

Benchmark Blindspots: A systematic audit of documentation decay in TPAMI's datasets

IEEE Transactions on Pattern Analysis and Machine Intelligence



Introduction

Abstract

A structured audit of 75 top-cited TPAMI vision papers (2009–2024) reveals that 37% of essential dataset annotation metadata is missing. Critical gaps span annotator recruitment, training, compensation, and inter-rater reliability (IRR). While a few benchmarks demonstrate best practices, most provide unverifiable “ground truth,” undermining reproducibility and fairness.

Questions

- How transparent are TPAMI papers about dataset annotation practices?
- Do citation counts correlate with better documentation?
- Are recent papers more transparent?
- What metadata is most frequently missing?
- Are there consistent co-reporting patterns?

Research

Background

Machine learning hinges on trustworthy data. Yet many benchmark datasets—widely used and highly cited—are built through opaque annotation pipelines. Poor documentation undermines model reliability, reproducibility, and fairness.

Methodology

Sample: 75 TPAMI papers (2009–2024)

Checklist: 27 annotation metadata fields (adapted from Geiger et al.)

Phases:

- **Paper Selection** (Tab 1): Stratified sample across 3 time periods.
- **Dataset Compilation** (Tab 2): 838 datasets collected; 64 evaluated.
- **Annotation Audit** (Tab 3): Metadata extracted + analyzed via Cramer’s V, Spearman’s ρ , Pearson’s r .

Key Findings

37.03% of metadata fields were missing overall.

Field	Missing %
Labeller Population Rationale	76.6%
Prescreening	73.4%
Total Labellers	68.8%
Compensation	67.2%

Citation count **does not correlate** with better reporting
Documentation has **not improved** significantly over time

Correlated Gaps

Some metadata fields tend to be missing or present in tandem:

- **IRR & Metric** (Cramer’s $V = 0.81$)
- **Human Labels & Label Source** ($V = 0.7$)
- **Compensation & Training** ($V = 0.45$)

Examples of Good Practice

Datasets like **ImageNet-Real**, **Cityscapes**, **MPII** show >75% completeness- due to multi-institutional support, bencharm alignment and supplementary materials.

Responsible Research

- Open code and data.
- Reproducible pipeline; all scripts and annotations available.

Conclusion

Conclusion

Over 37% of key annotation details in TPAMI papers are missing—especially around who labeled the data and how. This threatens reproducibility and trust. While a few benchmarks show strong practices, most fall short. Clear, standardized reporting must become a baseline for future datasets.

Recommendations

- Require **checklist-based annotation statements**.
- Adopt standards like **Datasheets for Datasets**
- Extend audits to other venues

Contact info

Alex Despan,
a.despan@student.tudelft.nl