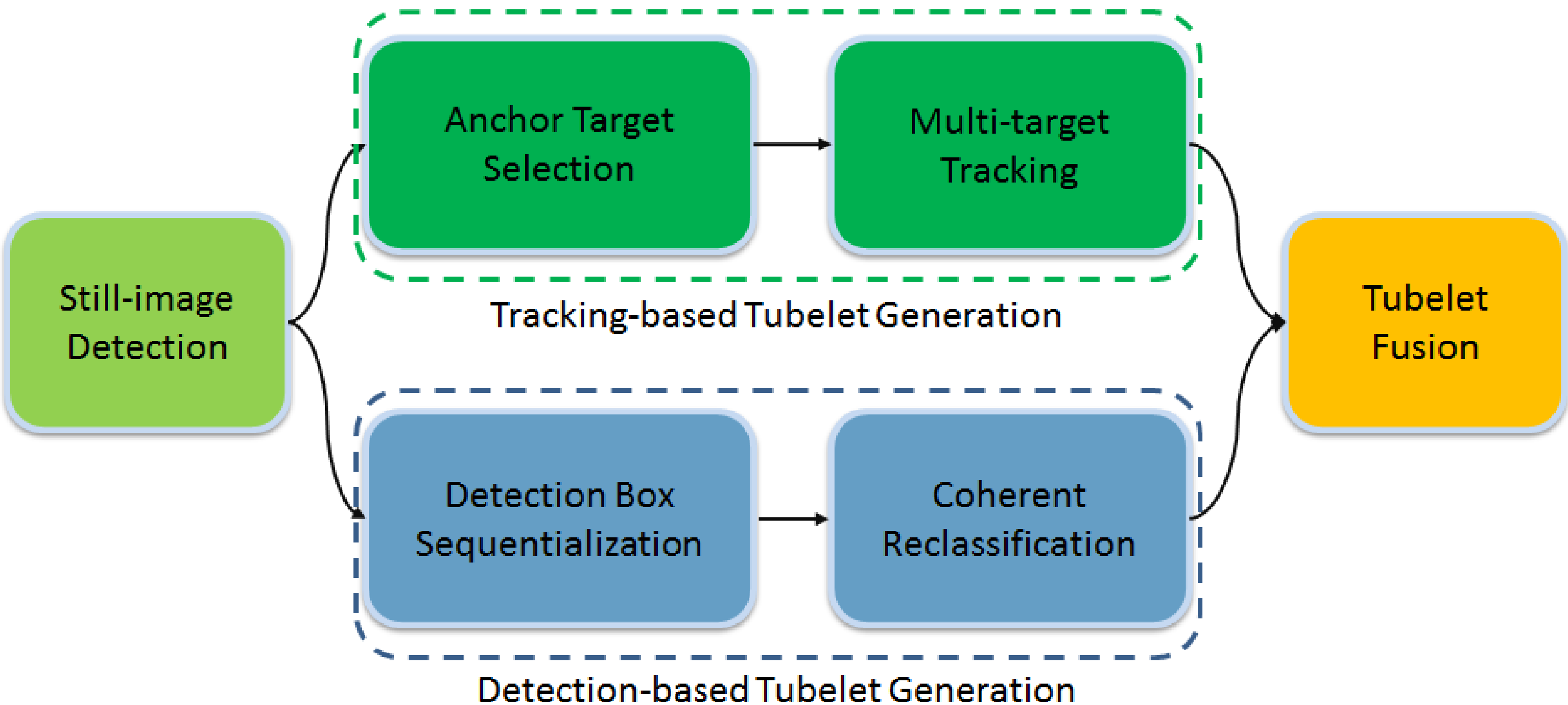
Fusion Tubelet

Improve location quality &

Recall missing objects

VID: Overview

- Two main contributions:
  - Tubelet generation: sequentialize detection box with optical-flow
  - Overlapping and successive tubelet fusion

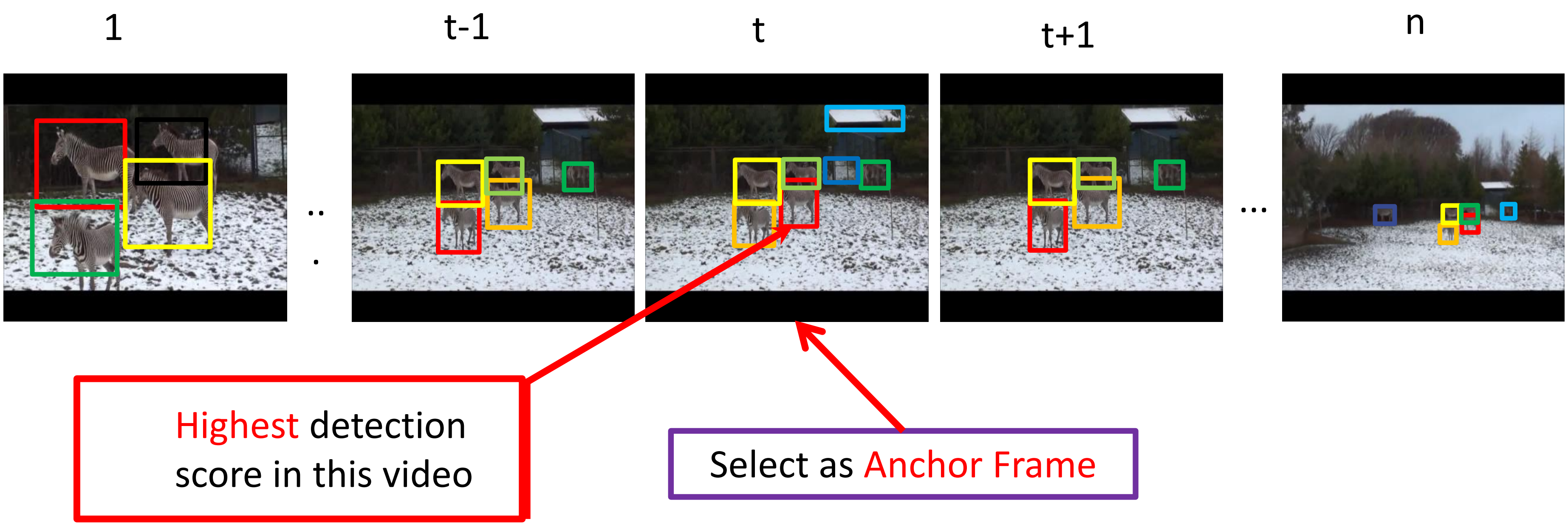


Still-image Detection

- Training Data
  - DET train&val + VID train (1/6)
- Architecture
  - Faster R-CNN [1] + ResNet [4]
  - Add RPN anchor with size = 64
- Model Ensemble
  - ResNet101, ResNet200
  - Weighting average for coordinate position and category score.

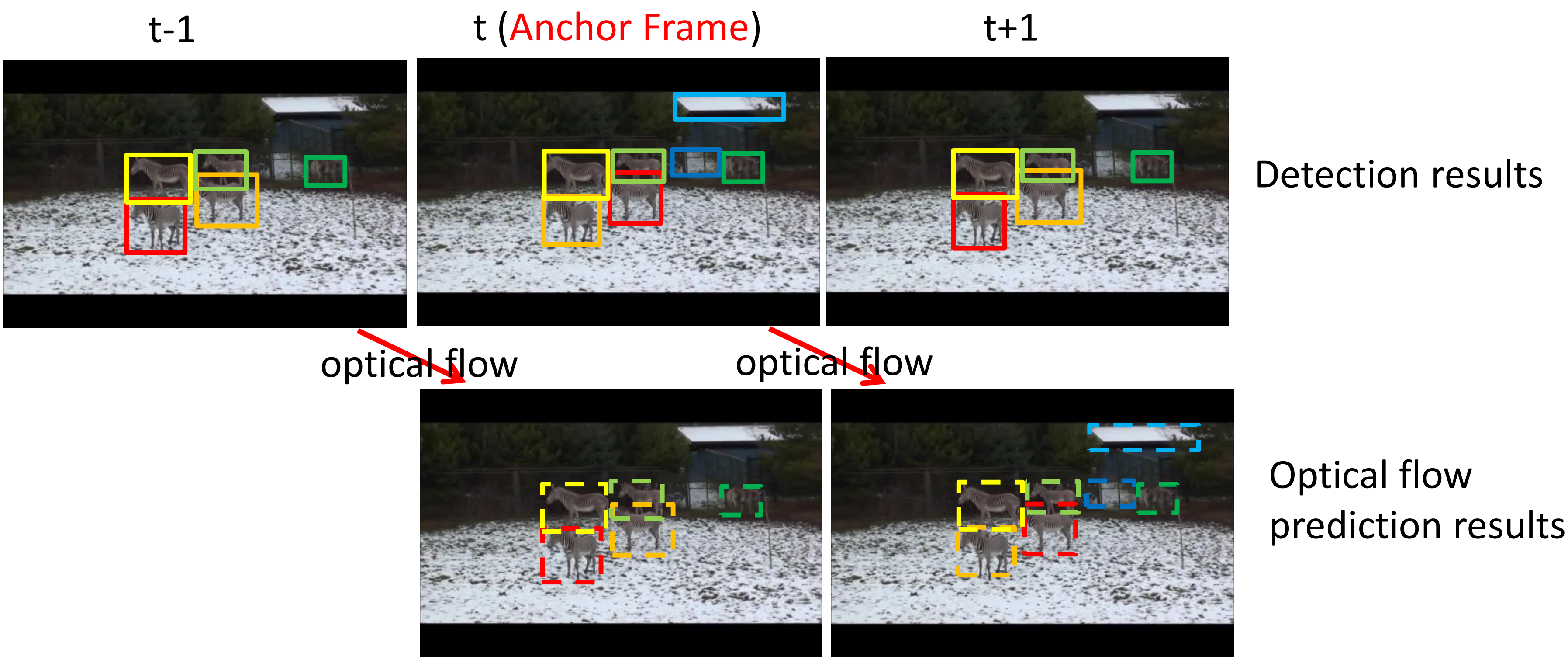
Tracking-based Tubelet Generation

- Anchor Frame Selection
  - Select the frame with highest detection score object



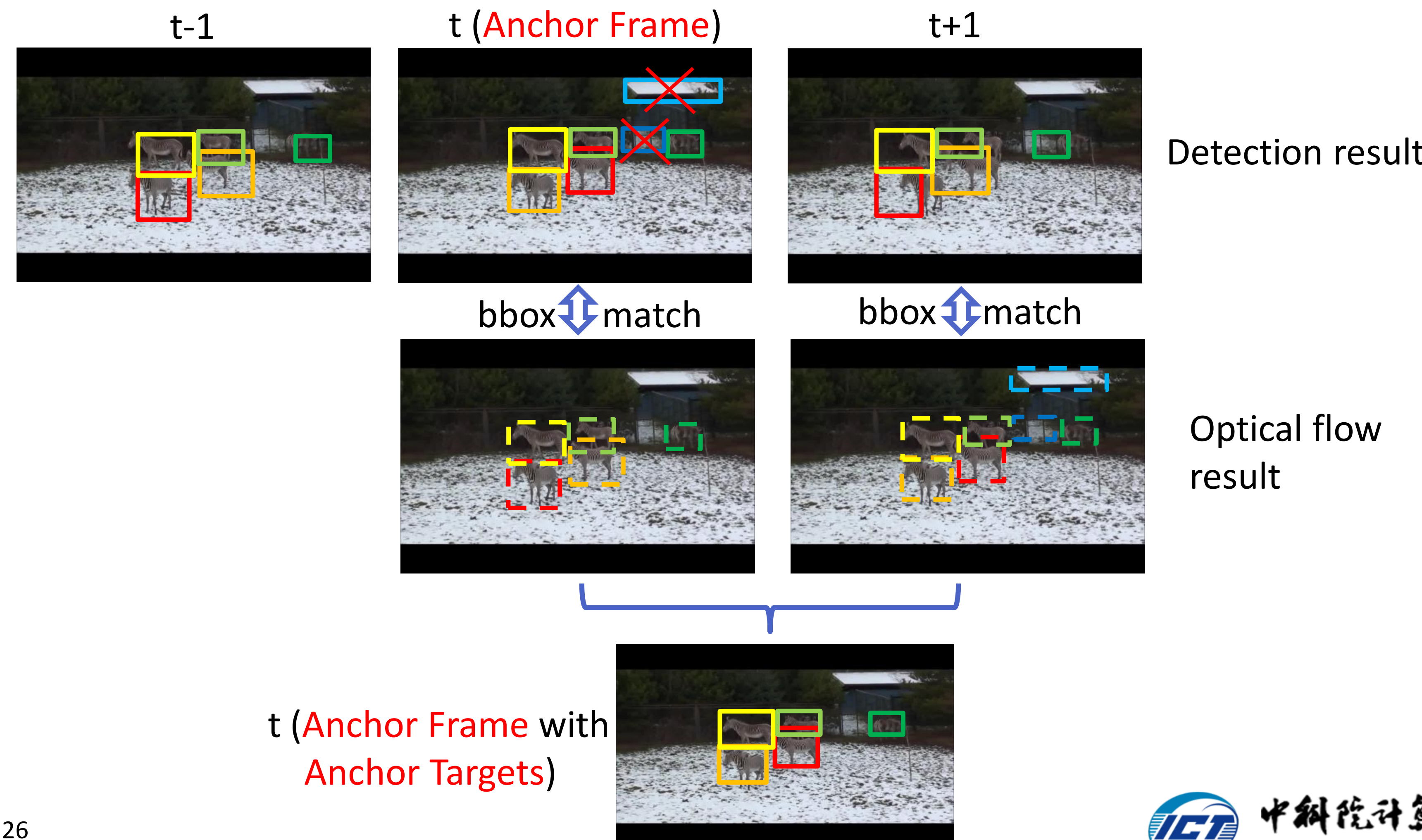
Tracking-based Tubelet Generation

- Anchor Target Selection
  - Exploit the adjacent information with optical flow [5] to determine the reliable anchor targets



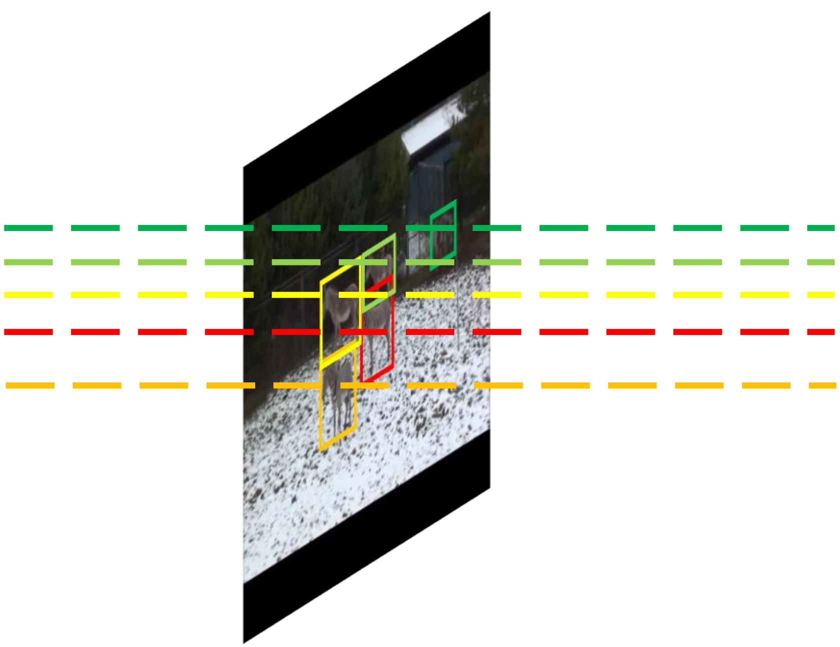
Tracking-based Tubelet Generation

- Anchor Target Selection: remove the unreliable

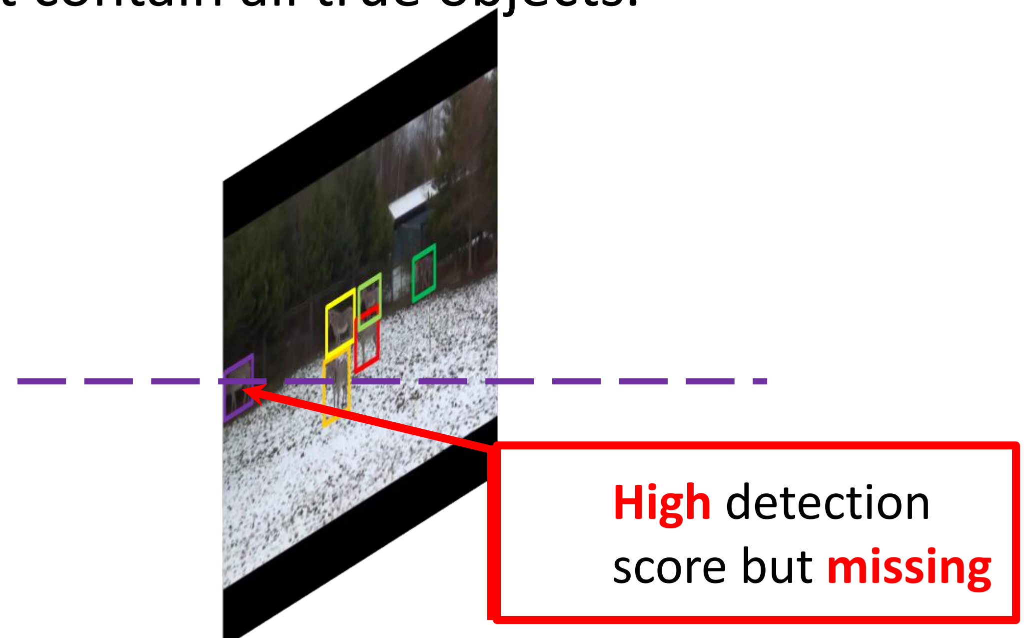


Tracking-based Tubelet Generation

- Multi-target tracking with **detection recall**
  - Allocate each anchor target with a MDNet tracker [6]
  - Track them in parallel
  - Then use detection results to recall missing tubelets since the anchor frame may not contain all true objects.



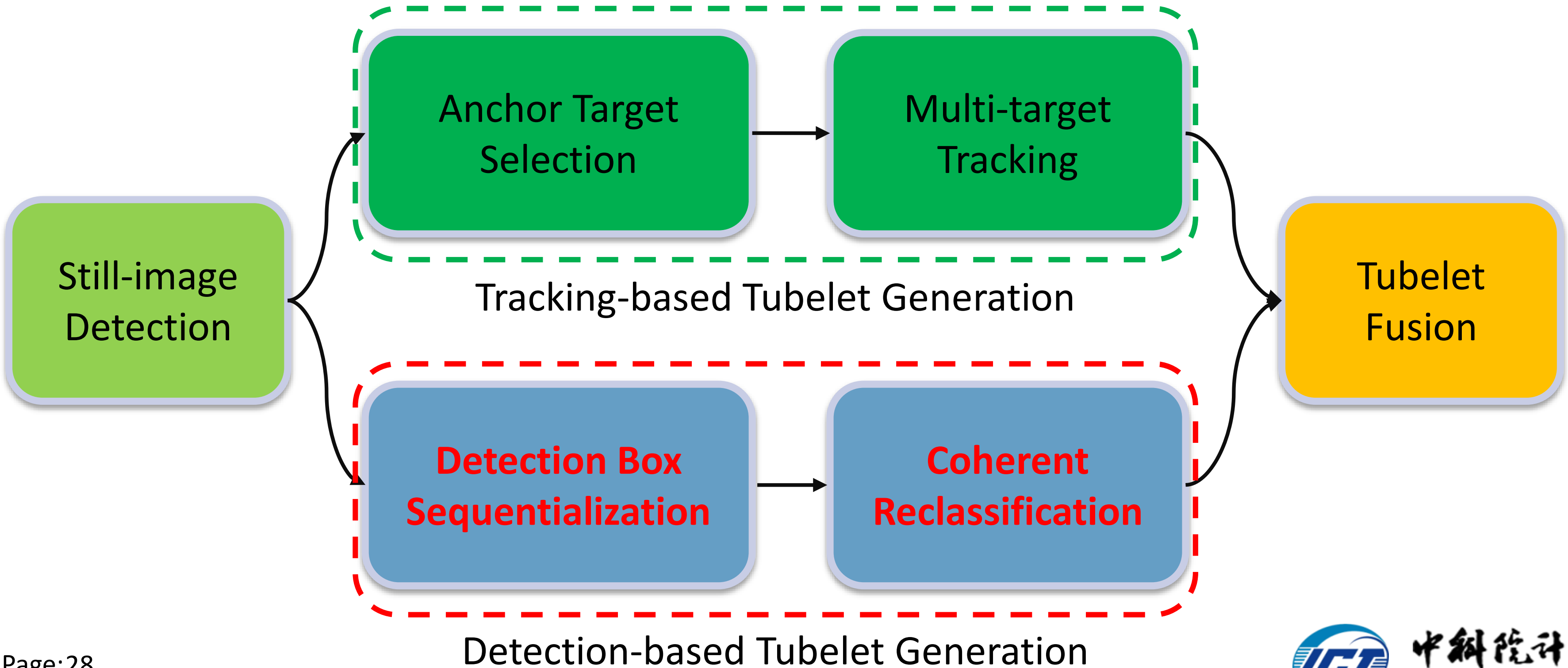
(a) existed tubelets



(b)Recall missing tubelets

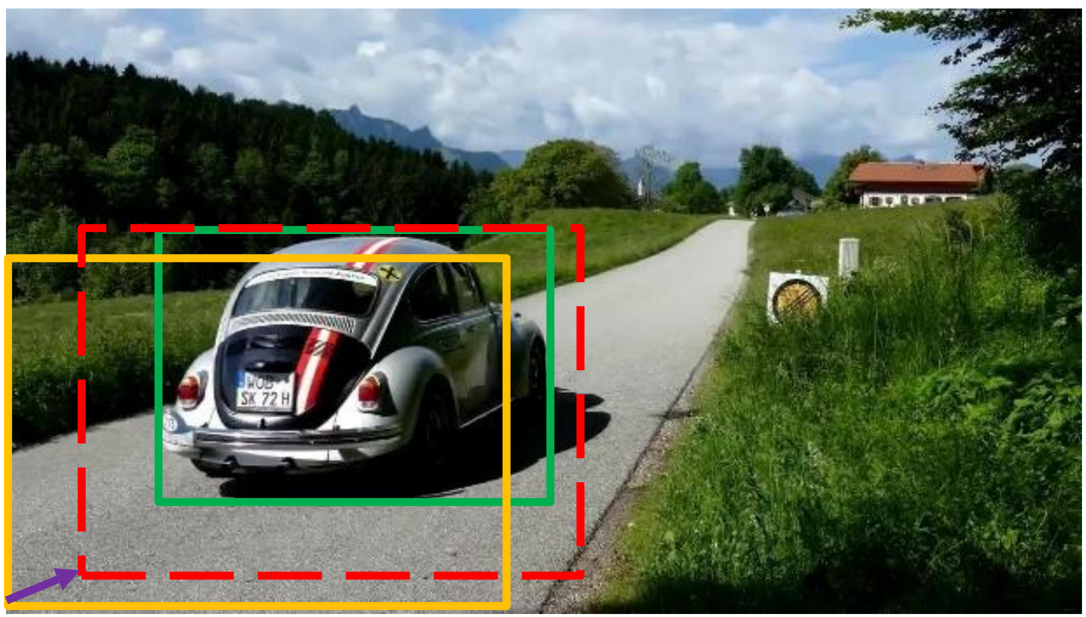
Detection-based Tubelet Generation

- Motivation
  - Overcome **drifting location** problem
  - Excellent object detectors (Faster R-CNN) can generate precise bounding box of high location quality



Detection-based Tubelet Generation

- Adjacent Checking:
  - By optical-flow for precise tubelet



Current Detection Box

Prev Detection Box

Prev Prediction Box

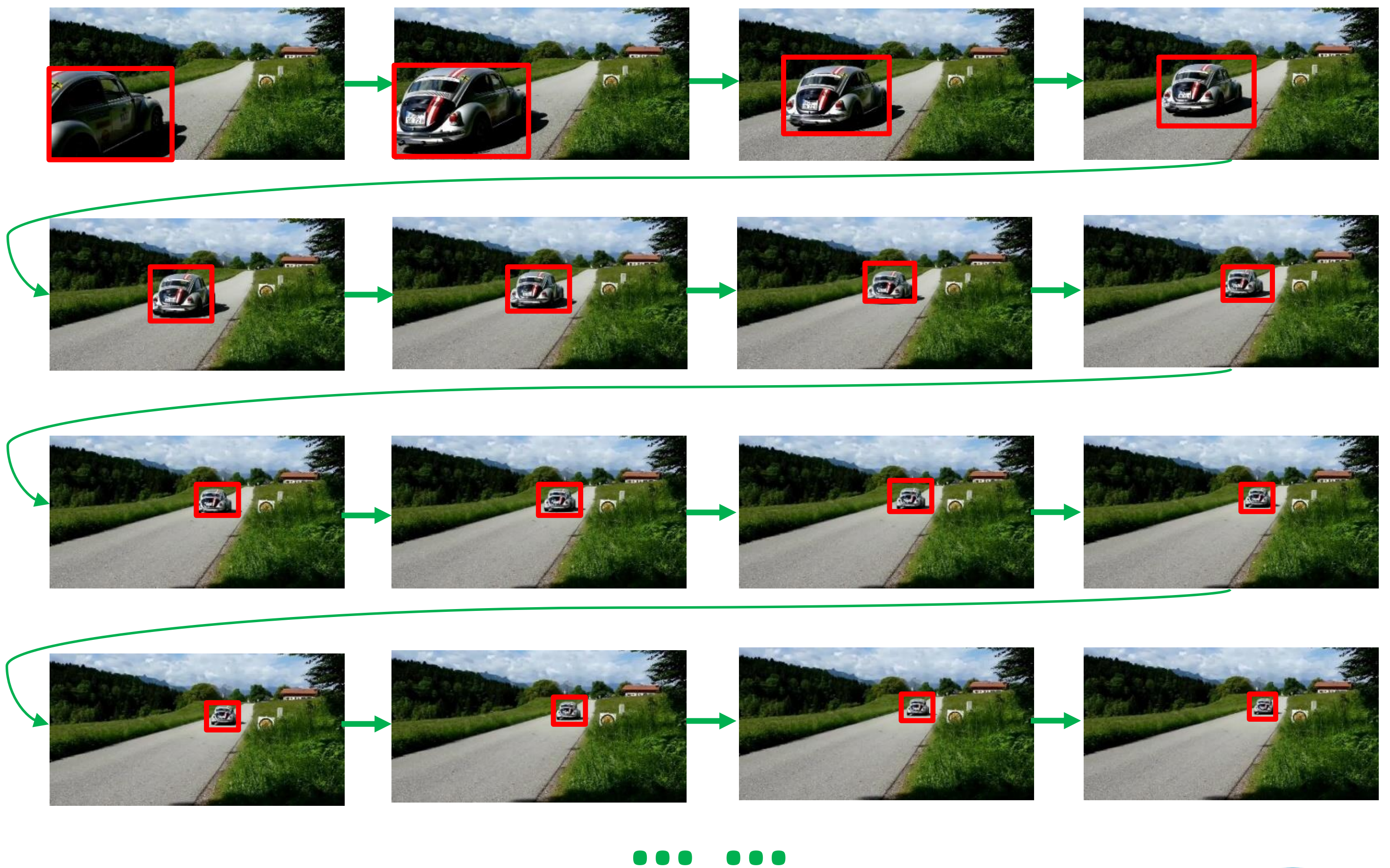
Optical Flow

Current Frame

If IoU (Red, Green) > a given threshold: same tubelet
else: a new tubelet

Detection-based Tubelet Generation

- Detection Box Sequentialization



Detection-based Tubelet Generation

- Coherent reclassification
  - Use majority voting to get coherent categories throughout a given tubelet

	Start Frame							End Frame
Tub	C_p:S_1	C_p:S_2	C_p:S_3	C_p:S_4	C_p:S_5	…	C_p:S_{N-1}	C_q:S_{N-1}	C_p:S_N

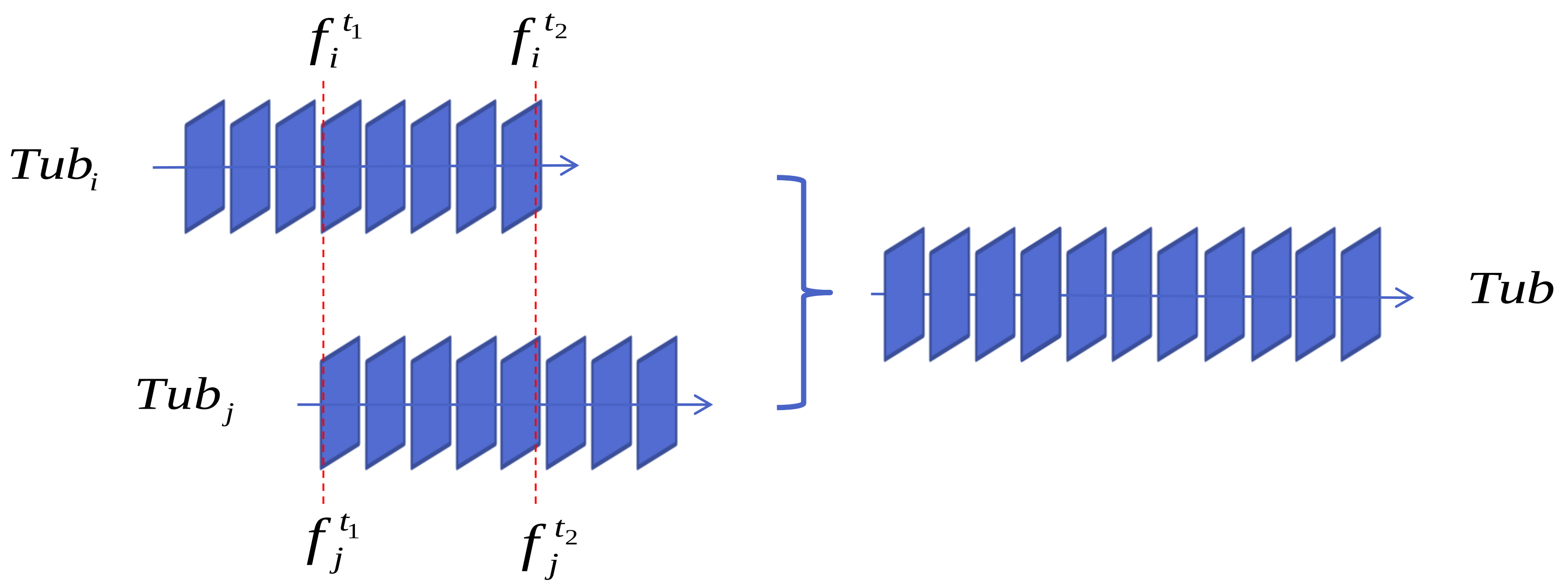
$$Tub_{cls} = \max_k \ \text{count}(C_k)\ (k \in [1,30]) \Longrightarrow$$ Tubelet Category: C_p

$$S_k = \frac{1}{N_k}\sum_{i=1}^{N_k} S_i \Longrightarrow$$ Tubelet Box Score: S

	Start Frame							End Frame
Tub	C_p:S	C_p:S	C_p:S	C_p:S	C_p:S	…	C_p:S	C_p:S	C_p:S

Tubelet Fusion

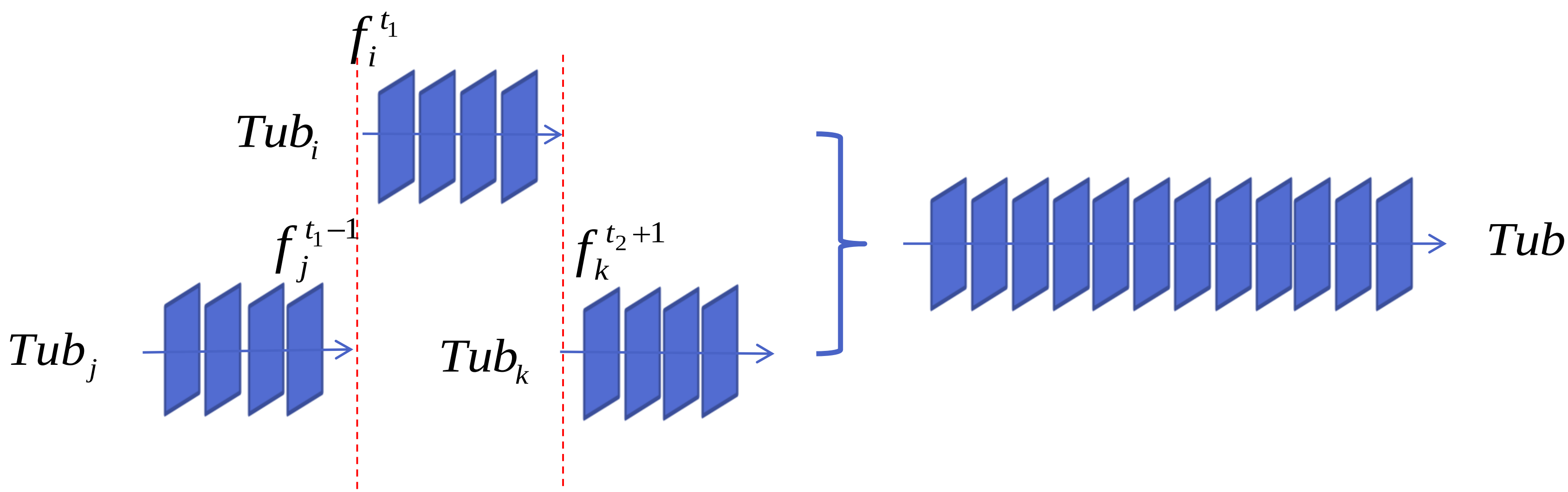
- Union Fusion
  - Merge overlapping tubelets



Union Fusion

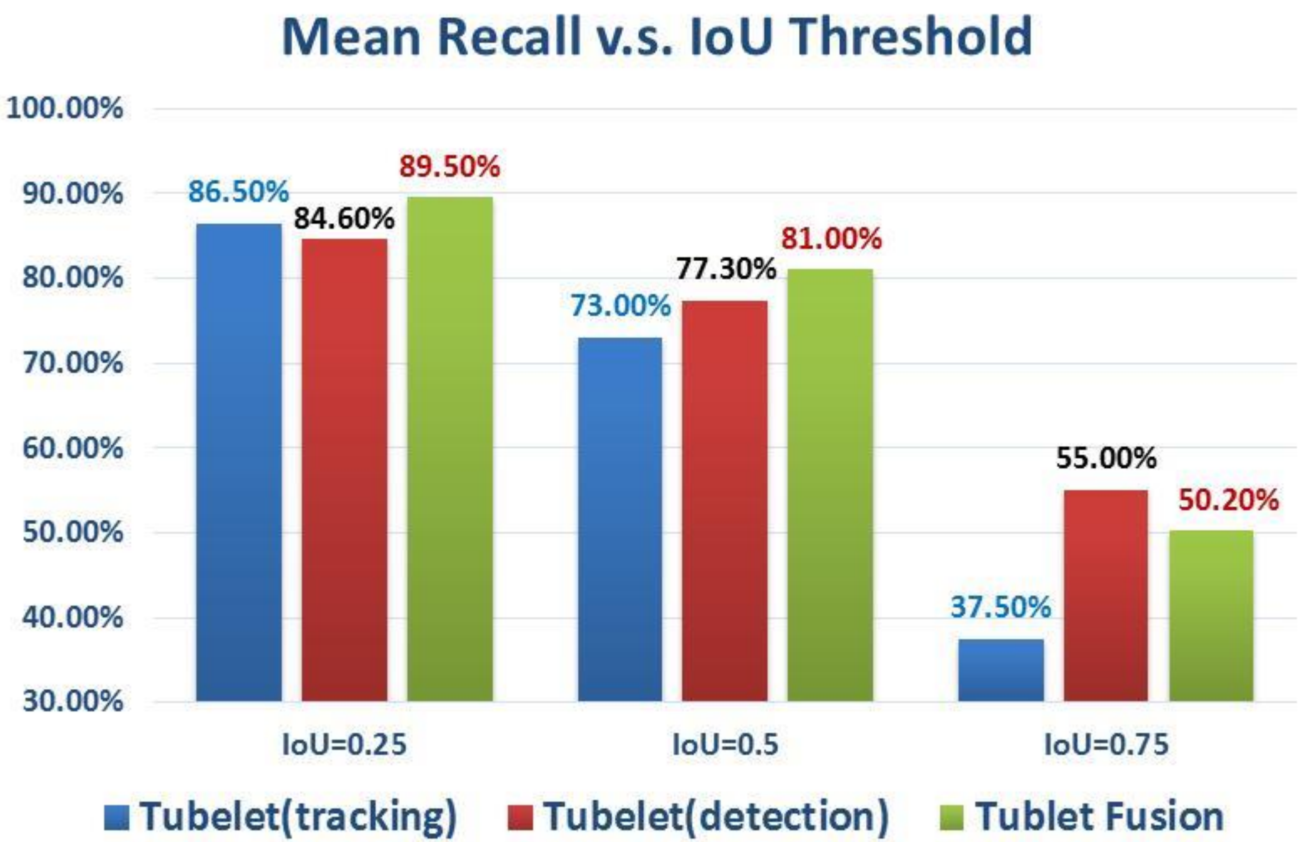
Tubelet Fusion

- Concatenation Fusion
  - Merge successive tubelets



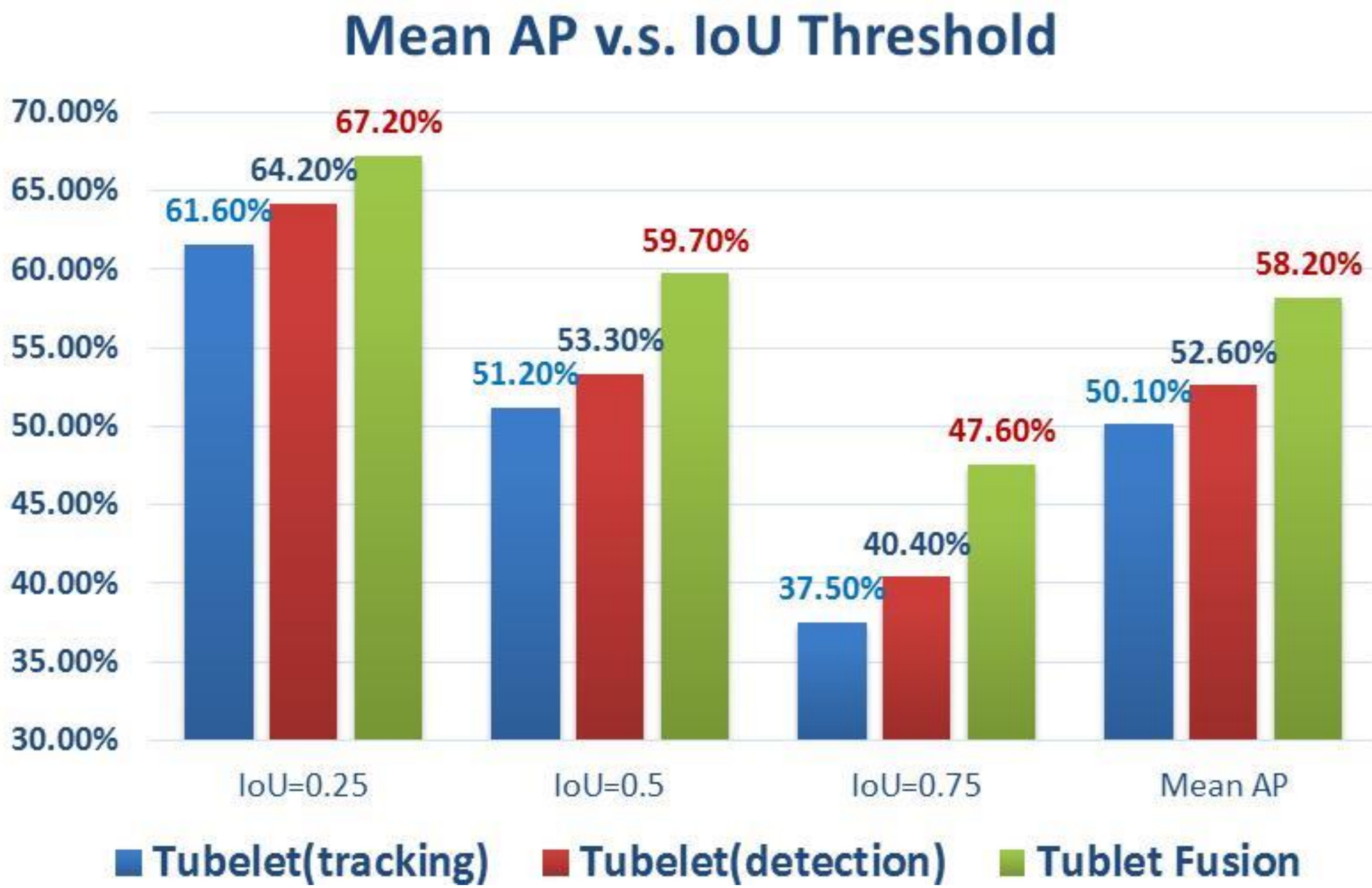
Concatenation Fusion

ILSVRC 2016 VID Val Results



Mean Recall:
Lower IoU: tracking-based is higher
Higher IoU: detection-based is higher

ILSVRC 2016 VID Val Results



Mean AP: Detection-based is higher than Tracking-based, Fusion is best!

References

- [1] Ren S, He K, Girshick R, Sun J. “Faster R-CNN: Towards real-time object detection with region proposal networks”, NIPS 2015: 91-99.
- [2] Gidaris, Spyros, and Nikos Komodakis. "Object detection via a multi-region and semantic segmentation-aware cnn model." ICCV 2015.
- [3] Yu, Fisher, and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions." ICLR 2016.
- [4] He K, Zhang X, Ren S, Sun J. “Deep residual learning for image recognition”, CVPR 2016.
- [5] Kang K, Ouyang W, Li H, Wang X. “Object Detection from Video Tubelets with Convolutional Neural Networks”, CVPR 2016.
- [6] Nam H, Han B. “Learning multi-domain convolutional neural networks for visual tracking”, CVPR 2016.

Welcome:

Questions and Comments

Thank You!