Learning Lightweight Pedestrian Detector with Hierarchical Knowledge Distillation

Rui Chen, Haizhou Ai, Chong Shang, Zijie Zhuang, Long Chen

Department of Computer Science and Technology, Tsinghua University

Email: chenr18@mails.tsinghua.edu.cn

2019/09/24

Outline

- Brief Recap:
  - Our Task: Accelerating Pedestrian Detection
  - Our Technique: Knowledge Distillation

- Our Proposed Method:
  - **Hierarchical** Knowledge Distillation for Pedestrian Detection

Brief Recap:

- Pedestrian Detection

- Knowledge Distillation

Recap: Pedestrian Detection

Pedestrian detection is mainly about **localizing** and **classifying** all the pedestrians in the still image or the video sequence.

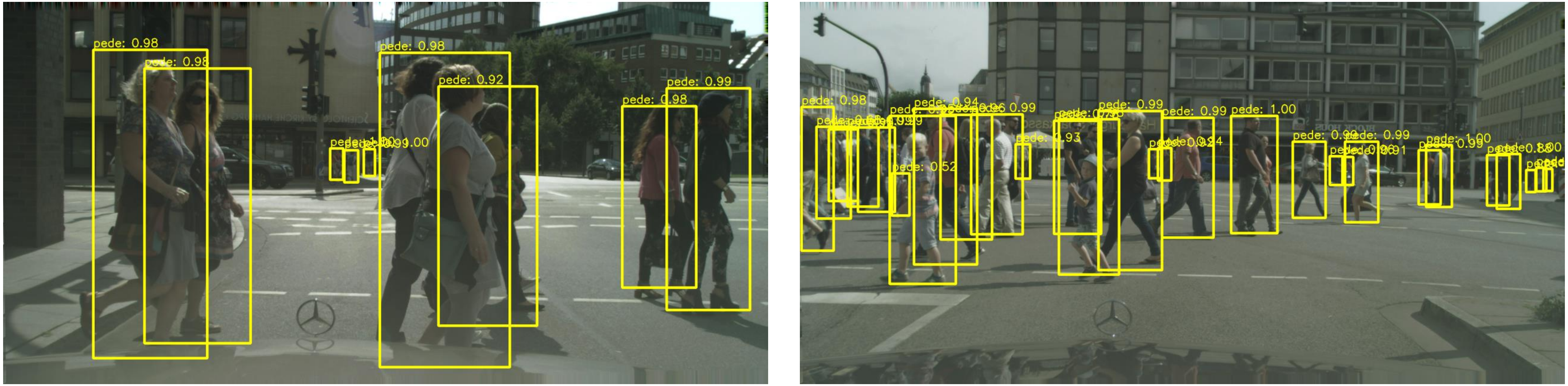


Fig. **1**. Detection results using our lightweight detector on the *Citypersons* dataset.

Recap: Knowledge Distillation

Knowledge distillation aims to improve a **lightweight** model’s performance by **learning from** a well-trained but **cumbersome** model

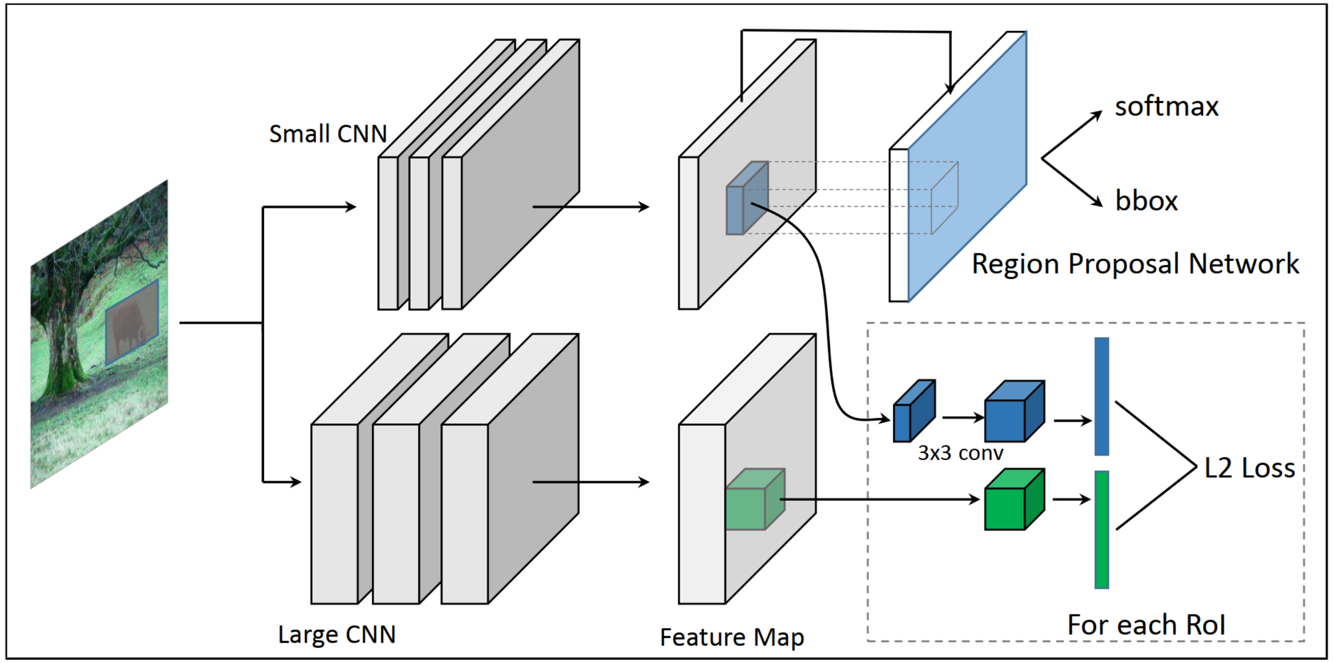


Fig. **2**. An example of knowledge distillation in object detection, mimicking feature by proposal sampling.

Problems we need to tackle:

1. Current CNN-based pedestrian detectors with heavy backbones (*VGG-16*, *ResNet50*) indeed have relatively good performances on this task.

- But their **computational burdens** lead to deployment problem in real-world applications.

2. Current model compression techniques or directly training a lightweight model from scratch.

- May not be easy to **implement** (quantization), or cannot satisfy the need for detection **accuracy**.

Motivation:

1. Knowledge distillation is a simple, yet effective technique for improving detection speed with minor performance degradation.

2. Multiple intermediate supervision is a direct way for better transferring knowledge from a teacher model to a student model.

3. Multi-level feature distillation should have a more precise spatial response than the single-level feature distillation.

Our Proposed Method:

- Overall Architecture
- Hierarchical Distillation
- Experiment Result

The Pipeline of Our Framework

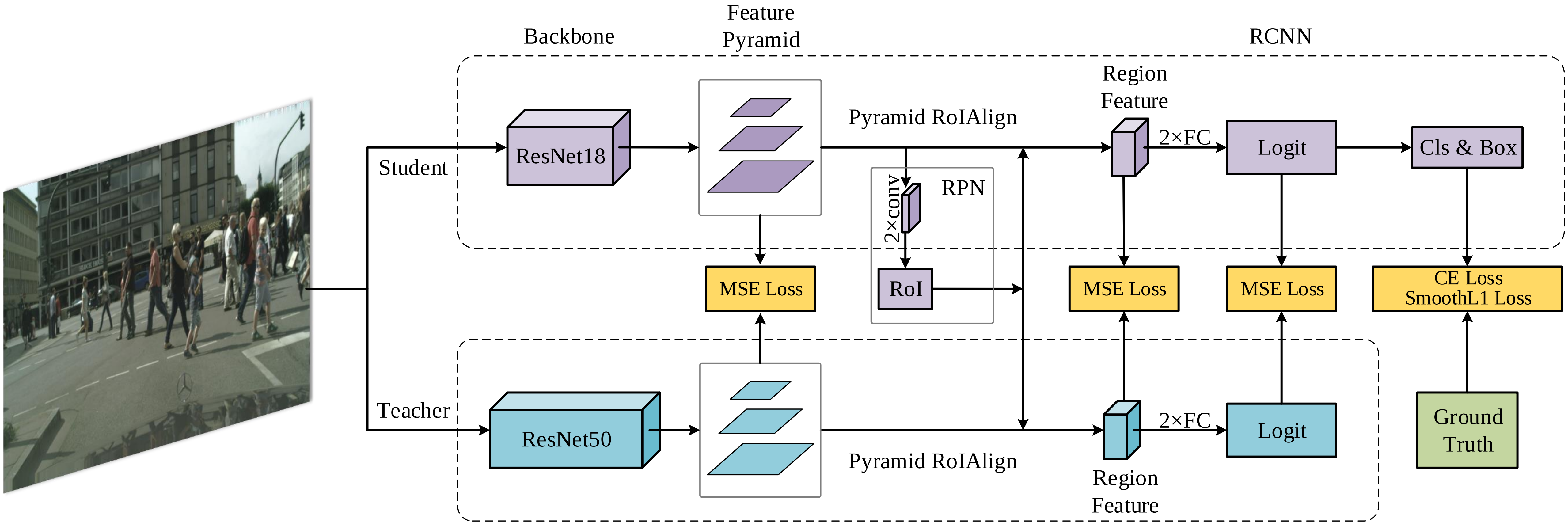
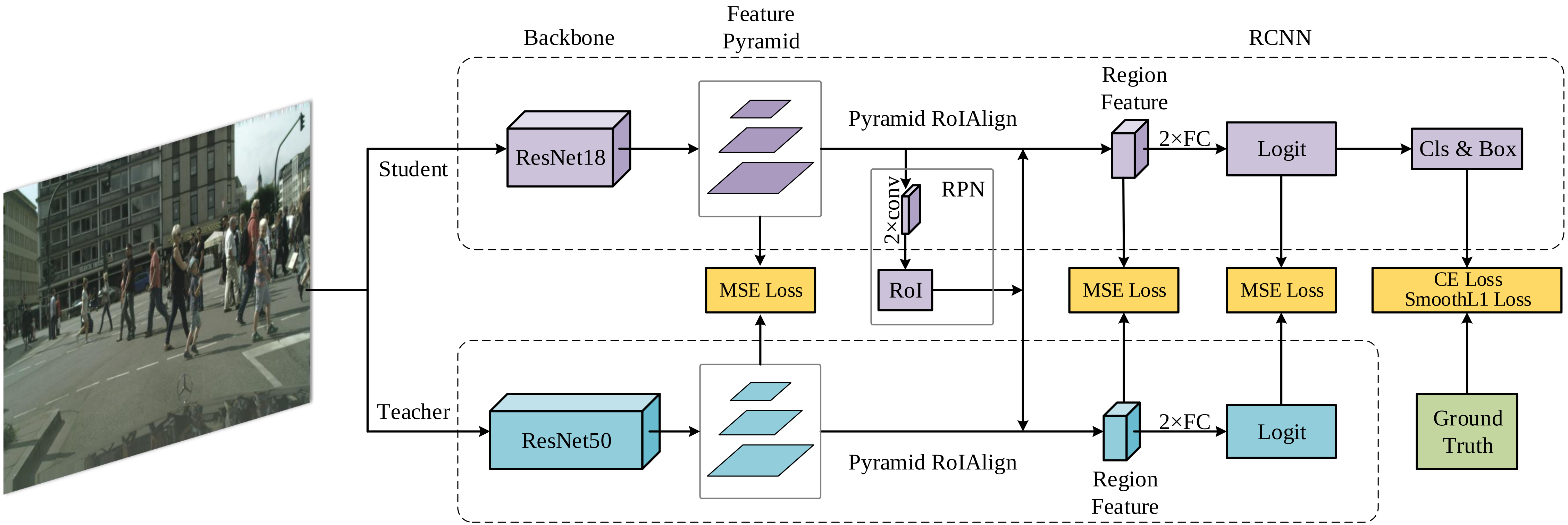


Fig. **3**. The proposed framework. The stronger ResNet50 detector teaches the ResNet18 to obtain a better representation. Multiple distillations, i.e. MSE Loss, are performed in the pipeline. Note that the RPN Loss is omitted in the figure.

The Pipeline of Our Framework



- We need to pre-train a teacher model using the same dataset.
- The RoIs generated by the student model are shared with the teacher model.
- The parameters of the teacher model are frozen when we train the student model.

Hierarchical Distillation

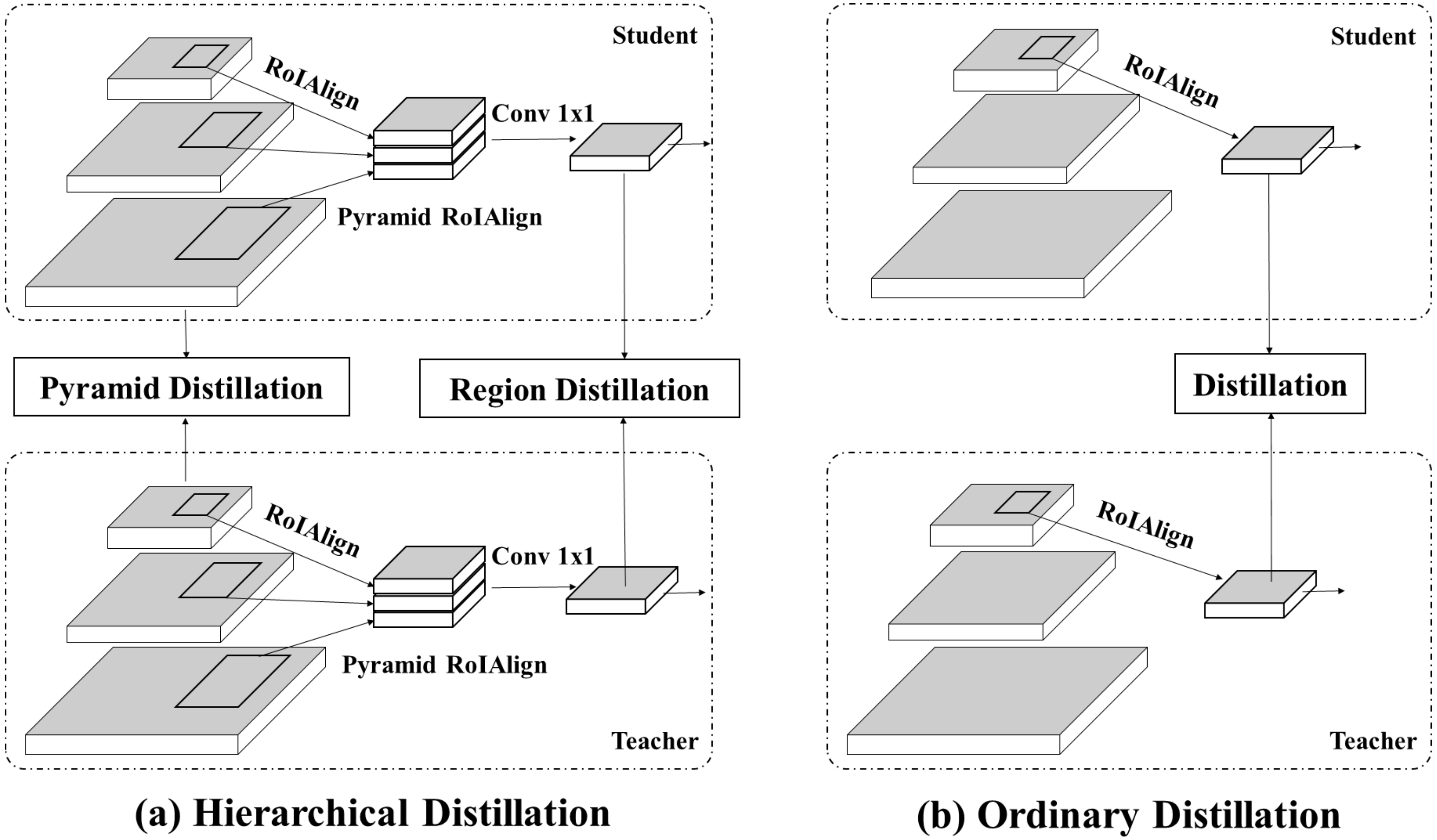


Fig. 4. Illustration of our proposed hierarchical knowledge distillation (a) versus the ordinary distillation (b).

Hierarchical Distillation:

For each RoI we crop features from **all levels** of the feature pyramid, followed by a concatenation and conv 1 × 1 operation.

Ordinary Distillation:

A certain level of feature in the feature pyramid is first found according to the area of each RoI, then the region feature is cropped within that level.

The Hierarchy of Our Framework:

Horizontally: The *Pyramid Distillation*, *Region Distillation* and *Logit Distillation* formulates our hierarchical distillation framework.

Vertically: Hierarchy is naturally inherited from the feature pyramid, that is, our *Pyramid Distillation* is performed on each level of feature pyramid, which ranges from low-level features with precise spatial response to high-level features with strong semantic meanings.

Loss Function for Knowledge Distillation

Pyramid Distillation Loss:

$$L_{PD}(P^s, P^t) = \frac{1}{N_{PD}} \sum_{i=2}^{5} \left\| P_i^s - P_i^t \right\|_2, \qquad (1)$$

Region Distillation Loss:

$$L_{RD}(R^s, R^t) = \frac{1}{N_R} \left\| R^s - R^t \right\|_2, \qquad (2)$$

Logit Distillation Loss:

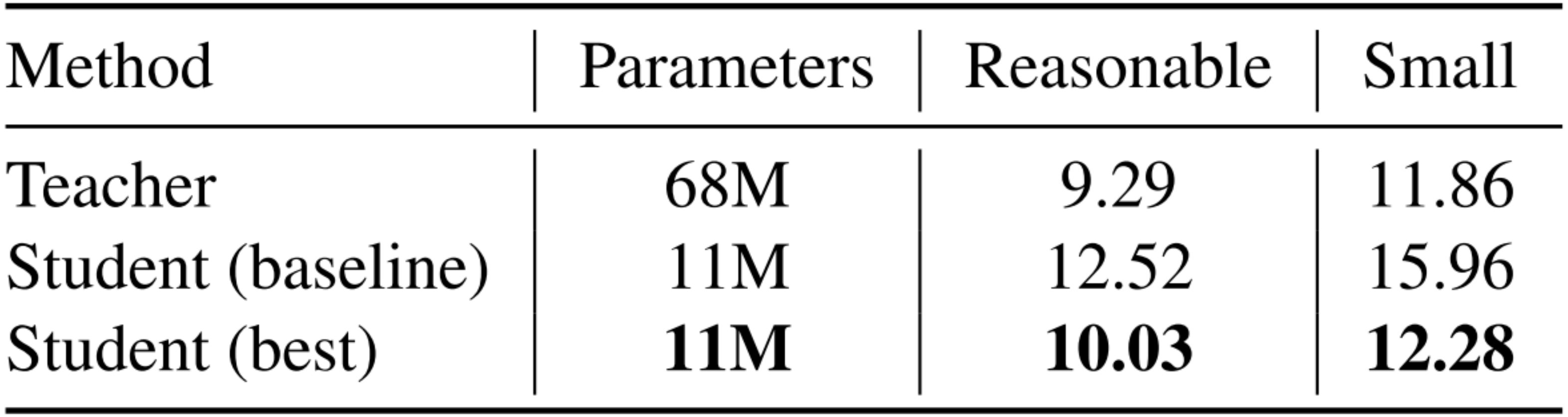
$$L_{LD}(G^s, G^t) = \frac{1}{N_G} \left\| G^s - G^t \right\|_2, \qquad (3)$$

Final hierarchical distillation objective:

$$L_{dist} = \lambda_{PD} L_{PD} + \lambda_{RD} L_{RD} + \lambda_{LD} L_{LD}, \qquad (4)$$

Experiment Result

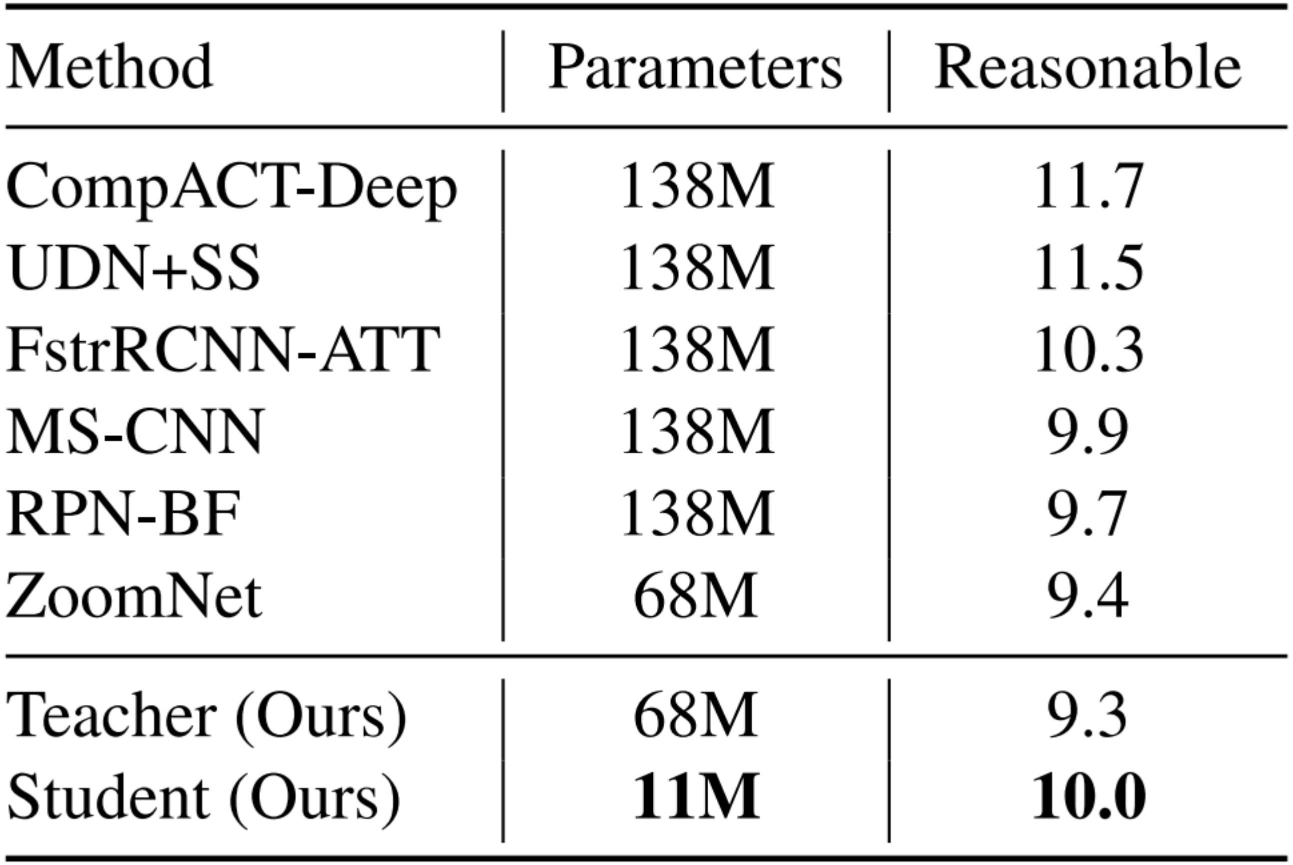
Table. 1. Comparison between student and teacher model. Performance is measured by MR (in %), lower value is better.



Method	Parameters	Reasonable	Small
Teacher	68M	9.29	11.86
Student (baseline)	11M	12.52	15.96
Student (best)	**11M**	**10.03**	**12.28**

Experiment Result

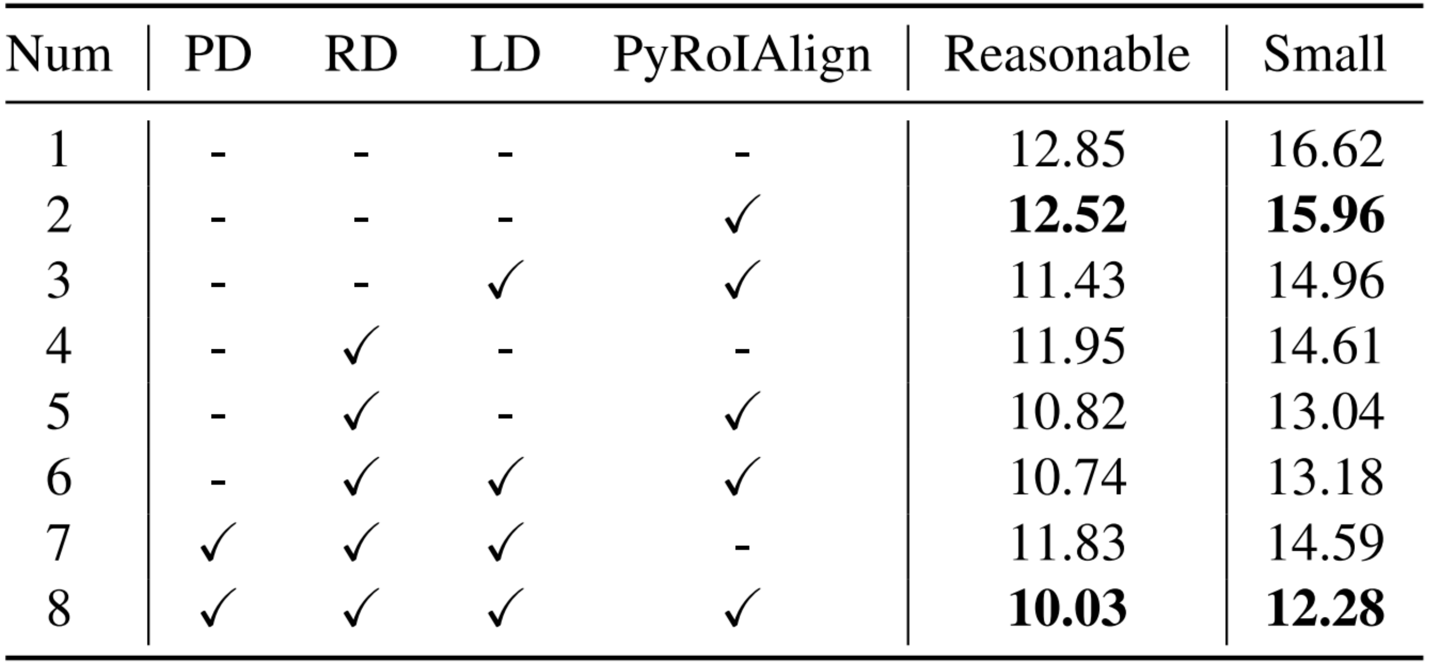
Table. 2. Comparison with other state-of-the-art methods.



Method	Parameters	Reasonable
CompACT-Deep	138M	11.7
UDN+SS	138M	11.5
FstrRCNN-ATT	138M	10.3
MS-CNN	138M	9.9
RPN-BF	138M	9.7
ZoomNet	68M	9.4
Teacher (Ours)	68M	9.3
Student (Ours)	**11M**	**10.0**

Ablation Study

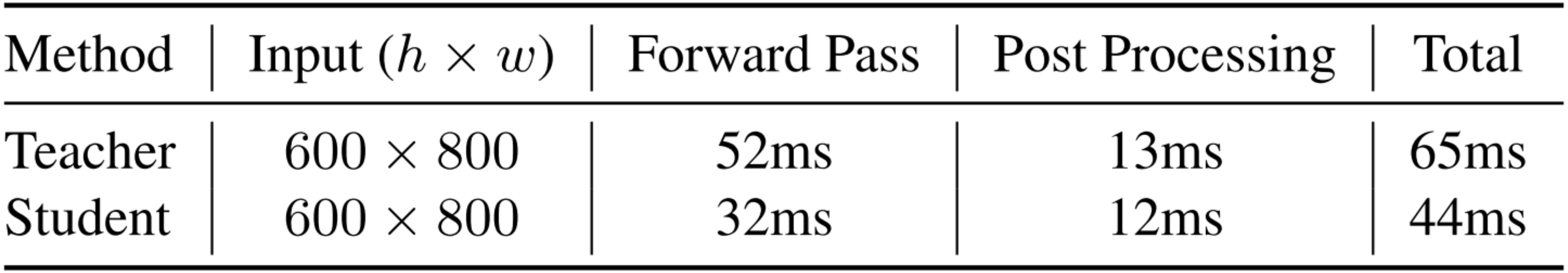
Table. 3. Comparison of multiple intermediate supervision.



Num	PD	RD	LD	PyRoIAlign	Reasonable	Small
1	-	-	-	-	12.85	16.62
2	-	-	-	✓	**12.52**	**15.96**
3	-	-	✓	✓	11.43	14.96
4	-	✓	-	-	11.95	14.61
5	-	✓	-	✓	10.82	13.04
6	-	✓	✓	✓	10.74	13.18
7	✓	✓	✓	-	11.83	14.59
8	✓	✓	✓	✓	**10.03**	**12.28**

Detection Speed

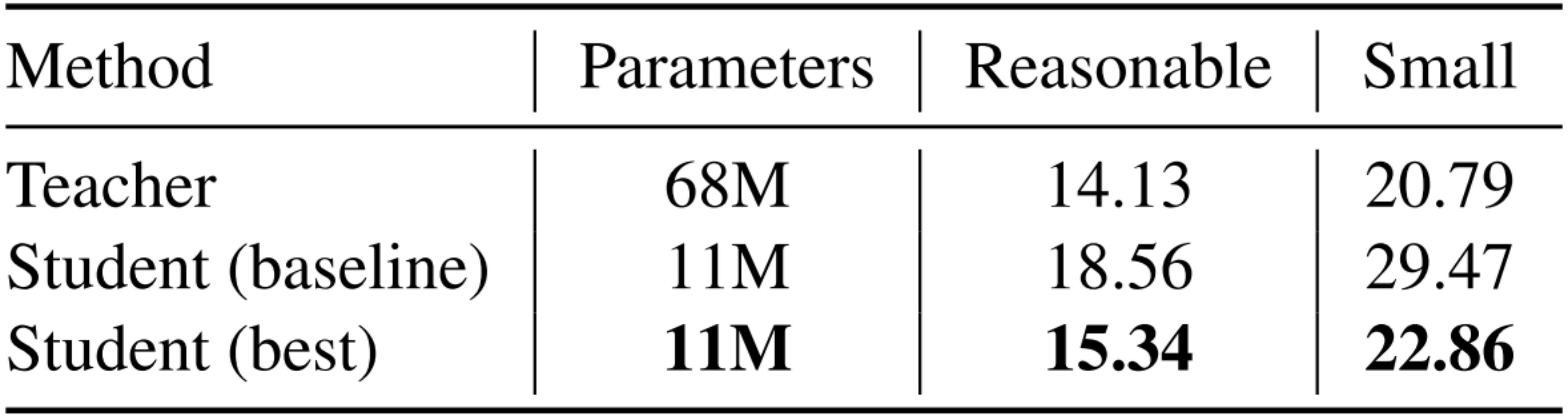
Table. 4. Comparison of inference time per image. ‘Forwd. Pass’ denotes the time for forward pass, and ‘Post Proc.’ means the time for post processing. Also the input size of image is of shape $h \times w$.



Method	Input ($h \times w$)	Forward Pass	Post Processing	Total
Teacher	600×800	52ms	13ms	65ms
Student	600×800	32ms	12ms	44ms

Experiment Result on *Citypersons* Dataset

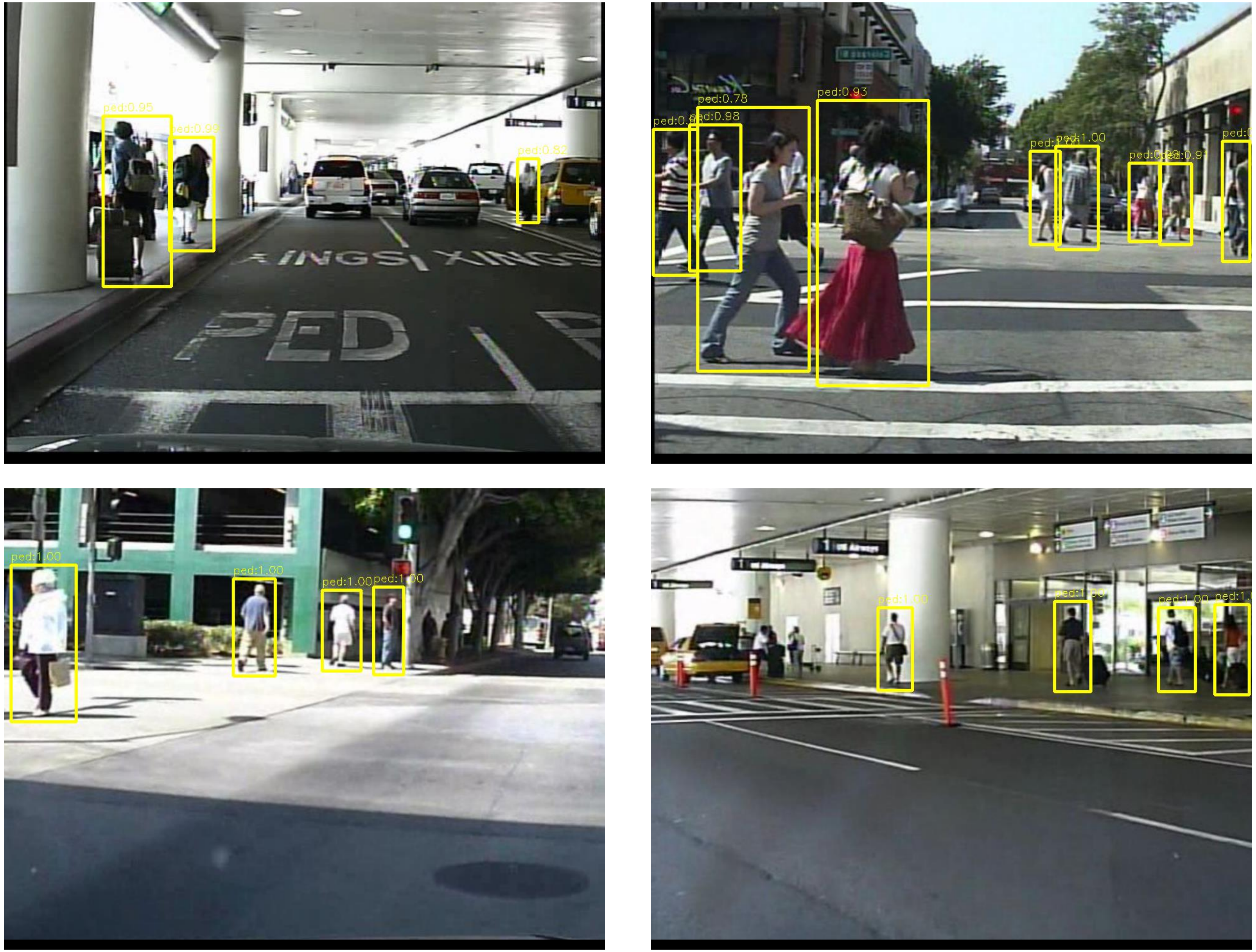
Table. 5. Comparison between student and teacher model on the *Citypersons* dataset. Also measured by Miss Rate (MR, in %).



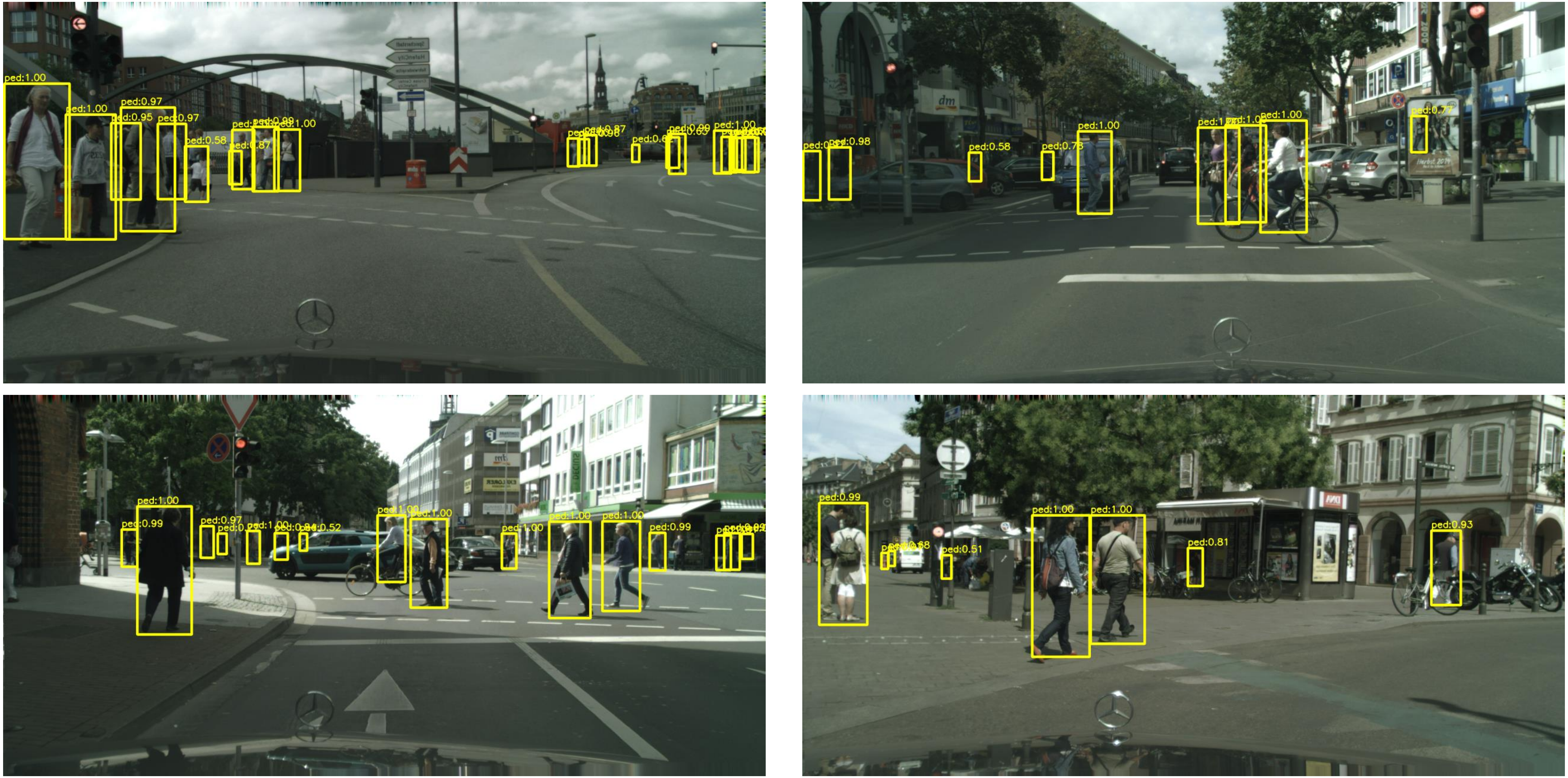
Method	Parameters	Reasonable	Small
Teacher	68M	14.13	20.79
Student (baseline)	11M	18.56	29.47
Student (best)	**11M**	**15.34**	**22.86**

清華大學
Tsinghua University

Detection Results on *Caltech* Dataset



Detection Results on *Citypersons* Dataset



Q & A

Codes are available here:

https://github.com/RuiChen96/MaskRCNN-Knowledge-Distillation

Email:

chenr18@mails.tsinghua.edu.cn

Thanks for your attention.