Multi-task Self-Supervised Object Detection via Recycling of Bounding Box Annotations

2019-07-08

Wonhee Lee (wonhee@vision.snu.ac.kr)

SEOUL NATIONAL UNIV.
VISION & LEARNING

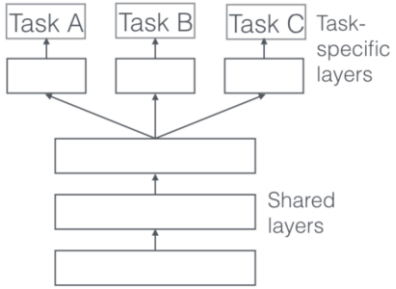
① Multi-task ② Self-Supervised Object Detection via ③ Recycling of Bounding Box Annotations

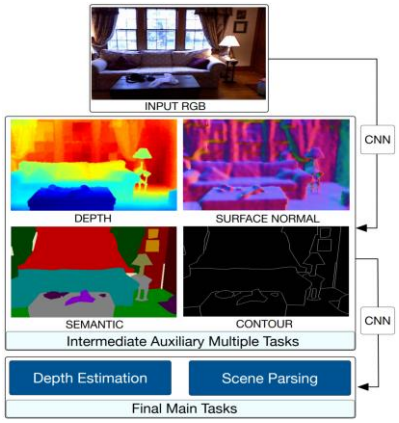
Objectives

- Improving the object detection accuracy
- The two requisites for object detection
  - Novel detector & backbone network
  - Lots of bounding box annotations
- Additional ways
  - Multi-task learning
  - Self-supervised learning
  - Annotation reusing

Introduction - MTL

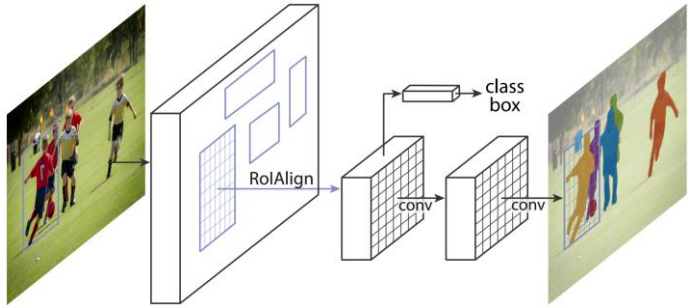
- Multi-task learning (MTL)
  - Cooperative feature learning by several tasks
  - Synergetic when tasks are related to each other





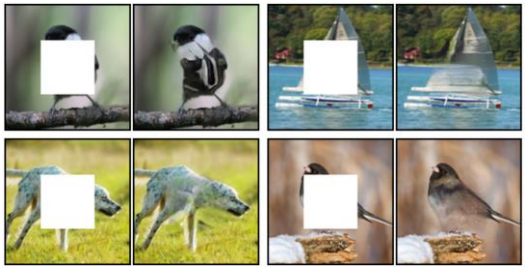
Black: Yes, Necktie: Yes, Gender: Male, Strips: No, Yellow: No, White: No, Skin Exposure: No ...

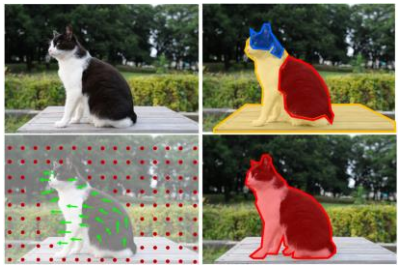
Black: Yes, Necktie: No, Gender: Female, Strips: Yes, Yellow: No, White: Yes, Skin Exposure: Yes ...



Introduction - SSL

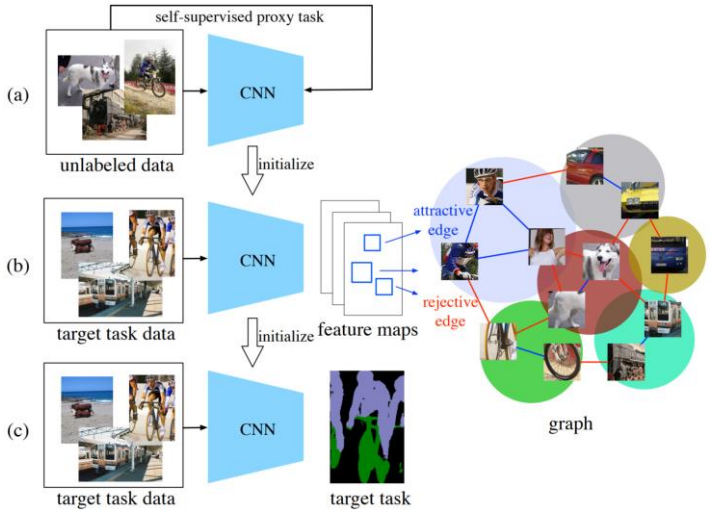
- Self-supervised learning (SSL)
  - annotated without any human effort
- Mostly for pretraining
  - performance: random init. < self-supervision pretraining < ImageNet pretraining



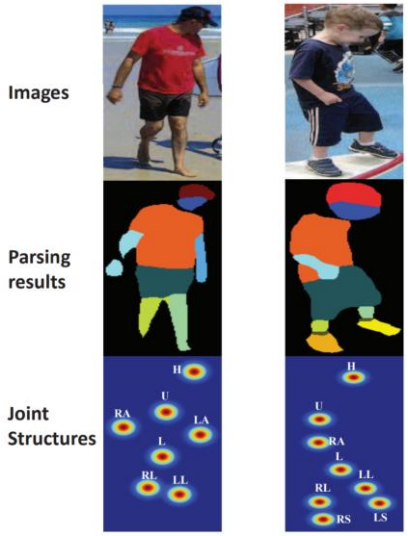


Introduction - Annotation Recycling

- Why Recycling?
  - used to create detection loss
  - used again to create losses of auxiliary tasks
  - extract information in the label to the extreme
- To boost up the main task



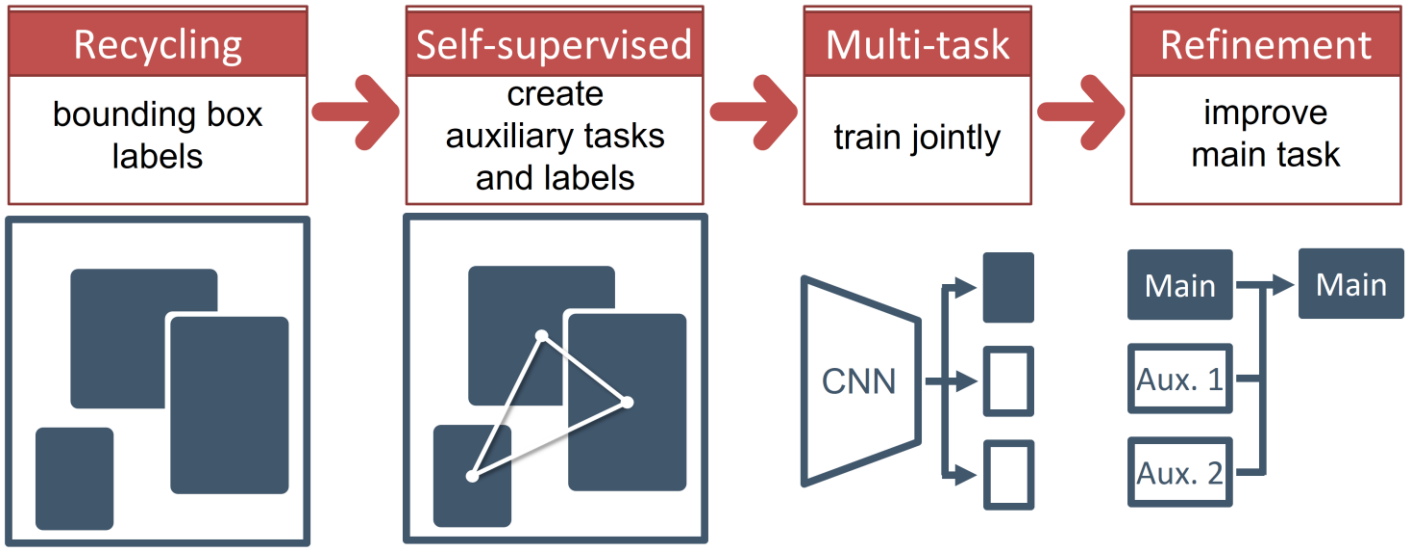
Mix & Match task by recycling segmentation label



Joint structure prediction by recycling segmentation label

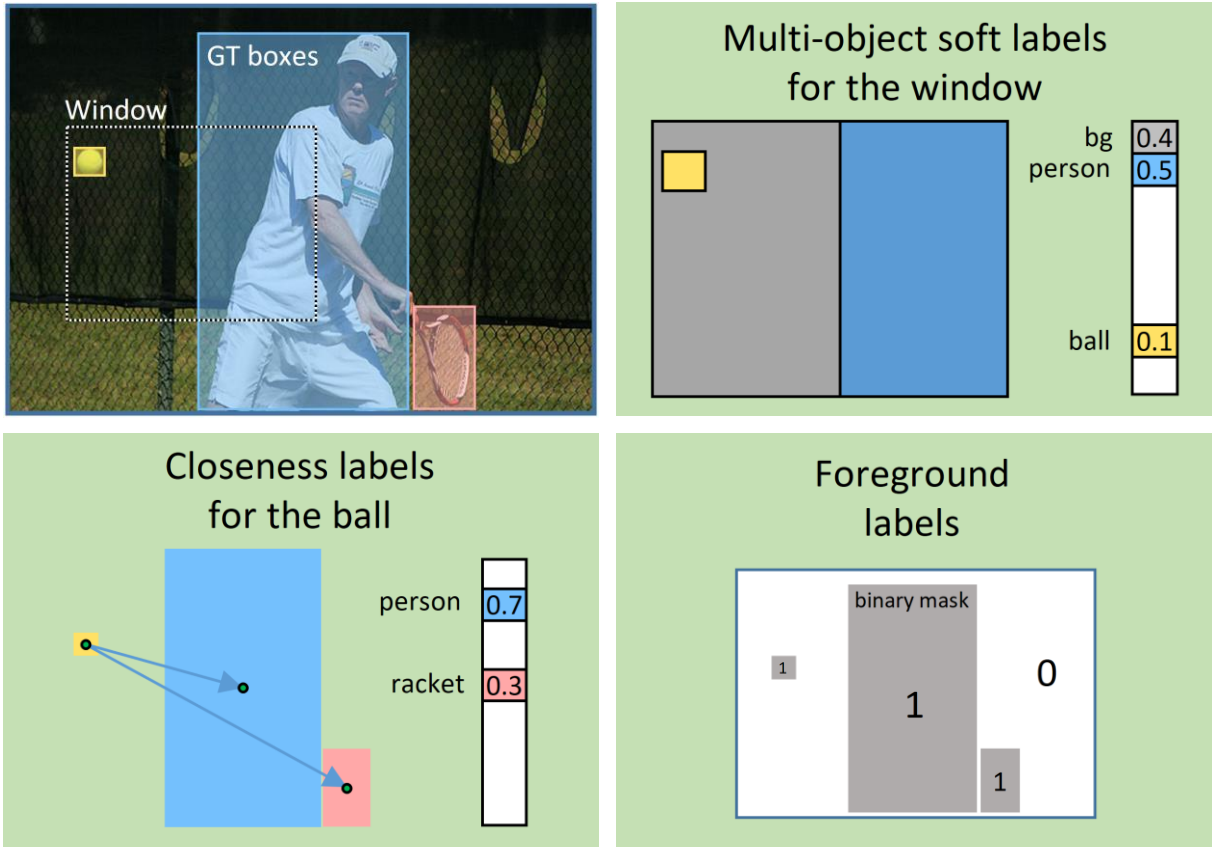
Our Idea

- Start from the conventional detection setting with BBox annotations
- Extract additional information in the annotations (Recycling)
- Create three auxiliary tasks that predict it (SSL)
- Cooperative feature learning (MTL)
- Refine the classification result using the output of auxiliary tasks



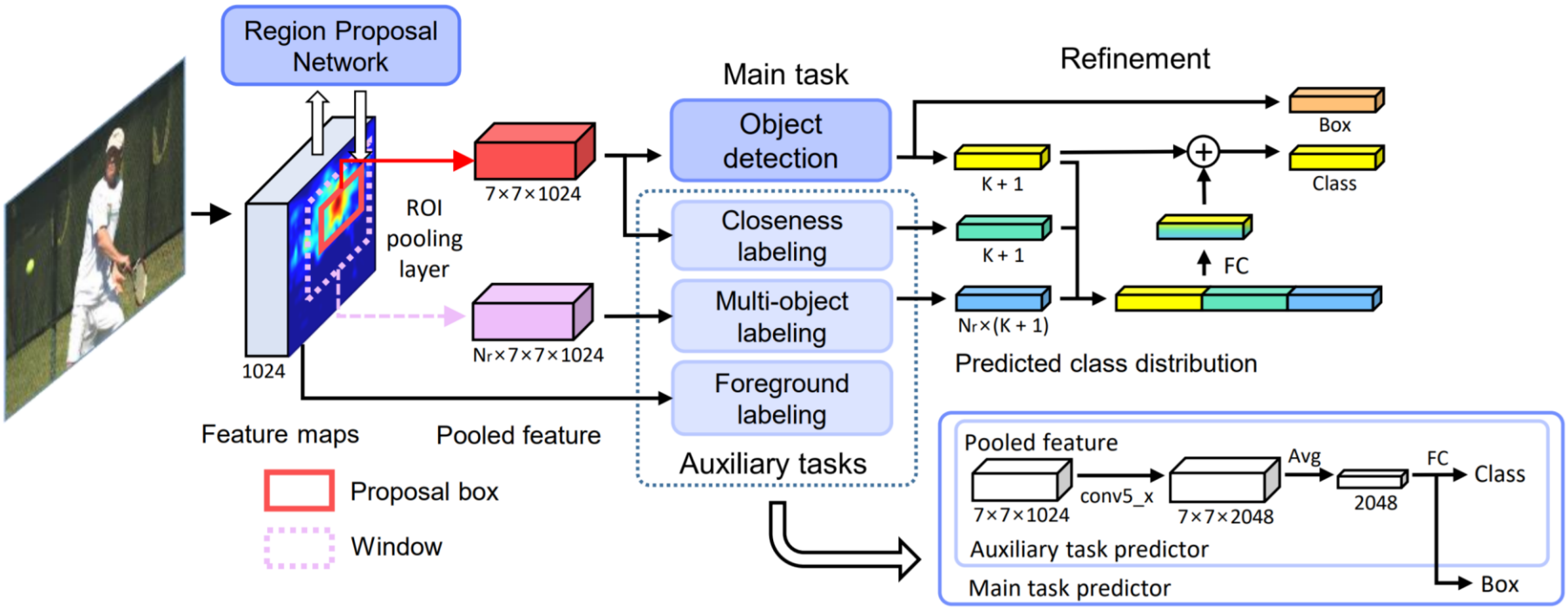
Approach - 3 auxiliary tasks

- Multi-object labeling
- Closeness labeling
- Foreground labeling



Approach - architecture

- How object detection improves?
  - Multi-task learning: cooperative feature learning
  - Refinement: using contextual information in the auxiliary task output



Experiments

- 2.0 mAP ↑ on average

Dataset	VOC			COCO	
Training	07	07+12		17 *train*	
Test	07	07	12	17 *val*	17 *test-dev*
Baseline	77.0	81.7	75.3	32.7	32.8
+ Task1	**78.9**	**83.8**	77.4	34.1	34.2
+ Task2	77.3	83.0	76.0	33.3	33.5
+ Task3	77.0	82.0	75.1	32.9	32.8
+ Task1,2	78.5	83.7	77.3	34.5	34.6
+ Task1,2,3	78.7	83.7	**77.5**	**34.6**	**34.7**

Dataset

Training	07	07+12	
Test	07	07	12
Baseline	73.4	78.6	72.1
+ Task1	74.3	80.1	74.0
+ Task2	73.5	78.7	72.2
+ Task3	73.3	78.4	71.9
+ Task1,2	**75.0**	80.4	**74.2**
+ Task1,2,3	74.7	**80.6**	73.9

Base Detector (R-FCN)

Backbone	MobileNet [23]			Inception-ResNet-v2 [44]		
Training	07	07+12		07	07+12	
Test	07	07	12	07	07	12
Baseline	61.2	68.6	62.0	80.7	84.3	78.2
+ Task1	63.4	**71.3**	**64.5**	81.7	85.9	**80.5**
+ Task2	62.5	69.3	62.6	81.0	84.8	79.0
+ Task3	61.3	68.8	61.7	80.6	84.2	78.3
+ Task1,2	**63.9**	70.9	**64.5**	**81.8**	**86.1**	80.1
+ Task1,2,3	63.8	70.8	64.4	**81.8**	86.0	80.0

Backbone Networks

Training	07	07+12	
Test	07	07	12
Baseline	77.0	81.7	75.3
+ MTL	78.0 (+1.0)	83.0 (+1.3)	76.7 (+1.4)
+ Refinement	78.3 (+1.3)	82.7 (+1.0)	76.4 (+1.1)
+ Both	**78.7 (+1.7)**	**83.7 (+2.0)**	**77.5 (+2.2)**

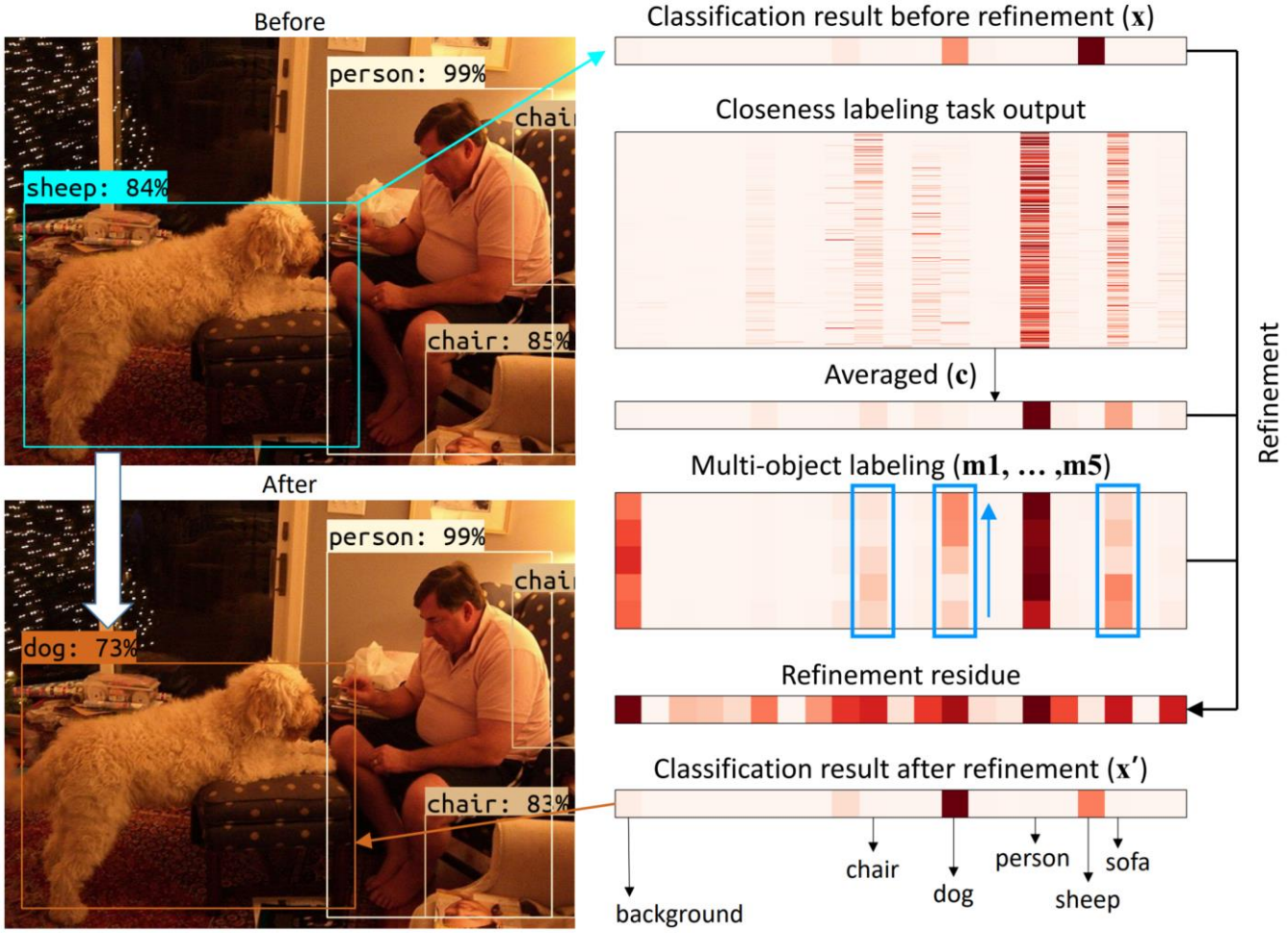
Ablations

Contributions

- A first attempt to recycle BBox annotations for object detection
- Orthogonal to any proposal-based detection models
- Improvement (mAP 2.0↑ on avg) in multiple architectures

Visualization - Refinement

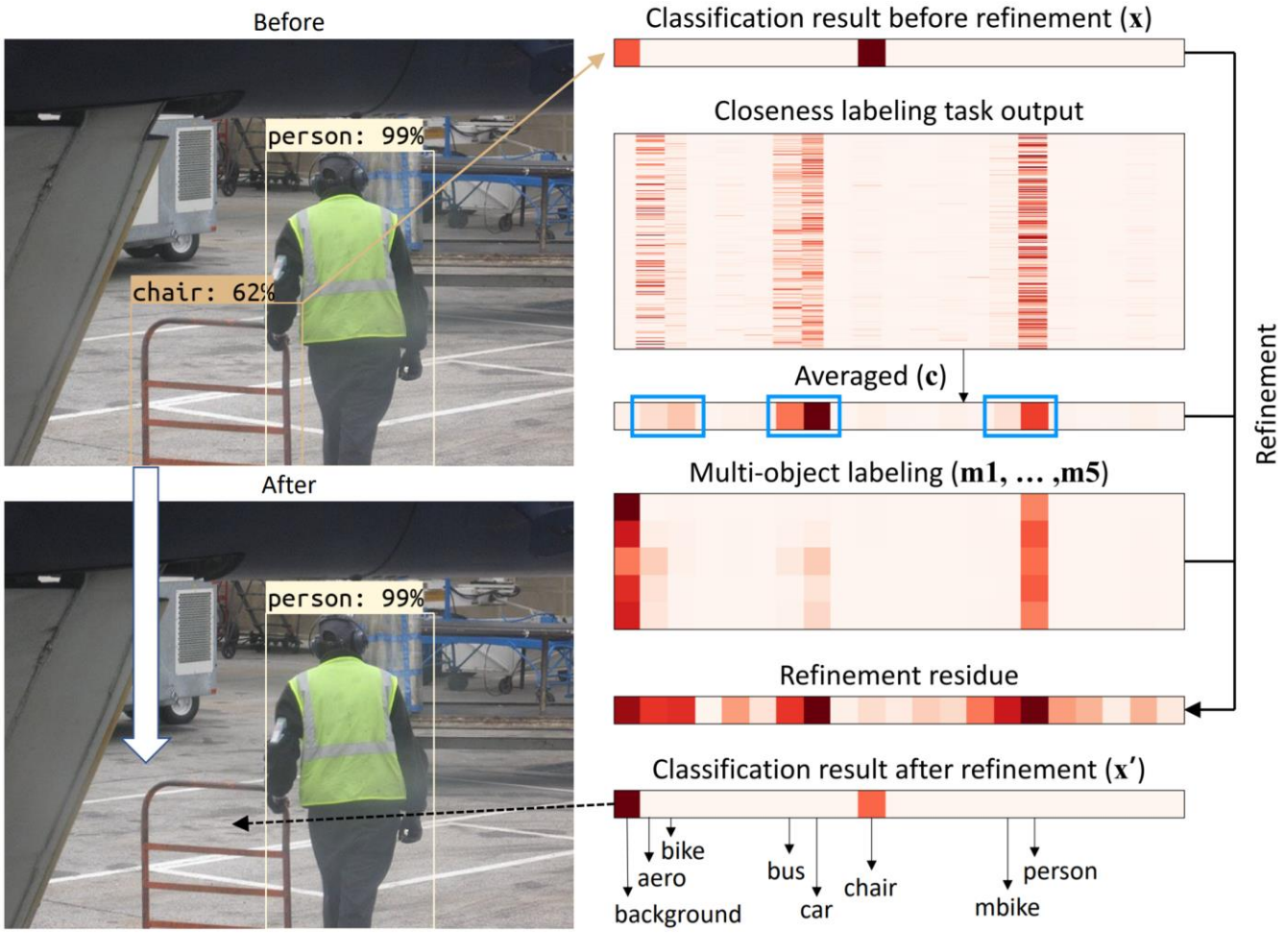
- By multi-object labeling task



(a) Classification result changed from "sheep" to "dog"

Visualization - Refinement

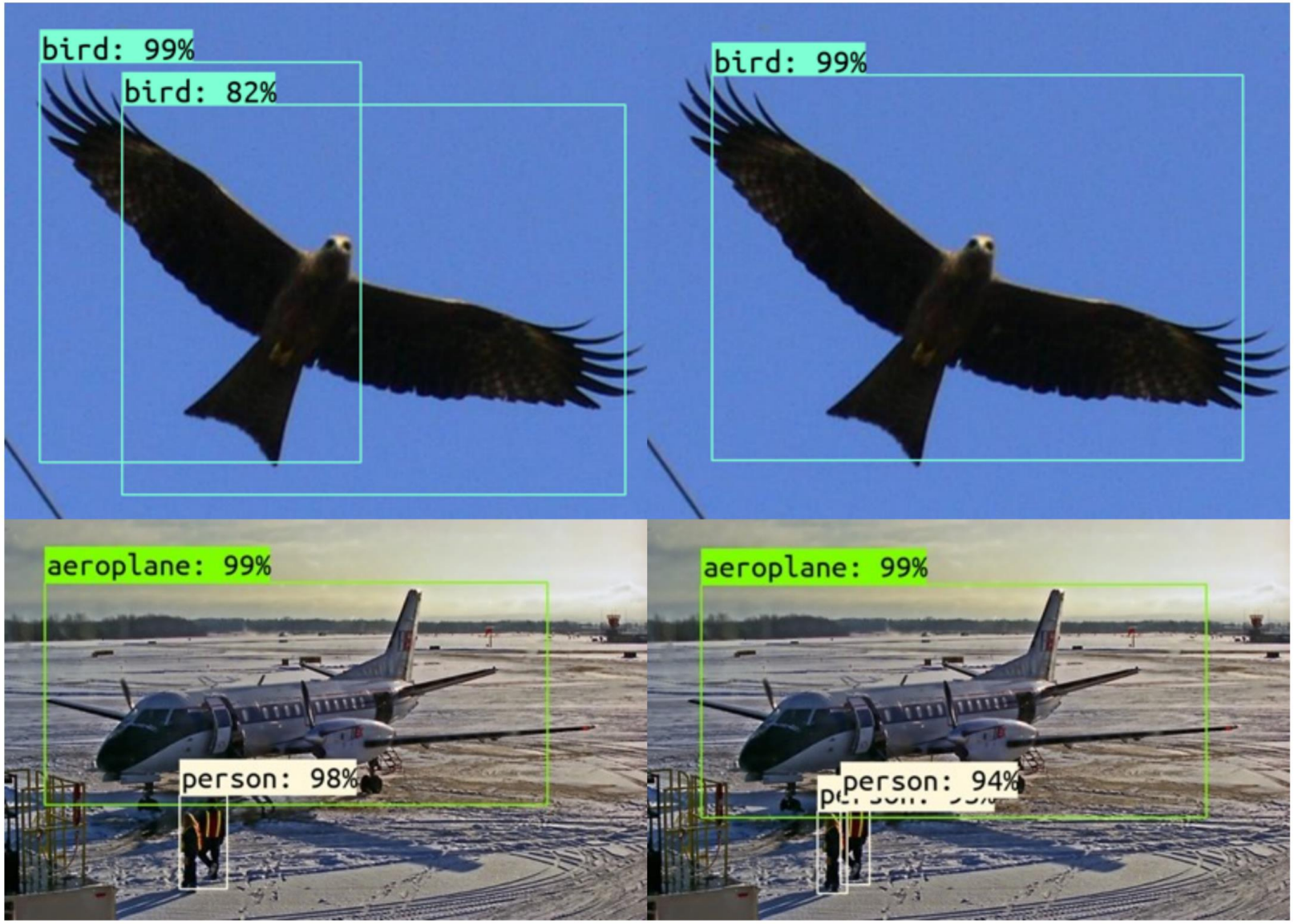
- By closeness labeling task



(b) Classification result changed from “chair” to “background”

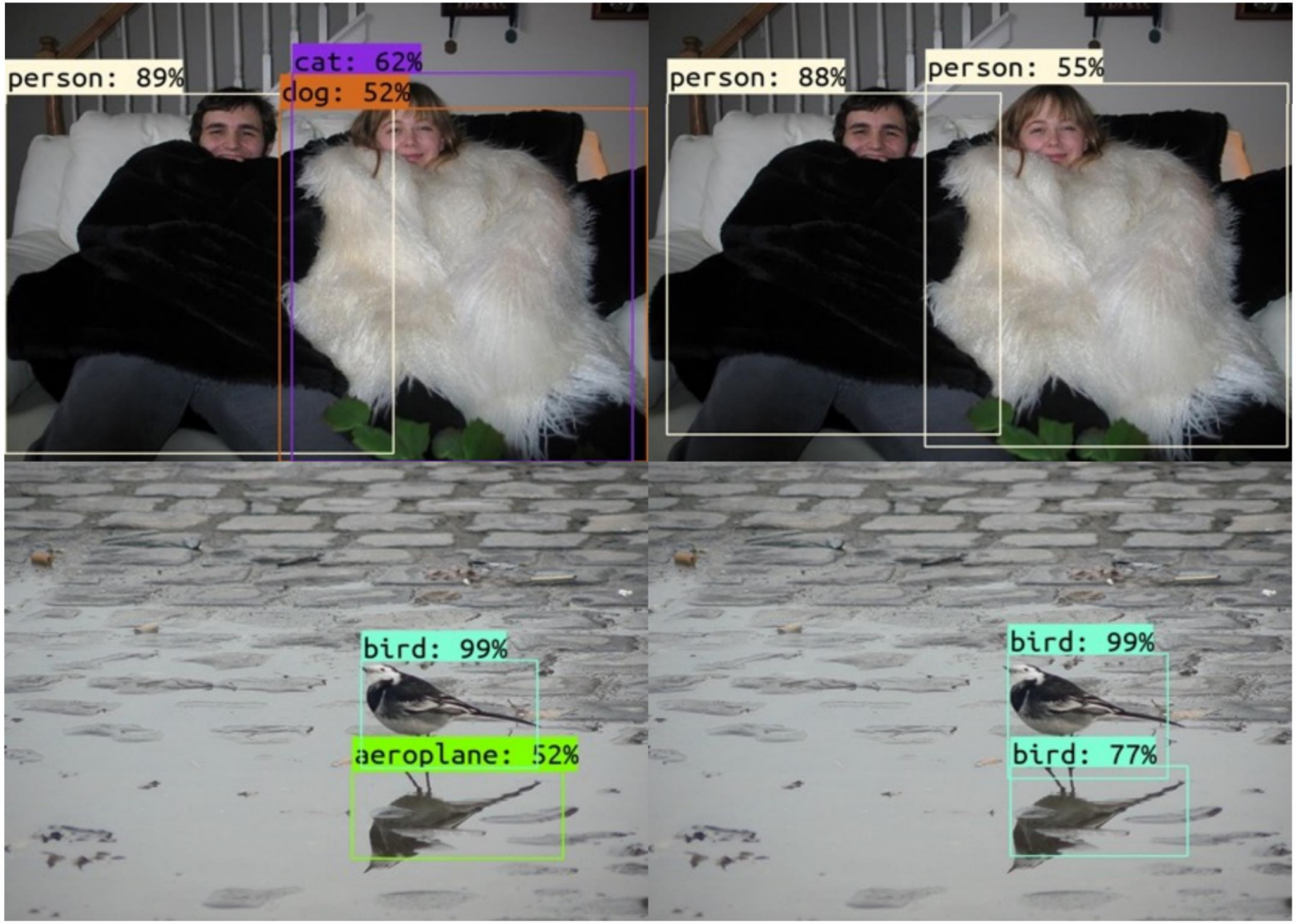
Visualization

Localization



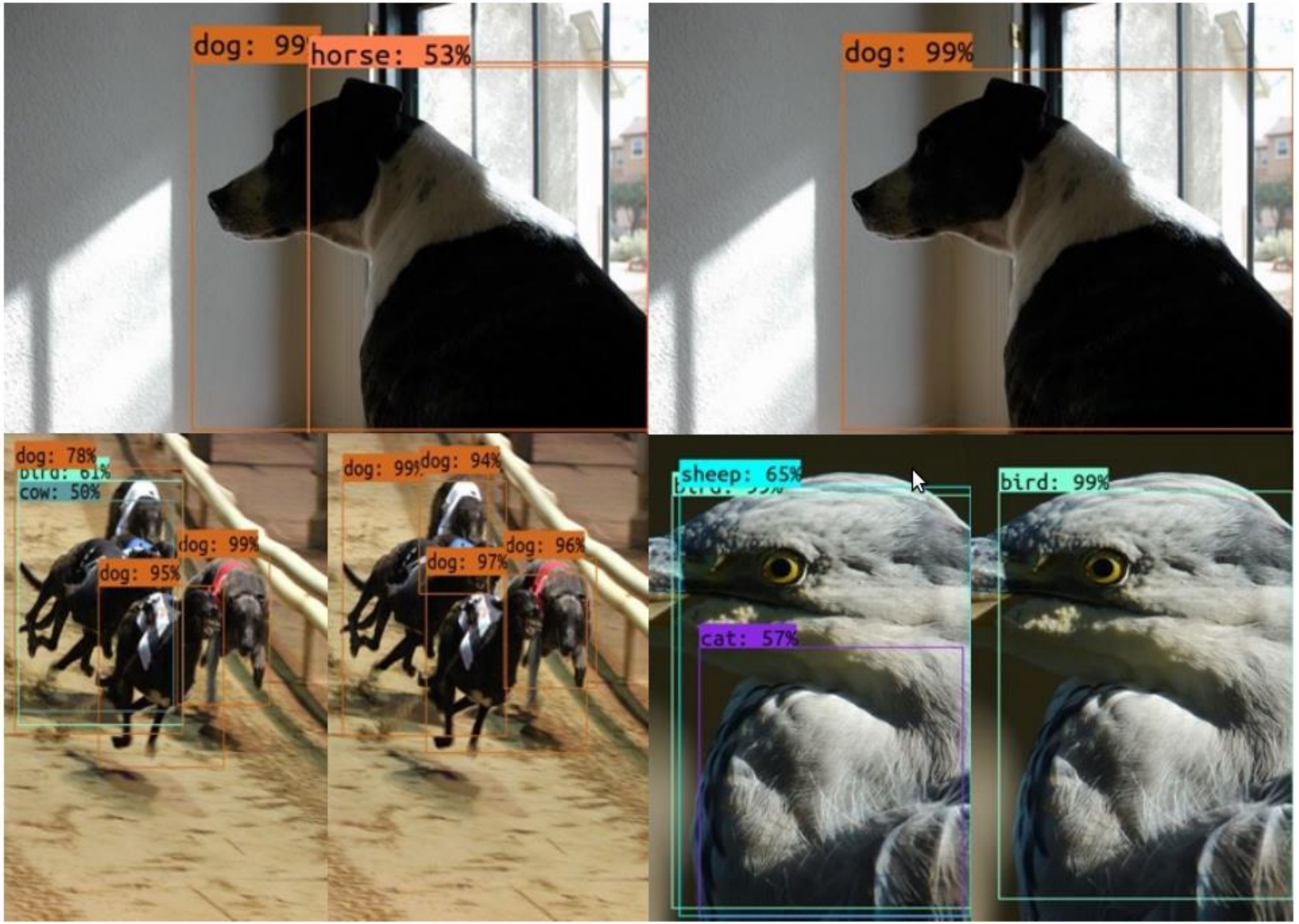
Visualization

Classification



Visualization

Redundancy

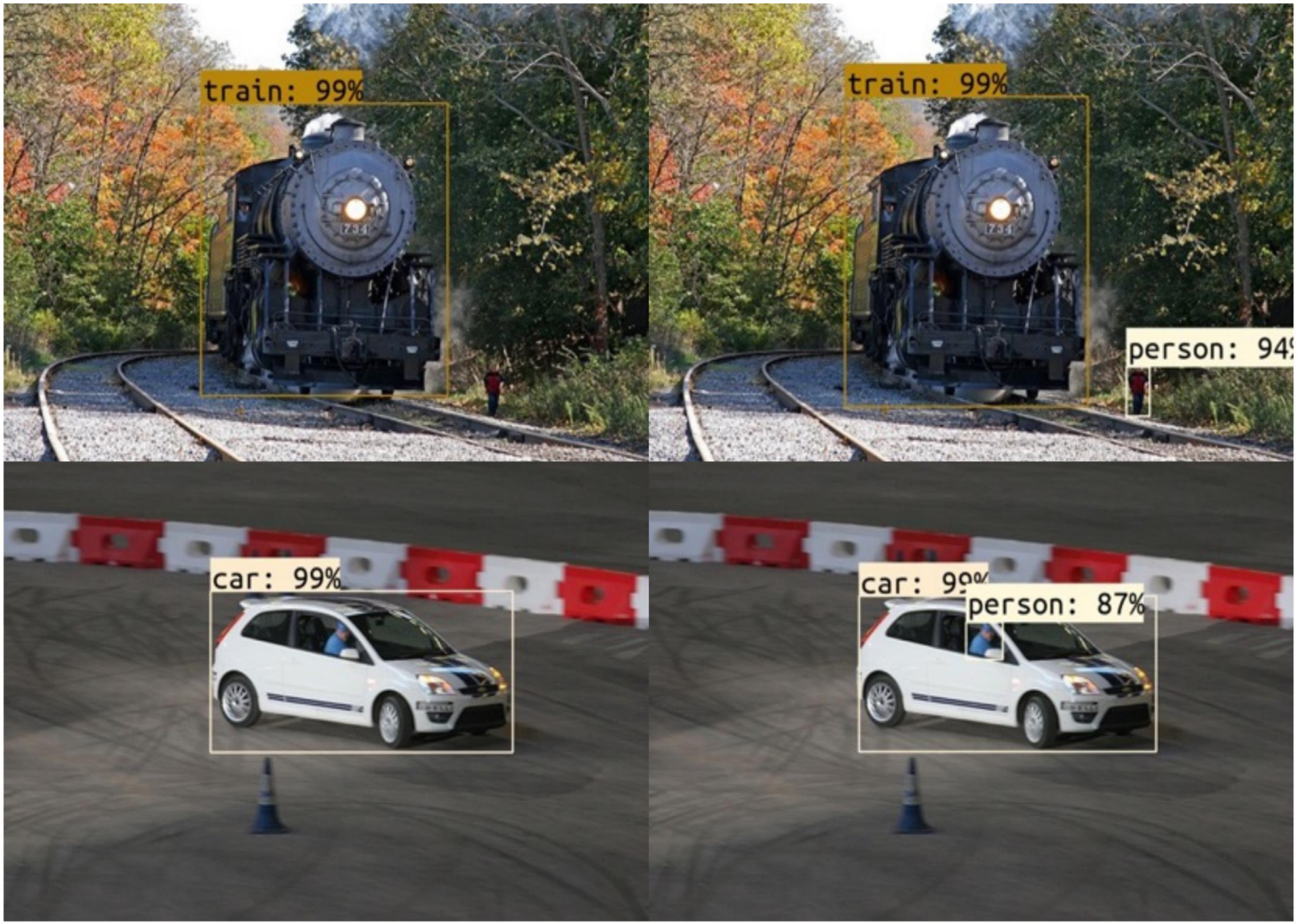


Visualization

Background

Visualization

False Negative



Q & A

Thank You!