

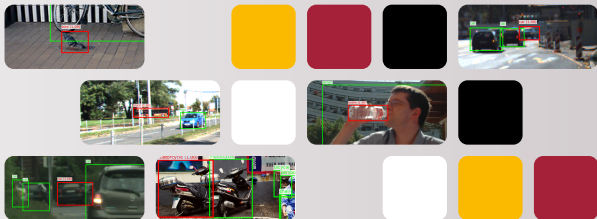
Object Detection Benchmarks are Incomplete: The Role of Label Errors and Annotation Uncertainty

Sarina Penquitt¹, Jonathan Klees¹,
Antonia van Betteray¹, Parssa Jashnieh¹,
Peter Stehr², Matthias Rottmann¹,
Lars Schmarje³

¹Institute of Computer Science, Osnabrück University

²Department of Mathematics, University of Wuppertal

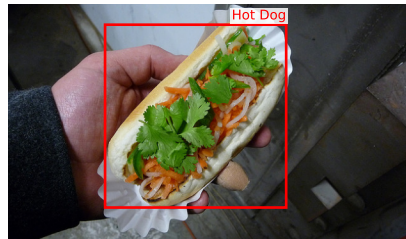
³Wayve



Why Are Object Detection Benchmarks Incomplete?

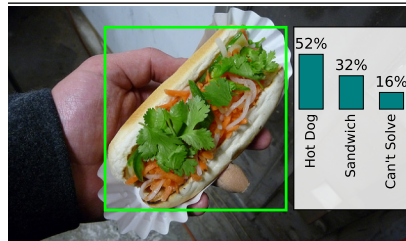
Annotation Incompleteness

- ▶ Small, occluded, and crowded objects are often missing
- ▶ Correct detections of these objects may be counted as false positives



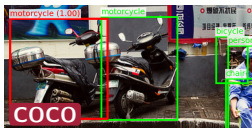
Missing Annotation Uncertainty

- ▶ Hard labels ignore class ambiguity
- ▶ Overconfident annotations may introduce biases
- ▶ This limits uncertainty-aware evaluation



Our Approach & Contributions

- We re-annotate four established datasets:



Re-Annotation Pipeline

- **High recall:**
Capture as many valid objects as possible
- **Soft labels:**
Aggregate class annotations from multiple annotators

Uncertainty-Aware Object Detection Benchmark

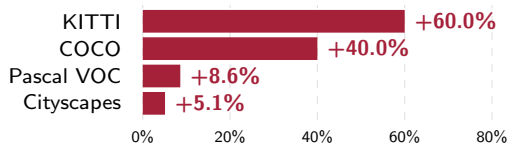
- **Annotation quality:**
Investigate its impact on model evaluation
- **Average Similarity (AS):**
Extend AP to soft labels

Real Label Error Detection Benchmark

- **Real label errors:**
Identify errors by comparing original and validated GT
- **Method evaluation:**
Evaluate methods on real rather than synthetic errors

Re-Annotation Results & Label Errors

Increase in Annotated Objects



False Negative Rate (%)

Dataset	Orig.	Valid.
Cityscapes	11.1	3.8
COCO	13.1	6.1
KITTI	41.5	11.6
Pascal VOC	18.6	9.3

Label Errors in Original Ground Truth (Counts)

Dataset	Missing	Misaligned	Classification
Cityscapes	5,249	2,385	85
COCO	16,219	7,736	254
KITTI	22,710	7,167	1,026
Pascal VOC	3,654	2,210	52

Missing annotations dominate label errors, while re-annotation substantially improves completeness

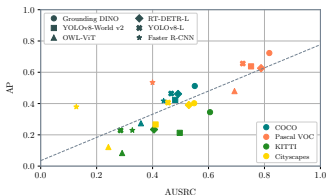
Object Detection Benchmark

Impact of Annotation Quality

KITTI (AP _{50:95} , %)	Orig.	Valid.
Grounding DINO	29.2	39.4
YOLOv8-L	74.7	26.7

- **Model comparison:** The pretrained YOLOv8-L drops substantially on validated GT, while Grounding DINO improves

Uncertainty-Aware Evaluation

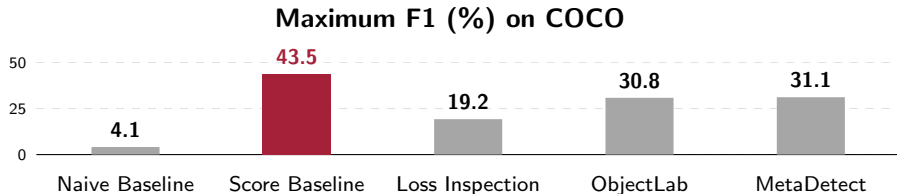


- **Average Similarity (AS)** generalizes AP to soft labels
- **AS vs. AP:** Strong positive correlation
- **Uncertainty alignment:** AS captures agreement with soft labels, unlike AP

Current object detectors are misaligned with human uncertainty

Label Error Detection Benchmark

- ▶ **Automatic label error detection:** Methods that identify potential label errors are often evaluated on synthetic errors
- ▶ **Our benchmark:** We identify real label errors across four datasets to evaluate existing methods



- ▶ **Across datasets:** Score baseline achieves the highest maximum F1 on all four datasets

Detecting real label errors remains challenging, even for state-of-the-art methods

Conclusion

- ▶ **Annotation completeness:** Missing objects are the dominant source of label errors in established object detection benchmarks
- ▶ **Annotation uncertainty:** Soft labels enable uncertainty-aware evaluation, capturing both detection performance and alignment with human uncertainty
- ▶ **Object detection benchmark:** Benchmark performance strongly depends on annotation quality
- ▶ **Label error detection benchmark:** Existing methods struggle to reliably detect real label errors

**High-recall, uncertainty-aware ground truth is essential
for more reliable object detection benchmarks**

Acknowledgment

Thank you for your attention!



With funding from the:



Federal Ministry
of Research, Technology
and Space

 **OSVIA**
Osnabrück Vision AI Lab

