A diagnostic toolkit to quantify consistency and refine label quality in object detection and other tasks.

KαLOS finds Consensus:

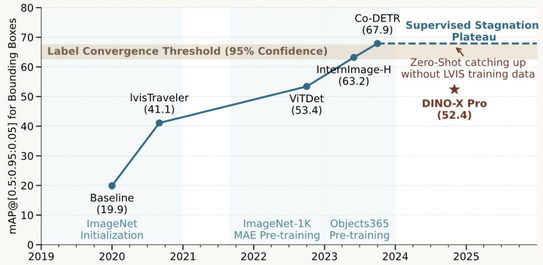
A Meta-Algorithm for Evaluating Inter-Annotator Agreement in Complex Vision Tasks

David Tschirschwitz and Volker Rodehorst

Bauhaus-Universität Weimar, Germany

I. Motivation and Problem Description

1. Motivation: Overcome Stagnation in Object Detection



History of top-performing models on LVIS v1.0 val, showing performance improvements stagnating at the label convergence threshold.

2. Problem: Contradicting Annotations cause a Bottleneck

Label noise creates a **Supervised Stagnation Plateau**, capping model progress. Improving data quality is now more critical than scaling architectures.

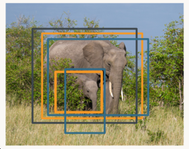
Our Solution: KαLOS

We introduce **KαLOS** to unify Inter-Annotator Agreement (IAA) across complex tasks. It identifies systematic errors in human, synthetic and auto-annotated data.

LVIS Example: Disputed 'Train' topology (Lead car vs. entire consist). How do we quantify consensus for such conceptual ambiguity?

II. Method: Evaluate Dataset Quality

1. Input: Active Disagreement



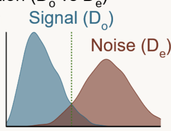
Rater A
Rater B
Rater C

2. Principled Configuration

Data-Driven Calibration (D_o vs D_e)

τ^* Calibration Anchor

d_{loc} Distance Function

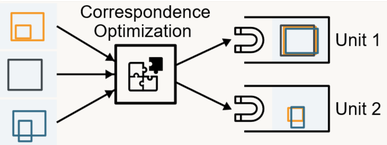


Empirical Defaults

ψ_{soft} Cost Function

S Greedy Solver

3. KαLOS Execution



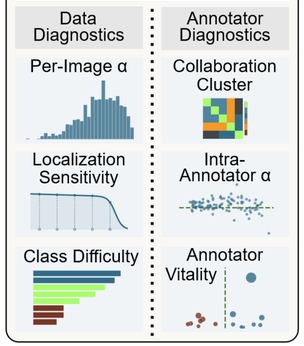
Reliability Matrix

	Unit 1	Unit 2
Rater A	Elephant	Elephant
Rater B	Elephant	No Object
Rater C	Elephant	Hippo

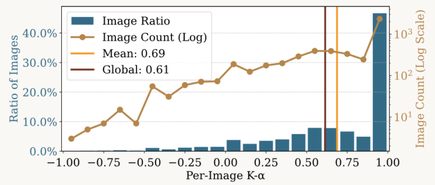
Agreement Calculation

$$\alpha = 1 - \frac{D_o}{D_e}$$

4. Downstream Analysis



5. Example: Per-Image α on LVIS



Beyond a Single Score (Diagnostic Pipeline):

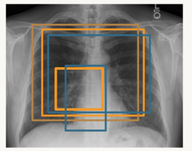
- Guideline Audit: Identify ambiguous classes.
- Rater Vitality: Measure individual consistency.
- Precision Limits: Quantify spatial sensitivity.

GitHub:

III. Method Validation

1. Error Extraction

Decomposing Human Variance



Rater A
Rater B
Rater C

Δ_{loc}

Δ_{cls} ? Cls A vs Cls B

Δ_{um}

Δ_{topo}

2. Noise Modelling

Data → Fit → Test

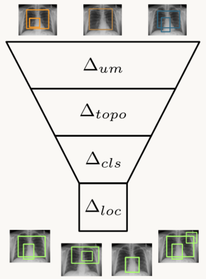
Parametric Drivers (e.g. Student's t, Von Mises)

Validated Proxies (Foundational Models)

Dual-Layer Framework: Create data-driven noise generator derived from empirical extraction process.

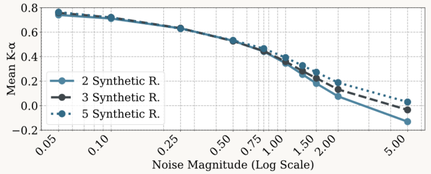
3. Generator Composition

Reference Annotations



Synthetic Annotations

4. Evidence Based on Synthetic Test-Suit



We validate the core properties of KαLOS (monotonic decrease, full-range utilization, and no sudden jumps) on a synthetic testbed with increasing noise magnitudes, based on an empirically derived noise generator.

www.uni-weimar.de