Behind the Labels: Transparency Pitfalls in Annotation Practices for Societally Impactful ML

Author: Claudia Scorția (c.scortia@tudelft.nl), Supervisor & Responsible Professor: Andrew M. Demetriou, dr. Cynthia Liem

Introduction

- ML models in critical areas like healthcare and autonomous driving rely on high-quality annotated data, yet annotation practices are often underreported, risking bias and unreliability.
- This study analyzes annotation and dataset reporting in top computer vision research, focusing on IEEE/CVF Conference on Computer Vision and Pattern Recognition publications.

Background

- Building on Geiger et al. (2020), who highlighted poor transparency in annotation reporting, this project examines how annotations are collected, validated, and disclosed in ML research claiming societal impact.
- Prior studies [2][3] reveal systemic issues in annotation quality and interpretation, emphasizing errors in learning from data and challenging assumptions about human annotation reliability.

Research Question

What are the data collection and reporting practices of human annotations/labels in societally impactful ML research published at IEEE/CVF CVPR?

Sub-questions:

RQ1: What reporting elements count as transparent dataset documentation?
RQ2: How are human annotations collected in these papers?
RQ3: How is annotation quality assessed?

Methodology

This is a systematic literature review of 25 CVPR papers sampled from each three time periods:

- 2023-2024
- 2020-2024
- 2010-2024

Each CVPR paper was analyzed to identify the datasets used. After gathering all the results, the dataset papers were reviewed in order of their importance.

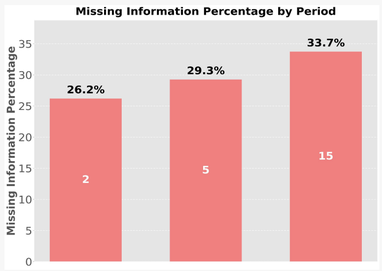
Each dataset paper was analyzed with focus on three aspects:

- **items**: outcome; annotations per item; original labels; produced by humans or not; overlap.
- **annotators**: the background of annotators; the recruitment and prescreening methods employed; training or qualification protocols; compensation models ; any quality control and any other reported information.
- **annotation practices**: annotation schema mentioned and its rationale.

Following the structured annotation of each dataset paper, we conducted a focused analysis across the three key dimensions. This step aimed to identify common trends, highlight gaps in transparency, and assess the consistency of reporting practices. For analyzing the results we used the code at: https://github.com/Gargant0373DatasetAnalysis

Findings

- *Deep residual learning for image recognition* is the most cited paper in CVPR by far.
- The community is changing from Pascal VOC 2007 & 2012 to COCO



CVPR papers show an improvement in the datasets used. Each period represents the CVPR papers published within that time, not the datasets. →

When inspecting each field for each dataset, top 10 most undocumented fields are:

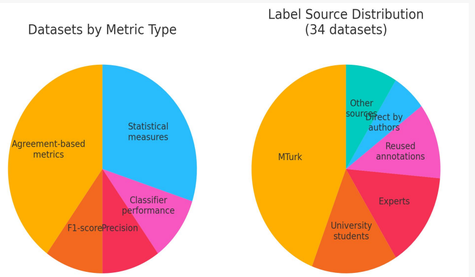
- **Total labellers** (67.21%),**Labeller Population Rationale** (62.30%), **Prescreening** of the annotators (52.46%), **Compensation** (40.98%), **IRR** (40.98%), the **Metric** used (40.98%), and the **Training** offered to the annotators (39.34%)
- **Sample size** (54.10%) and its **rationale** (45.90%) and the **Label Threshold** (39.34%).

Fields with the lowest missing rate when documented:

- Prescreening: 23.72%
- Annotators per item: 26.65%
- Item Sample Size Rationale: 28.16%
- Compensation: 28.46%
- Labeller Population Rationale: 28.57%

Fields with the greatest impact on completeness:

- Overlap: 33.0 pp
- Formal Instructions: 28.2 pp
- A priori Annotation Scheme: 19.51 pp
- Discussion: 2.18 pp
- A priori sample size: 0.87 pp



Discussion

RQ1: Datasets that documented prescreening practices reported a small amount of missing information on our metadata checklist items. Also, datasets discussing the overlap of labels and the formal instructions showed a significant impact on documenting the annotation process.

RQ2: The high use of Amazon Mechanical Turk for recruiting annotators points out once more the importance of prescreening and training the people hired for annotating a new dataset.

RQ3: No dataset used standard psychometric measures like Cohen's Kappa, Fleiss' Kappa, or Krippendorff's Alpha. This absence of consistent quality metrics highlights a significant gap in how datasets are validated and undermines trust in label reliability. →

Implications of Machine Learning

Standardizing annotator reporting via a single template—such as Datasheets for Datasets [4] or Data Cards [5] —ensures every dataset release includes key details on demographics, qualifications, training, pay, and reliability. Embedding this template into publication, repository, and tool workflows automates metadata capture and promotes transparency, reproducibility, and trust in AI systems.

Conclusion & Future Work

~30 % metadata gap: Key annotator details (count, training, screening, quality checks) are routinely missing.

Core fields matter: Reporting formal instructions, prescreening and overlap strongly boosts overall transparency.

Mandate templates: Integrate a single annotation-reporting template (e.g., Datasheets or Data Cards) into publication and repository workflows.

Next steps: Broaden scope beyond CVPR and perform qualitative studies of annotation decisions.

References

[1] R. S. Geiger, D. Cope, J. Ip, M. Lotosh, A. Shah, J. Weng, and R. Tang, "'Garbage in, garbage out' revisited: What do machine learning application papers report about human-labeled training data?,"Quantitative Science Studies, vol. 2, no. 3, pp. 795–827, 2021. Doi: 10.1162/qss_a_00144.

[2] J. Hullman, S. Kapoor, P. Nanayakkara, A. Gelman, and A. Narayanan, "The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning," in Proc. 2022 AAAI/ACM Conf. AI, Ethics, and Society (AIES '22), Oxford, United Kingdom, 2022, pp. 335–348. doi: 10.1145/3514094.3534196.

[3] L. Aroyo and C. Welty, "Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation", AIMag, vol. 36, no. 1, pp. 15–24, Mar. 2015.

[4] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Vaughan, Hanna Wallach, III Daume'e, and Kate Crawford. Datasheets for datasets. Communications of the ACM, 64, 03 2018.

[5] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, page 1776–1826, New York, NY, USA, 2022. Association for Computing Machinery.