Master in
Computer Vision
Barcelona

M5 Project: Cross-modal Retrieval

Week 2

Introduction to Object detection and Instance Segmentation with Detectron2

Rubèn Pérez Tito
rperez@cvc.uab.cat

Ernest Valveny
ernest@cvc.uab.cat

P2 Obj Det and Seg. With Detectron 2

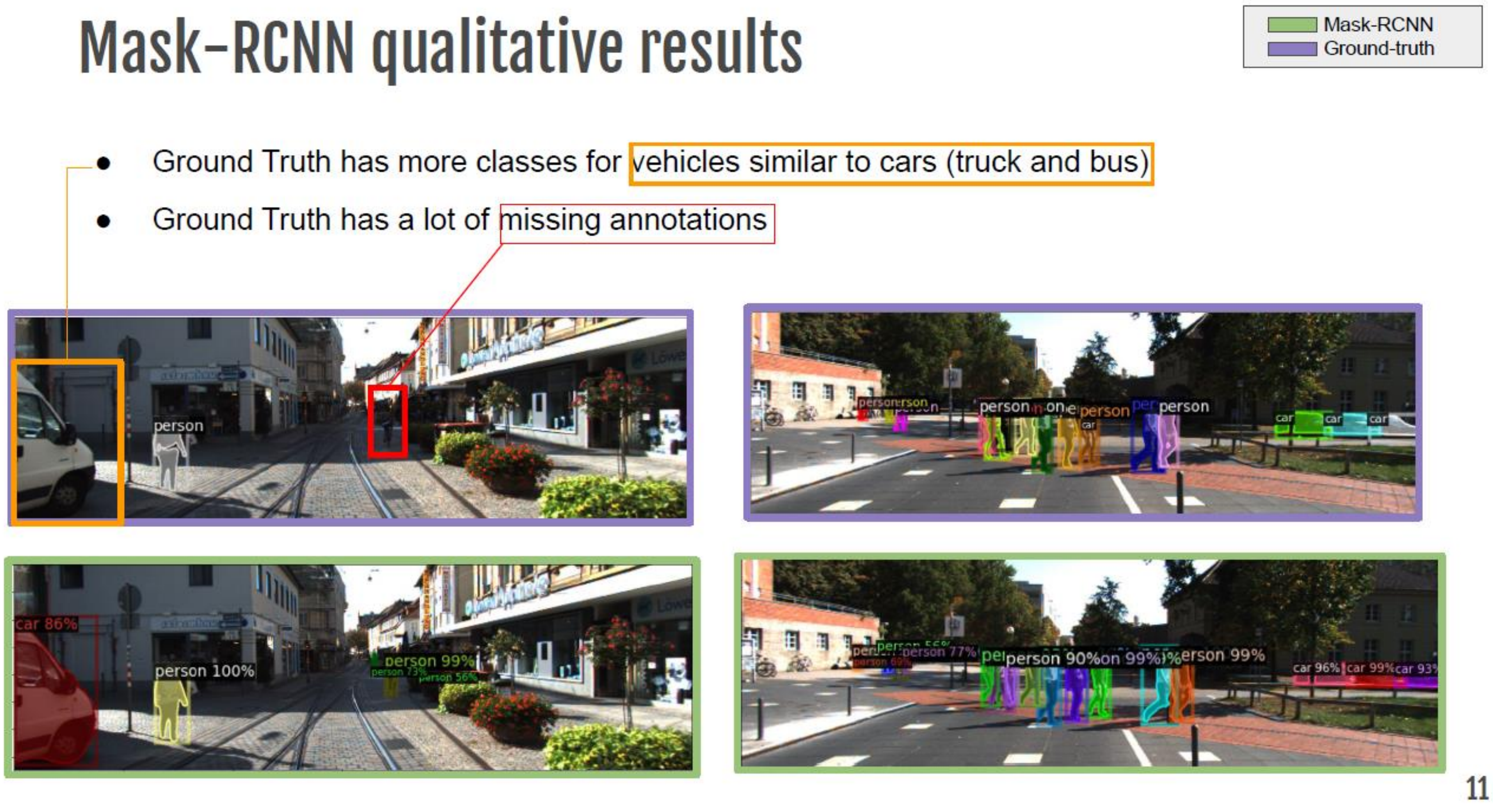
Metrics

- Explain what the metric is evaluating, don’t write just a formula.
- If we have different metrics, it might be interesting to do a specific analysis on some of them (if there is something relevant to say). Does pre-training affect the same both big and small objects?

Splits

- **New splits**. How you set-up your experiments is very important. How did you divide the two original splits into the three train, val and test sets?
  - Random split.
  - Based on sequences.
  - Keep the same un/balanced data distribution.

P2 Qualitative results



Mask-RCNN qualitative results

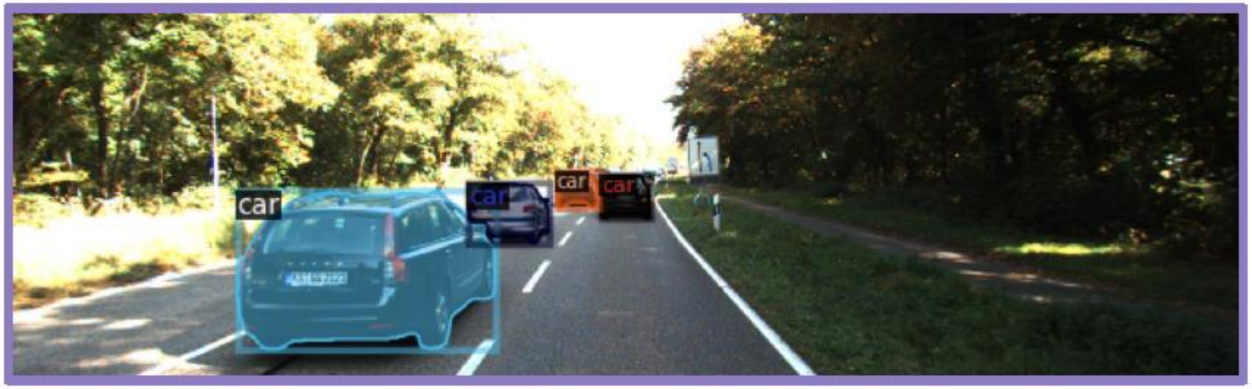
🟩 Mask-RCNN
🟪 Ground-truth

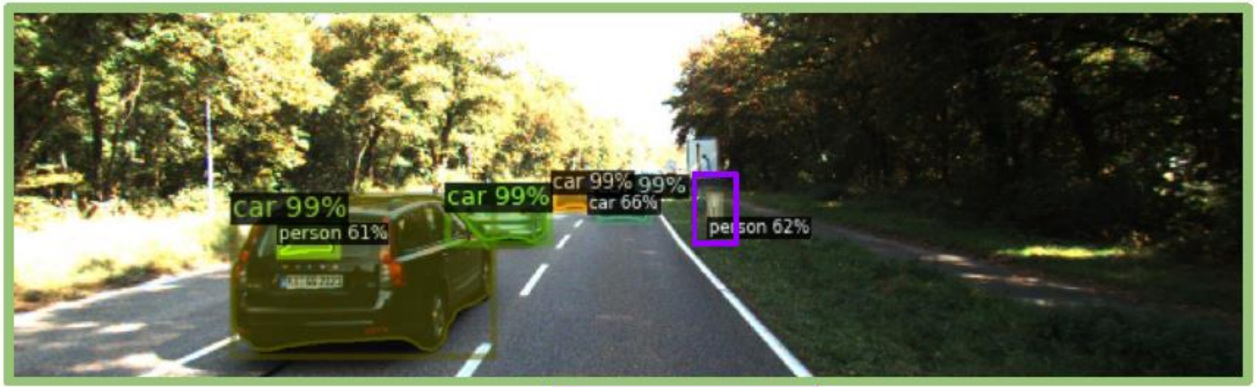
- Ground Truth has more classes for vehicles similar to cars (truck and bus)
- Ground Truth has a lot of missing annotations

P2 Qualitative results

Mask-RCNN qualitative results

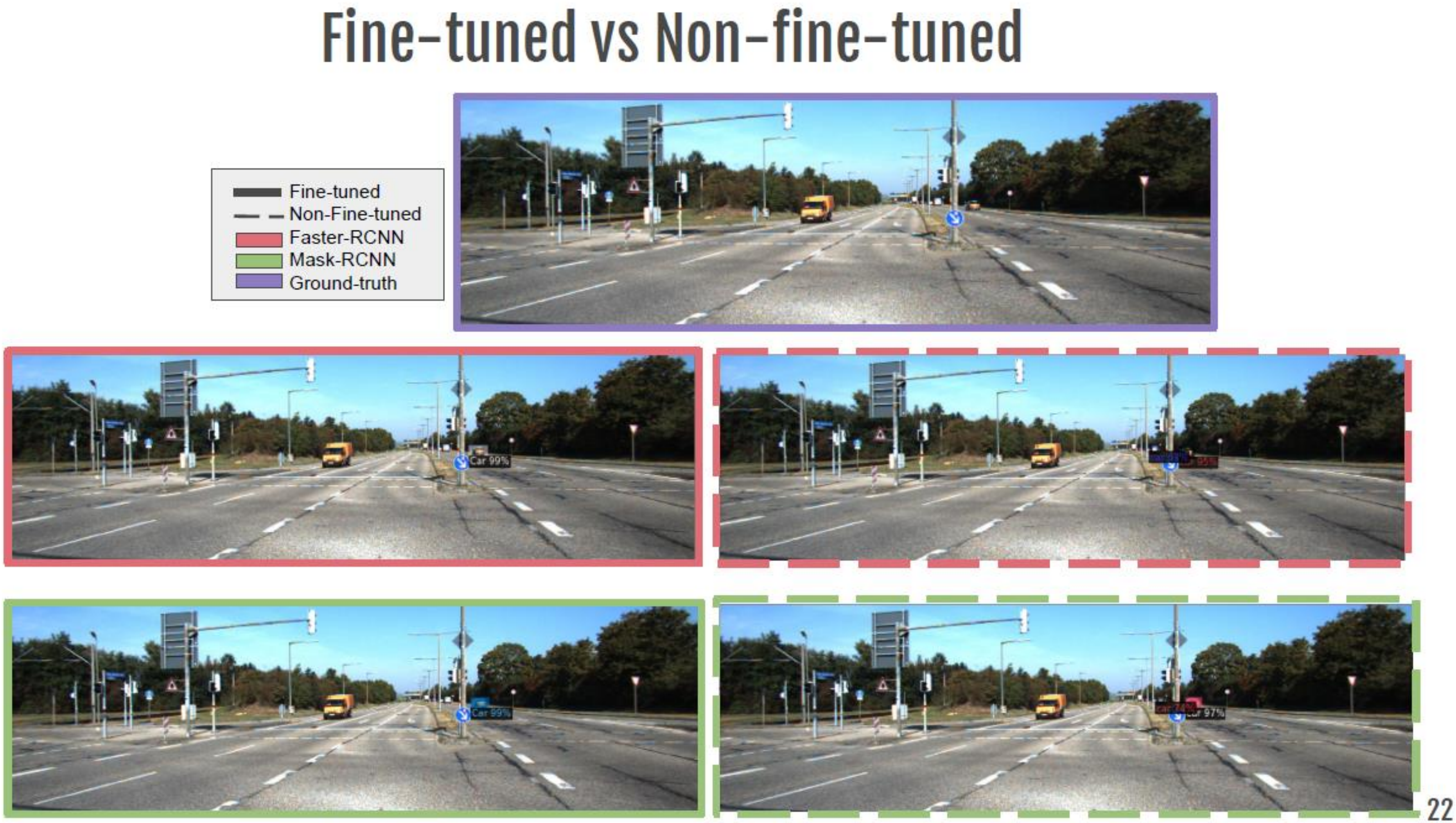
🟩 Mask-RCNN
🟪 Ground-truth





- Prediction sometimes confuses signs with class person

P2 Qualitative results



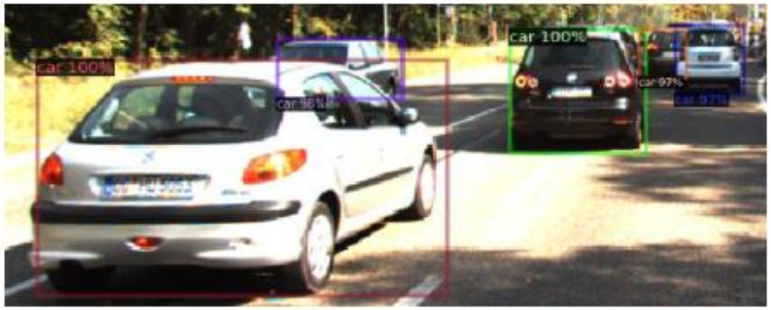
Fine-tuned vs Non-fine-tuned

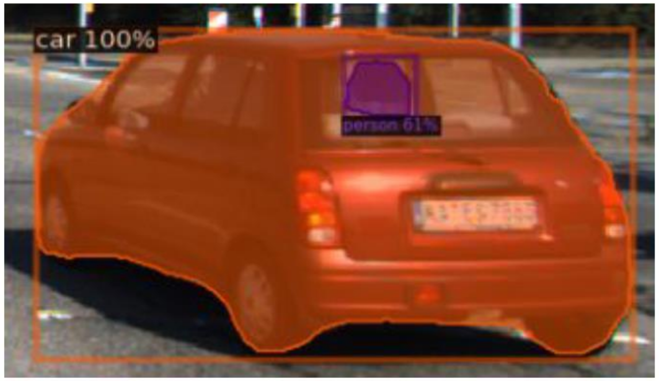
- ▬ Fine-tuned
- – – Non-Fine-tuned
- 🟥 Faster-RCNN
- 🟩 Mask-RCNN
- 🟪 Ground-truth

Using the same images helps to better assess model’s/experiments’ differences

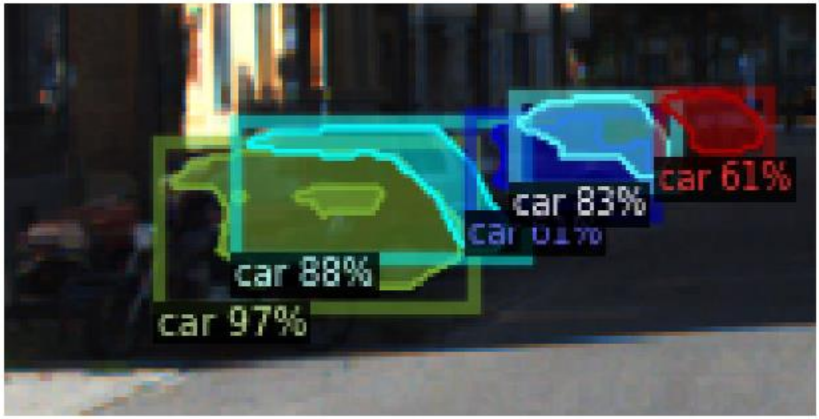
P2 Qualitative results

Good examples: Faster R CNN





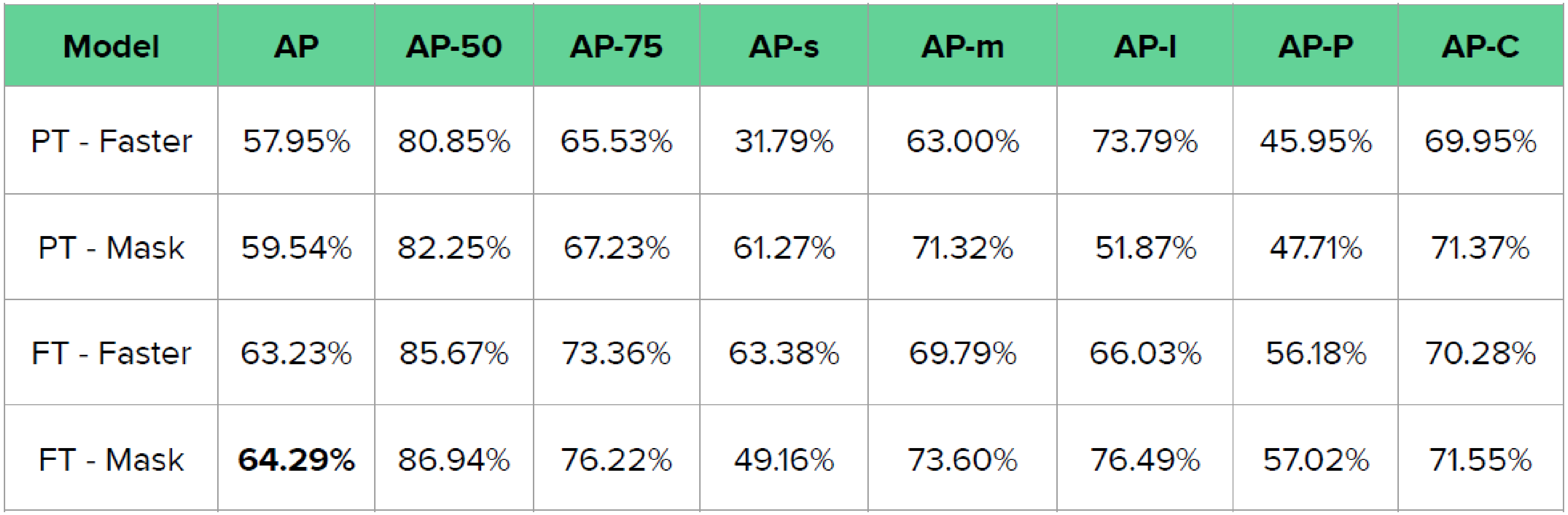
Bad examples



Show good and bad predictions for qualitative results

P2 Quantitative results

Grouping all the results in one table helps to better compare the different models / experiments.



Model	AP	AP-50	AP-75	AP-s	AP-m	AP-l	AP-P	AP-C
PT - Faster	57.95%	80.85%	65.53%	31.79%	63.00%	73.79%	45.95%	69.95%
PT - Mask	59.54%	82.25%	67.23%	61.27%	71.32%	51.87%	47.71%	71.37%
FT - Faster	63.23%	85.67%	73.36%	63.38%	69.79%	66.03%	56.18%	70.28%
FT - Mask	**64.29%**	86.94%	76.22%	49.16%	73.60%	76.49%	57.02%	71.55%

P2 Training / inference time

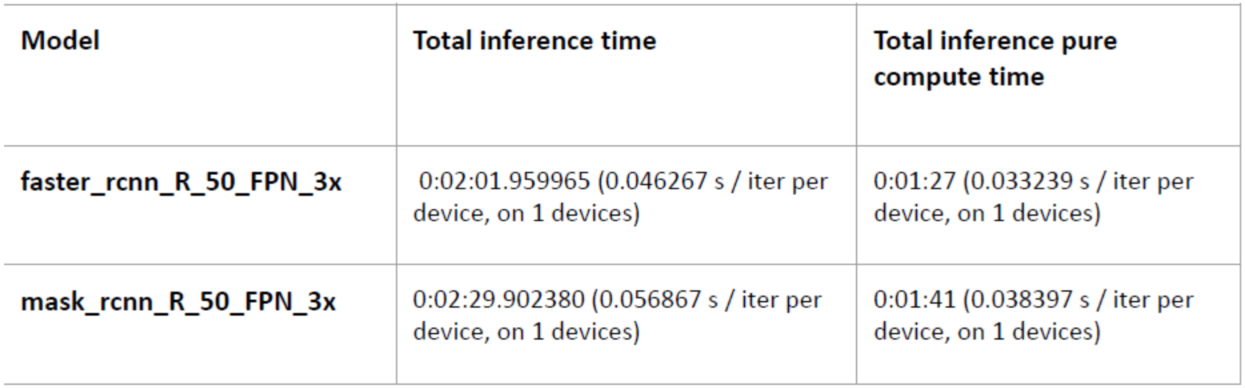
Using a RTX 3090 GPU

Training time:

	Faster RCNN	Mask RCNN
Total time	1h 35min	1h 55min

Inference time:

	Faster RCNN	Mask RCNN
Time per image (Sec)	0.0399	0.0489



Model	Total inference time	Total inference pure compute time
faster_rcnn_R_50_FPN_3x	0:02:01.959965 (0.046267 s / iter per device, on 1 devices)	0:01:27 (0.033239 s / iter per device, on 1 devices)
mask_rcnn_R_50_FPN_3x	0:02:29.902380 (0.056867 s / iter per device, on 1 devices)	0:01:41 (0.038397 s / iter per device, on 1 devices)

P2 Conclusion

- Mask R-CNN performs better for detection that Faster R-CNN.
- Fine-tuned models perform better than pretrained ones.
  - At least quantitative-wise.
- Cars are easier to detect/segment.
  - What if we balance the dataset to have the same samples for cars and pedestrians?
- Problems with distant, occluded or blurred objects.
  - Domain adaptation might be tricky: person != pedestrian