CENTER FOR BIOSYSTEMS AND BIOTECH DATA SCIENCE

Esla Timothy Anzaku, Mohammed Aliy Mohammed, Stefan Magez, Sofie Van Hoecke, Arnout Van Messem, and Wesley De Neve

UNCERTAINTY-AWARE EVALUATION OF DEEP LEARNING OBJECT DETECTORS UNDER SCARCE AND EVOLVING TEST DATASETS

Problem: Tracking deep learning model performance is challenging when train/test datasets evolve

Motivation: Datasets evolve, e.g., annotations get refined, new species and under-represented samples get added

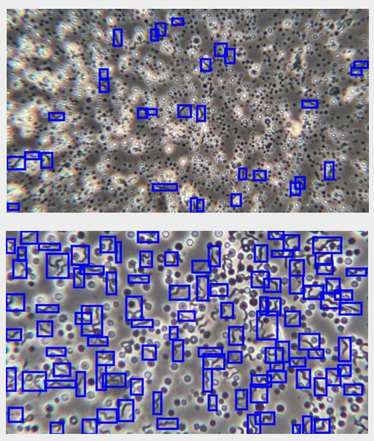
Challenge: Standard metrics (e.g., mAP, recall, precision) can be misleading across dataset versions

Our Approach: We introduce an uncertainty-aware methodology to tracking model effectiveness

Demonstration: Applied to 8 object detection models across 3 dataset versions for trypanosome parasite detection

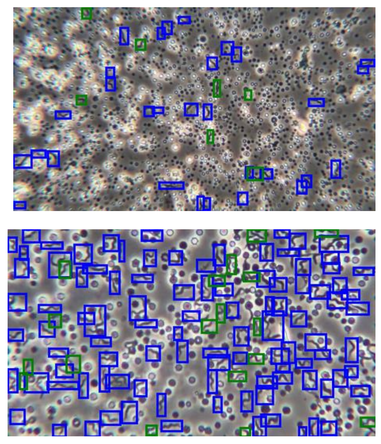
Example Images from the Dataset Versions

Version 1



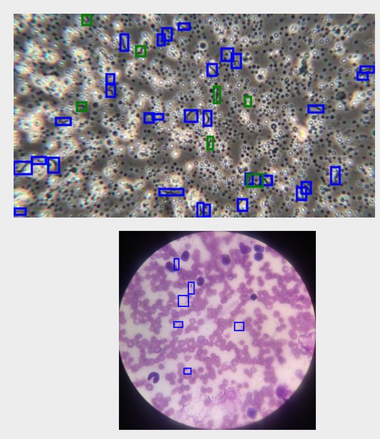
- Tryp dataset [1]
- Published in 2023
- *Trypanosoma brucei brucei*
- Label-free
- Thick blood smears
- Published in 2023

Version 2



- Same images as Version 1
- Labels enhanced in 2023/2024

Version 3



- Contains Version 2
- Contains *Trypanosoma cruzi* [2]
- Thick and thin smears
- Label-free
- Giemsa stained

1. Conventional Evaluation Approach

How it is Done

- Multiple versions of a single dataset are rare
- When they exist
  - Model comparisons are typically based on mAP due to precision-recall trade-offs
  - Models are evaluated on only one version, or
  - Models are evaluated on all versions and reported separately (sometimes averaged)

Limitations

- **Dataset Bias:** Different versions may introduce various biases (e.g., varying degree of annotation errors, inclusion of easy/difficult images, class imbalance)
- **Performance Ambiguity:** Observed improvements or declines may reflect dataset changes rather than true model capability
- **Metric Limitations:** Conventional metrics often fail to identify clear winners across dataset versions

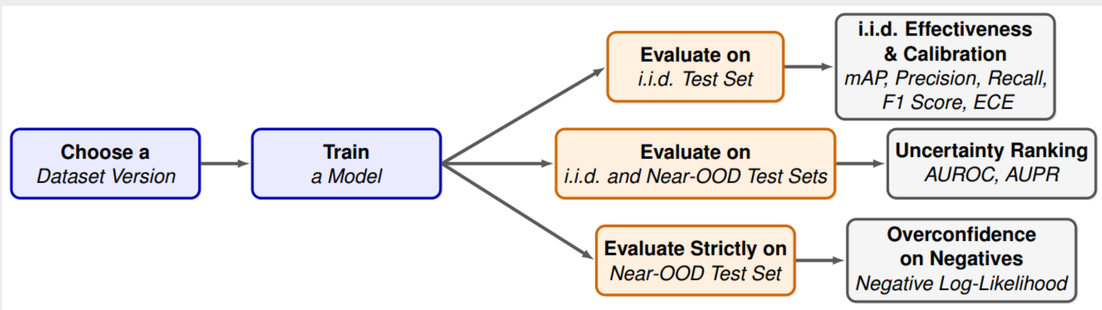
Implications

- **Static Dataset Bias:** Reliance on fixed datasets persists, though high-quality test data are crucial
- **Model Selection Risk:** Models better aligned with reliability expectations may be overlooked in favor of SOTA models optimized for standard metrics

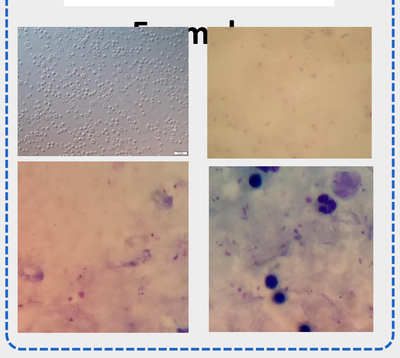
2. Proposed Evaluation Approach (cont.)

Methodology

- We leverage near-OOD images and negative log-likelihood to complement model effectiveness and uncertainty ranking metrics

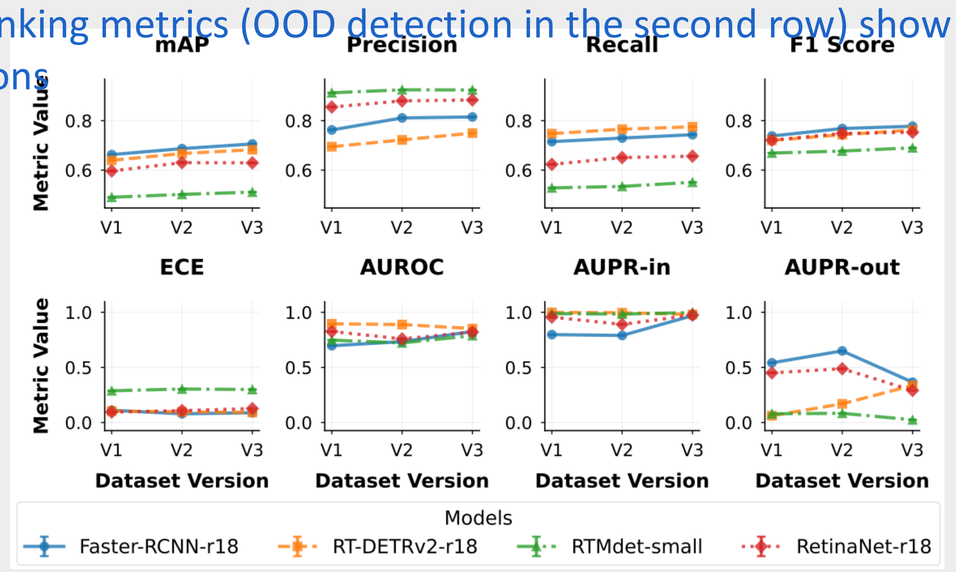


Near-OOD



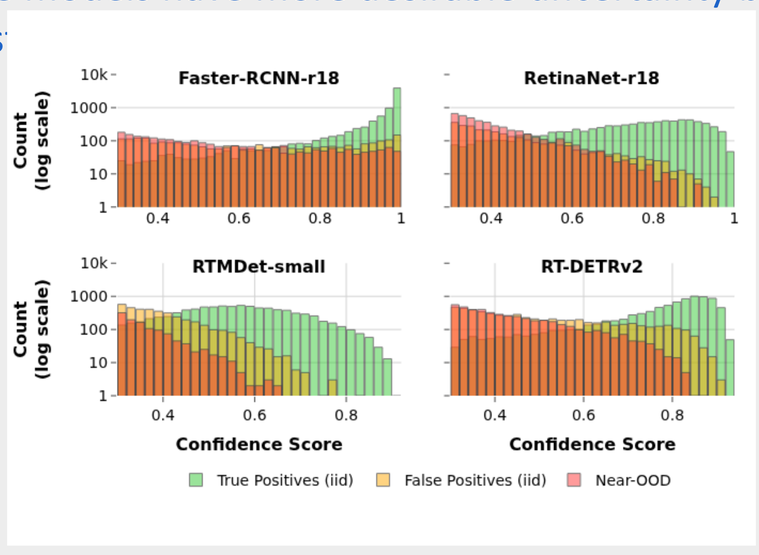
Findings (1/2)

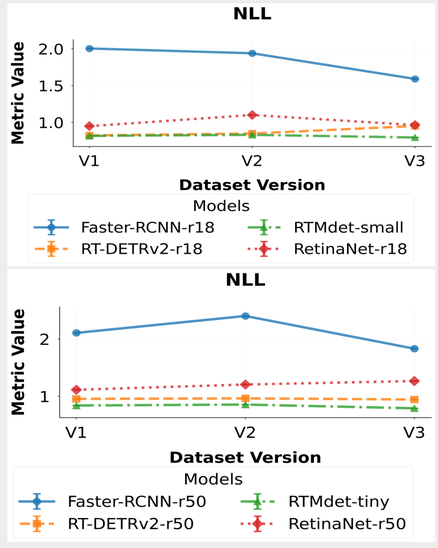
- Models appear to be smoothly improving across dataset versions under i.i.d. metrics (top row)
- However, ranking metrics (OOD detection in the second row) show more variations across versions



Findings (2/2)

- Better performing under i.i.d. metrics does not necessarily translate to better uncertainty behavior (e.g., Faster-RCNN)
- Some models have more desirable uncertainty behavior and are worth further inves





2. Proposed Evaluation Approach

The Intuition is to:

- Complement, rather than replace, conventional metrics
- Track properties of reliable models expected to remain invariant across dataset versions
- Expand evaluation rigor by providing additional signals without compromising existing assessment
- Leverage images plausible during deployment but lacking target objects; referred to as near-OOD

[1] Anzaku *et al.* Tryp: a dataset of microscopy images of unstained thick blood smears for trypanosome detection, 2023.

[2] Morais *et al.* Automatic detection of the parasite *Trypanosoma cruzi* in blood smears using a machine learning approach applied to mobile phone images, 2022.

Esla Timothy Anzaku@ugent.be

GHENT UNIVERSITY

GHENT UNIVERSITY GLOBAL CAMPUS

VUB VRIJE UNIVERSITEIT BRUSSEL

LIÈGE université Mathematics

MICCAI2025 Daejeon REPUBLIC OF KOREA