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

Alex Despan
Supervisors: Andrew Demetriou, dr. Cynthia Liem
EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Alex Despan
Final project course: CSE3000 Research Project
Thesis committee: dr. Cynthia Liem, Andrew Demetriou, dr. Jie Yang

An electronic version of this thesis is available at http://repository.tudelft.nl/.

*Transactions on Pattern Analysis and Machine Intelligence, one of the most influential Machine Learning venues

Abstract

High-impact vision research still rests on datasets whose labels arrive via opaque, rarely documented pipelines. To understand how serious the problem is inside a large venue, we audited 75 TPAMI papers (2009-2024) that rely or introduce datasets. Each dataset was coded against a 27-item checklist adapted from *Garbage in, Garbage out*, spanning annotator recruitment, training, compensation, overlap-resolution and more.

Across the corpus, 37% of the expected annotation metadata is missing; the rate changes little between recent (2022-24) and older cohorts. The scarcest fields are labeller-population rationale (76.6% absent), prescreening criteria (73.4%), total annotators (68.8%), compensation (67.2%) and training procedures (62.5%). Documentation quality shows virtually no correlation with a paper's citation impact, suggesting community prestige does not buy transparency.

A handful of well—curated datasets achieve >75% completeness, proving that thorough documentation is possible when incentives align. The median TPAMI benchmark still ships with an unverifiable "ground truth", threatening the reproducibility and fairness claims of downstream models.

We advocate that journals and conferences require a concise, checklist-based annotation statement, mirroring existing ethics and reproducibility forms, to ensure future vision systems are built (and evaluated) on transparent, trustworthy data foundations.

1 Introduction

1.1 Background and Motivation

Machine learning systems depend fundamentally on the quality of their data, used both to train and evaluate the model's performance. Often, the reference labels are noisy, biased, or even unreliable, which can result in models that produce misleading predictions. As Geiger et al. [35] observed, much of the machine learning research focuses on algorithmic development once a "state-of-the-art" dataset is available, ignoring the possibility that the dataset itself is not trustworthy. This carries serious implications—such as the continual masking of biases—and undermines our ability to understand when and why automated systems fail in real-world scenarios.

Despite the crucial role of human-generated annotations in constructing these datasets, reporting practices around how these labels are collected—who produces them and what quality-control measures are applied—are often virtually nonexistent [37]. In their systematic survey of applied ML papers, Geiger et al. [35] found that most publications offer almost no information about their annotation protocols. This lack of transparency makes it nearly impossible to determine whether a model's claimed performance reflects real-world robustness or merely overfits on biases embedded in the data collection process.

Several proposals have aimed to improve dataset documentation standards in machine learning. Notably, the "Datasheets for Datasets" framework [32] and "Data Statements for Natural Language Processing" [6] offer structured templates that encourage researchers to report key aspects of dataset creation, including annotation methods, annotator demographics, and quality control. However, uptake remains inconsistent, and many high-profile publications

in leading venues continue to omit even basic information.

A central concept in assessing annotation quality is inter-rater reliability (IRR), which captures the degree of agreement between annotators beyond chance. Despite its importance, IRR is rarely reported or discussed in detail in most machine learning papers. Without it, the robustness of models trained on such data cannot be properly judged.

This study uses multiple statistical measures to evaluate associations between metadata fields. For categorical variables (e.g., *Yes*, *No*, *No information*), we rely on **Cramer's V** C, which quantifies the strength of association between two nominal variables using the chi-squared statistic. It ranges from 0 (no association) to 1 (perfect association), and does not assume any ordering or distribution. For fields with ordinal or ranked values, such as Likert-style responses or quality scores, we use **Spearman's** ρ C, which measures monotonic relationships based on rank-order correlation. When both variables are continuous and normally distributed, we apply **Pearson's** r C, which captures linear relationships between them. These three metrics together allow for flexible and appropriate analysis across the varied data types present in the annotation schema.

This persistent gap between the importance of "ground truth" and the absence of systematic reporting standards motivated the current study. While previous work has surfaced these problems across broad domains, little is known about how these issues manifest in top-tier machine learning venues. This thesis seeks to fill that void.

1.2 Problem Statement

Research Question: *How transparent are TPAMI's machine learning researchers about the data they use in applications?*

The aim of this research is to systematically assess how clearly and consistently annotation practices are reported in TPAMI publications that introduce or use datasets. In particular, the study investigates the extent to which papers disclose details about how labels were collected, who the annotators were, how they were selected and trained, whether they were compensated, and how the reliability of their labeling was measured.

To address the research question, we formulated several subquestions. First, we ask to what extent TPAMI papers provide metadata about annotator recruitment, training procedures, and compensation schemes. Second, we examine whether inter-rater reliability (IRR) is reported, and if so, whether authors specify the metric used to assess agreement. Third, we investigate whether there is a relationship between a paper's citation impact and the completeness of its annotation documentation. Fourth, we explore whether newer publications show improvement in transparency compared to older ones. Finally, we assess whether some types of datasets—such as those tied to public benchmarks or multi-institutional efforts—tend to demonstrate better reporting practices.

To answer these questions, we conducted a structured audit of 75 TPAMI papers published between 2009 and 2024. We examined 64 datasets mentioned in these papers and evaluated them against a 27-item checklist adapted from Geiger et al.'s "Garbage In, Garbage Out" framework. This audit quantifies the current state of annotation documentation and

highlights where and how transparency falls short. By doing so, the study offers evidence-based recommendations for improving the reliability and reproducibility of machine learning benchmarks through standardized, checklist-based reporting requirements. Additionally, we analyze correlations between fields in the checklist to identify patterns in reporting behavior—such as whether authors who report annotator compensation are also more likely to describe recruitment or training procedures.

2 Methodology

To systematically assess how TPAMI publications report annotation workflows, our methodology is divided into three phases discussed below. Data collection was done in Google Sheets and in this section and the following, by referring to "Tab X" we are talking about a specific sheet in our database. The description of these sheets is as follows:

1. **Article Selection (Tab 1)**: Selecting a representative sample of 75 articles across three time periods using a Scopus query.
2. **Dataset Compilation and Ranking (Tab 2)**: Identifying and ranking 214 unique datasets used within those articles.
3. **Annotation Metadata Extraction (Tab 3)**: Extracting structured metadata on annotation practices from the selected papers.

These steps form a reproducible framework for evaluating the completeness and transparency of dataset reporting in the machine learning literature. This division is inspired by the *Datasheets for Datasets* framework by Gebru et al. [32], which advocates for lifecycle transparency in dataset creation, labeling, and documentation.

2.1 Tab 1: Article Selection

We selected 75 TPAMI articles across three time periods, spanning the most recent 2, 5, and 15 years, with 25 papers drawn from each. Given the project's 10-week timeframe, analyzing a larger number of papers was deemed impractical. This temporal division allows us to evaluate both historical and recent reporting practices. The year 2024 marked the endpoint of data extraction, as 2025 publications were not yet fully indexed for some venues. We decided to sort the papers based on citation counts, as this should lead to analyzing the most influential datasets.

We used Scopus, as recommended by our supervisor, to simplify venue disambiguation. While several academic databases were available (e.g., Google Scholar, ArXiv), we selected Scopus as it is query-based and easily reproducible. Furthermore, Scopus allows us to discard the deduplication process for a specific time interval. On April 25, 2025, in order to extract data, we executed the queries that can be found in the appendix B.
This query returned only papers from TPAMI. Unlike my colleagues who had to filter for relevant venues, I only applied a publication year filter.

The results were exported to Google Sheets (hereafter, *Tab 1*). We filtered out:

- Survey papers or meta-studies that did not use datasets.

- Duplicate entries across the three time intervals.

In total, 16 survey papers and 5 duplicates were removed. The remaining 54 formed the basis of my dataset analysis.

2.2 Tab 2: Dataset Compilation and Ranking

From the final set of 54 articles, we extracted all datasets mentioned and compiled them into a new sheet, referred to as *Tab 2*. We identified 838 unique datasets across all publications and recorded the following metadata for each one of them:

- Dataset name
- Digital Object Identifier (DOI)
- Corresponding publication URL
- Notes

Each dataset was then ranked using a citation-weighted usage metric across time periods:

$$\text{Score}_{d,t} = \sum_{p \in P_{d,t}} \text{Citations}(p) \tag{1}$$

where:

- $\text{Score}_{d,t}$ is the usage score for dataset d during time period t.
- $P_{d,t}$ is the set of papers in t using dataset d.
- $\text{Citations}(p)$ is the number of citations for paper p.

The top 20 ranked datasets per period were compiled into a *Dataset Leaderboard*, which guided the selection of datasets for annotation review. The citation-weighted ranking method might favor older or more general-purpose datasets, and this potential skew must be acknowledged.

2.3 Tab 3: Annotation Metadata Extraction

In the final phase, we reviewed all selected datasets to extract structured metadata about annotation workflows. This information was documented in *Tab 3*, capturing three core categories:

- **Labellers**: What was the amount of annotators? Was the IRR calculated? How many labellers were there per item? Did the labellers receive any form of compensation?
- **Items**: What is the source of the annotated items? Was the amount of items decided before the annotation process? Why is this population of items chosen?
- **Annotation Schema**: Is there a reason for using a particular annotation schema? Was it decided beforehand?

These attributes were heavily inspired by Geiger et al's annotation schema for "Garbage in, Garbage Out Revisited" [34]. On top of their work, we introduced more attributes. The annotation schema was developed collaboratively with my colleagues and supervisor over a three-week period. Each dataset was annotated only once by one of our team members. All ambiguities were resolved through team discussions. The complete annotation schema we used can be found in the appendix A.
Finally, the data was extracted from *Tab 3* with the help of a script we wrote. This script is built in Python and has Analyzer modules that help us interpret the data we collected.

2.4 Analysis

To complement the metadata extraction process, we performed a structured quantitative analysis on the annotated fields from our 27-item checklist. This analysis comprised two main components: quantifying missing information and investigating consistency patterns across fields.

For the **missing information quantification**, each metadata field was assessed across all datasets to determine the proportion of entries that lacked usable information. Values were considered missing if they matched any of the following patterns: `No information`, `Unknown`, `Unsure`, or `nan`. Entries explicitly marked as `Not applicable` were excluded from both the numerator and denominator in our calculations, ensuring that only applicable cases contributed to each field's missingness rate. This approach allowed us to isolate the most neglected aspects of dataset documentation, such as annotator recruitment, compensation, and training procedures.

In the second part, the **field-pair consistency analysis**, we investigated whether transparency in one aspect of the annotation pipeline predicted transparency in another. This was done by computing pairwise associations between logically related fields— for instance, whether reporting compensation correlated with reporting annotator training, or whether datasets that mention inter-rater reliability (IRR) also specify the metric used. For each pair, we calculated *Cramer's V* to measure the strength of association between the two categorical variables. These pairings were selected based on conceptual dependencies within the annotation process and provide insight into patterns of co-reporting or co-omission.

Several metadata fields in the annotation schema involved open-ended or free-text responses, which required normalization before statistical analysis. To ensure compatibility with categorical association metrics such as Cramer's V, these responses were discretized into standard categories. A custom mapping script was used to classify answers into *Yes*, *No*, or *Not applicable*, based on keyword matching and pattern detection. For example, responses containing phrases such as "no information," "unknown," "not reported," or empty values (including `nan`) were uniformly mapped to *No*. Similarly, explicit denials of relevance (e.g., "not applicable," "N/A") were mapped to *Not applicable* and excluded from the correlation analysis. This discretization ensured internal consistency while preserving the semantic intent of the original answers.

All computations were carried out using a custom Python script. The script included modular `Analyzer` components for generating summary statistics, computing correlation coefficients, and visualizing reporting patterns. This analysis framework enabled both fine-grained

and aggregate evaluation of annotation transparency across the TPAMI dataset corpus.

3 Findings

In this section we will be explaining the quantitative results from this study. We will keep the explanations ordered, starting from broad observations towards more fine-grained ones. The analysis will be performed, based on the relevance of the analyzed data, either on unique datasets (52 in total) or based on the time periods we extracted data from, which contain 64 datasets (52 unique ones and overlap).

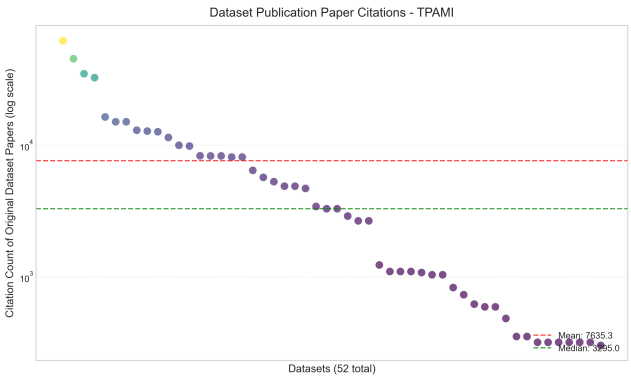

Figure 1: Citation distribution of TPAMI Papers

Figure 1 plots citation counts for the 52 dataset papers in Table 2.1. The smallest count is 302 and the largest is 62,201. On average, each paper has about 7,635 citations, with half of them below 3,295. The variation is high—a standard deviation of roughly 11,843—because a few top datasets pull the numbers upward. In fact, only 15 papers exceed twice the median (6,590 citations), yet those 15 account for over 60 percent of the total citations. Meanwhile, roughly 60 percent of the datasets cluster in the 800-8,000 range, forming a solid middle band beneath the handful of 'blockbuster' benchmarks.

3.1 Missing Information

Overall, 37.03% of the extracted fields were marked missing ("No Information"). When broken down by time period, the proportion of missing data is relatively similar—30.5% in the 2-year cohort, 40.6% in the 5-year cohort, and 39.6% in the 15-year cohort. It can be observed that in the last 2 years, the missing field rate is 10 percentage points lower. Additionally, missingness shows no meaningful relationship with citation impact: the Pearson[1] correlation coefficient is -0.1179, and the Spearman[2] rank correlation is

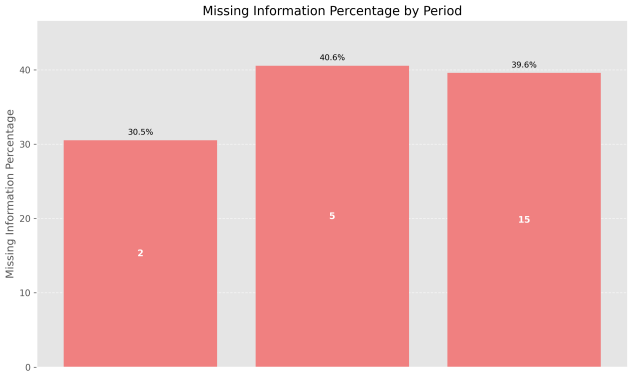

Figure 2: Missing Information per Period

[1] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient
[2] https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

-0.0632.

Across all 64 datasets, fields like Labeller Population Rationale (76.56% missing), Prescreening (73.44%), Total Labellers (68.75%), Compensation (67.19%), Training (62.50%), and Label Threshold (62.50%) are missing in well over three-fifths of cases. In other words, even though 86.54% of datasets report having human labels, almost none explain how those humans were selected, trained, compensated, or how reliability was measured. Only a small handful of datasets document a comprehensive annotation pipeline. This leads to two large problems. First, without transparency around how labels were produced and validated, it is difficult to interpret reported model performance, especially on models where subtle semantic distinctions or fine-grained categories are involved. Second, this barrier to reproducibility means that even if future researchers suspect a problem (e.g. bias in the labels), they would lack the baseline information needed to replicate or correct it. Until dataset creators adopt best practices for documenting annotator selection, instruction protocols, and reliability measures, the field will continue to rely on benchmarks whose foundational labels remain largely unverifiable.

3.2 Correlated Reporting Patterns

In this subsection, we explore whether, once a researcher documents one aspect of their annotation process, they tend to document the related aspects as well. This analysis explores whether transparency in one area of annotation documentation is predictive of transparency in others. Below we summarize the head-to-head findings for each pair of fields.

Overlap Synthesis vs Synthesis Description
Overlap synthesis refers to solving disagreements between multiple annotators for one item. When authors report any form of overlap synthesis (qualitative, quantitative, or other), they always describe the synthesis method (100% of cases). Conversely, if they omit overlap synthesis entirely, the synthesis description is also never provided. This perfect alignment (Cramer's V = 1.0) shows these two fields act as a single documentation unit.

Human Labels vs Original Labels
Among the 52 datasets with human-label information, 94.6% of fully human-labeled datasets declare their labels as "original" (collected by the dataset creators), while machine-labeled datasets (and those with missing human-label data) always default to "external" or "no information." The strong association (Cramer's V = 0.70) indicates that the decision to involve human annotators tightly predicts whether labels are primary (original) or secondary (external).
form **Compensation vs Annotator Training**
Of projects reporting monetary compensation, 55.6% also document that annotators received some formal training. In contrast, when compensation details are missing, 71.4% also omit training information. This moderate correlation (Cramer's V = 0.45) suggests that paying annotators is associated with the likelihood—but does not guarantee—that training procedures are described.

IRR vs Metric
Whenever inter-rater reliability (IRR) is reported, authors always specify the exact metric used (e.g., Cohen's K), and when IRR is omitted or marked not applicable, the metric field

is likewise blank. With a strong link (Cramer's V = 0.81), these two fields form an all-or-nothing pair for measuring annotation consistency.

Human Labels vs Total Labellers
Even among datasets fully labeled by humans, 67.9 % fail to specify how many annotators took part. Machine-labeled or undocumented human-label cases always coincide with "not applicable" or missing labeller counts. The strong association (Cramer's V = 0.72) highlights a common gap: authors note that humans labeled the data but frequently skip reporting the total number of annotators involved.

Looking at the missing information for these specific fields:

- **Labeller Population Rationale** is missing in 76.56% of papers, and **Prescreening** in 73.44%.
- **Total Labellers** is unreported in 68.75%, **Compensation** in 67.19%, and **Training** in 62.50%.
- Even when **IRR** is mentioned, the specific **Metric** is absent 56.25 % of the time.
- Mid-level fields such as **Overlap Synthesis** and **Synthesis Type** are skipped in 43.75% of papers.
- By contrast, only the most basic metadata (e.g. *Item Population*, *Item Source*, *Outcome*, *Human Labels*, *OG Labels*) achieve near-complete coverage (<5% missing).

These widespread omissions are serious violations of reproducibility rather than simply oversights. Any downstream evaluation, whether it be model correctness, bias assessment, or generalization claims, depends on an unexamined and ultimately unknown ground truth if three quarters of authors refuse to disclose the identities of their annotators, their training and compensation, or the methods used to measure consistency. TPAMI benchmarks will continue to spread ambiguity under the pretense of scientific advancement unless dataset creators treat annotation metadata with the same level of rigor as they do algorithmic descriptions and performance measures.

4 Discussion

Despite the overall gaps in annotation reporting, there is cause for optimism: a small but applaudable number of datasets used in TPAMI papers demonstrate that detailed documentation is achievable. In particular, datasets like ImageNet-Real [8], Cityscapes [18], SUN RGB-D [88], BSDS300 [74], and MPII [2]—each tied to formal benchmarks or multi-institutional efforts—consistently publish thorough annotation protocols, label-tool code, and inter-rater reliability procedures. Their authors put in the effort of having clear guidelines (e.g. how to handle occluded joints in MPII or ambiguous boundaries in BSDS300), making it straightforward for users to understand exactly how labels were produced.

These positive examples share three critical traits: alignment with public evaluation servers or challenges, strong institutional support (which often mandates transparency), and extensive supplementary material. Together, these factors ensure that every step of the labeling pipeline—from annotator selection and training to quality checks and compensation—is

clearly explained. As a result, these datasets achieved at least 75% completeness on our annotation-field checklist, in stark contrast to the broader collection where most fields remain blank.

However, it's important to recognize that even the best-reported papers prioritize benchmark performance over full transparency. While they excel at defining tasks, leaderboard metrics, and evaluation protocols, the underlying motivation remains to foster reproducible model comparisons rather than to provide a complete "datasheet" of every annotation decision. In other words, although some dataset papers do a good job on paper, what they have in common is a focus on benchmarks more than on transparency itself.

Building on the positive examples highlighted above, it is clear that detailed annotation reporting—while achievable—is still far from the norm. This gap underscores a critical truth: **transparency in machine learning is not optional, but foundational**. Without clear, consistent documentation of how data is labeled, we cannot reliably assess model performance, detect hidden biases or even reproduce experiments in new settings. In other words, the trust we place in "state-of-the-art" systems is only as strong as the transparency of their ground-truth foundations.

Transparent reporting of annotation practices begins with the very fundamentals of who created the labels and how reliable their judgments were. For example, knowing the total number of labellers is crucial. A small team may introduce biases, while a large, diverse pool promotes robustness. Striking that balance is impossible to gauge if 68.75% of papers simply omit this count. Likewise, reporting inter-rater reliability (IRR) scores without specifying the metric leaves readers unable to contextualize consistency: over half of the dataset papers report an IRR figure but not the metric used, basically stripping the statistic of meaning.

Equally important is documenting how many annotators labeled each item and the procedures for resolving disagreements. Multiple annotations per example foster confidence in the label, but the method of synthesizing overlap—be it majority vote or probabilistic aggregation—can systematically shift the dataset's character. Nearly 43.75% of papers fail to describe their overlap synthesis approach, preventing any meaningful reproduction or bias analysis.

Beyond these quantitative checks, transparency demands clarity on who those annotators were and how they worked. Details such as how the labellers were chosen, prescreening procedures, nature of training and instructions and even the compensation scheme directly influence label quality and ethical considerations. Yet 76.56% of papers do not justify their choice of annotators, 73.44% omit prescreening details, 62.5% ignore training protocols and 67.19% leave compensation unreported. This critical context, if absent, hides potential sources of systematic error and unfair labor practices.

The source and selection of items annotated are equally pivotal. Without knowing whether examples were drawn from public benchmarks, private collections or from the internet, one cannot asses dataset representativeness or generalization.

The lack of transparency around dataset annotation is not just a technical limitation—it raises important ethical and scientific concerns. When human annotation processes are

undocumented or poorly described, the underlying assumptions and labor behind machine learning benchmarks are hidden from scrutiny. This invisibility can obscure issues of bias, unfair labor practices, or methodological shortcuts that ultimately affect the fairness, reliability, and reproducibility of deployed models. Particularly in high-stakes domains such as medical imaging or surveillance, the absence of rigorous annotation protocols undermines public trust and can amplify harm to vulnerable populations.

In order to address these shortcomings, the machine learning community must standardize a reporting framework. By integrating a checklist of reporting on these fields into conference and journal submission requirements we can ensure consistent transparency across venues. Such a framework will not only improve reproducibility and comparability, but also foster accountability: researchers will know that every aspect of their annotation pipeline will be scrutinized and understood by reviewers and readers alike. Only through widespread adoption of a standardized reporting practice can we truly build machine learning systems whose performance and fairness rest on a solid, ethical foundation.

5 Responsible Research

In order to ensure the research made is reproducible, the dataset[3] and codebase [4] used to analyze the results is made public. For the data collection exact queries and date they were ran on are available in the appendix B.

One challenge that's difficult to fully control is the fact that citation counts can change over time. This might lead to small differences compared to when the data was originally gathered. Still, this shouldn't affect the overall results, since the most cited papers typically stay at the top regardless of slight changes.

The full methodology is explained in Section 2, where each step is described in enough detail to help others follow the same process. The annotation schema that was used to categorize the data is also included in Appendix A.

Large language models were used throughout the project to help with both paraphrasing and coding. GPT-4 was used to reword sections of text to make them clearer, while Claude 3.7 Sonnet was used to write parts of the analysis framework. Every piece of output, whether text or code, was carefully reviewed and often rewritten to make sure it was accurate and reliable.

6 Conclusions and Future Work

According to this audit, more than 37.03% of important annotation fields in TPAMI publications are still missing documentation. It is concerning that too many basic questions—like the overall number of labellers—go unanswered, indicating a widespread lack of precision in reporting.

[3] https://docs.google.com/spreadsheets/d/16MkuS-upEQxkAj-poZO5ggPqmu_UIDbwi7HWS3-21HE/edit?usp=sharing

[4] https://github.com/Gargant0373/DatasetAnalysis

Specifically, the bulk of papers lack fields like *Labeller Population Rationale* (76.56% missing), *Prescreening Procedures* (73.44%), and *Total Number of Labellers* (68.75%). Even techniques for guaranteeing label quality—*Training* (62.50% absent) and *IRR Metric* (56.25%)—are commonly left out, which erodes trust in the reliability of "ground truth."

That said, a minority of datasets demonstrate exemplary documentation, proving that comprehensive annotation metadata is achievable. These outliers illustrate best practices and can serve as templates for the wider community.

The *Datasheets for Datasets* framework [32], which recommends a comprehensive questionnaire encompassing annotator demographics, training processes, compensation schemes and inter-rater reliability measurements, is something I strongly endorse in order to close this transparency gap. For language data in particular, *Data Statements* [6] outlines a structured way to document collection circumstances, speaker demographics, and ethical considerations, thereby extending transparency all the way back to raw data acquisition. No important field would go unreported if such formal patterns were adopted, ideally enforced by journals and conferences.

Future work should expand this critique to other flagship venues (e.g. CVPR, ICCV) and continually update our dataset of documentation assessments. Moreover, routine community audits and public checklists could track progress and incentivize full disclosure. Ultimately, true machine learning excellence demands not only high performance metrics but also transparent, reproducible data foundations.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and Sabine SÃŒsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274â2282, November 2012.

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[4] P ArbelÃ¡ez, M Maire, C Fowlkes, and J Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898â916, May 2011.

[5] Pablo ArbelÃ¡ez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011.

[6] Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

[7] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798â1828, August 2013.

[8] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *CoRR*, abs/2006.07159, 2020.

[9] Lucas Beyer, Olivier J. HÃ©naff, Alexander Kolesnikov, Xiaohua Zhai, and AÃ€ron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. Introduces ImageNetâReaL, a reassessed labeling of ImageNet validation set for more reliable evaluation.

[10] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A highâdefinition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. Available online since AprilÂ 22,Â 2008.

[11] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1483â1498, May 2021.

[12] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2021.

[13] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172â186, January 2021.

[14] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834â848, April 2018.

[15] Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13467–13488, 2023.

[16] Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1â20, 2023.

[17] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. *arXiv preprint arXiv:1711.07846*, 2017. Introduces FMoW dataset with >1â¯M satellite images across 200+ countries and metadata annotations.

[18] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[19] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[20] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850â10869, September 2023.

[21] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005.

[22] Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1â1, 2021.

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[24] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743â761, April 2012.

[25] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295â307, February 2016.

[26] Frederik Ebert, Chelsea Finn, Alex X. Lee, and Sergey Levine. Selfâsupervised visual planning with temporal skip connections. *arXiv preprint arXiv:1710.05268*, 2017.

[27] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112â7127, October 2022.

[28] Mark Everingham, Luc Van Gool, Christopherâ¯K.â¯I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. Introduces VOC2007 dataset and benchmark evaluation procedures.

[29] J. Ferryman and A. Shahrokni. Pets2009: Dataset and challenge. In *2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 1–6, 2009.

[30] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154â180, January 2022.

[31] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652â662, February 2021.

[32] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, 2018.

[33] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[34] R. Stuart Geiger, Dominique Cope, Jamie Ip, Marsha Lotosh, Aayush Shah, Jenny Weng, and Rebekah Tang. âgarbage in, garbage outâ revisited: What do machine learning application papers report about human-labeled training data? *Quantitative Science Studies*, 2(3):795–827, 2021.

[35] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 325–336. Association for Computing Machinery, 2020.

[36] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Selfâsupervised structureâsensitive learning and a new benchmark for human parsing. *arXiv preprint arXiv:1703.05446*, 2017. Introduces the LIP dataset: 50,462 images annotated with 19 semantic human-part labels.

[37] Olav Engen Gundersen and Sverre Kjensmo. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 4787–4794, 2018.

[38] Meng-Hao Guo, Zheng-Ning Liu, Tai-Jiang Mu, and Shi-Min Hu. Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1â13, 2022.

[39] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4338â4364, December 2021.

[40] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):87â110, January 2023.

[41] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):386â397, February 2020.

[42] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, page 1956â1963. IEEE, June 2009.

[43] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2341–2353, 2011.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition*, page 346â361. Springer International Publishing, 2014.

[45] Joao F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583â596, March 2015.

[46] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5227â5244, August 2024.

[47] Timothy M Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J. Storkey. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1â1, 2021.

[48] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018.

[49] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562–1577, 2021.

[50] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562â1577, May 2021.

[51] Zilong Huang, Xinggang Wang, Yunchao Wei, Lichao Huang, Humphrey Shi, Wenyu Liu, and Thomas S. Huang. Ccnet: Criss-cross attention for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6896â6908, June 2023.

[52] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221â231, January 2013.

[53] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey, 2019.

[54] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4037â4058, November 2021.

[55] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1409â1422, July 2012.

[56] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160, 2015.

[57] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. Introduces CelebAâHQ dataset as highâquality version of CelebA (30â¯k images at 1024Ã1024).

[58] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013.

[59] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical Report TR-2009 TR-2009, University of Toronto, Department of Computer Science, Apr 8 2009. Introduces CIFARâ10 and CIFARâ100 datasets (60000 images, 10 classes).

[60] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563, 2011.

[61] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[62] Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, Apr 2022. Pictures of objects from 101 categories; collected in September 2003.

[63] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12581â12600, October 2023.

[64] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018.

[65] Zhizhong Li and Derek Hoiem. *Learning Without Forgetting*, page 614â629. Springer International Publishing, 2016.

[66] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292â3310, March 2023.

[67] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318â327, February 2020.

[68] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):978–994, 2011.

[69] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171â184, January 2013.

[70] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684â2701, October 2020.

[71] Canyi Lu, Jiashi Feng, Yudong Chen, Wei Liu, Zhouchen Lin, and Shuicheng Yan. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):925â938, April 2020.

[72] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fineâgrained visual classification of aircraft. arXiv preprint arXiv:1306.5151, arXiv, 2013. Introduces FGVCâAircraft dataset: 10â¯000 images covering 100 aircraft model variants in a hierarchical structure.

[73] Yu A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824â836, April 2020.

[74] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423 vol.2, 2001.

[75] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423 vol.2, 2001.

[76] Marc Masana, Xialei Liu, BartÅomiej Twardowski, Mikel Menta, Andrew D. Bagdanov, and Joost van de Weijer. Class-incremental learning: Survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513â5533, May 2023.

[77] Shervin Minaee, Yuri Y. Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1â1, 2021.

[78] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.

[79] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing*, pages 722–729, 2008.

[80] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420, 2009.

[81] Hossein Rahmani, Mohammed Bennamoun, and Qiuhong Ke. Human action recognition from various data modalities: A review. February 2021.

[82] Rene Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623â1637, March 2022.

[83] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137â1149, June 2017.

[84] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li FeiâFei. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*, 2014. Describes the 2012 ILSVRC benchmark with 1.2â¯M images across 1,000 classes.

[85] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1â14, 2022.

[86] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640â651, April 2017.

[87] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multiâclass object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009.

[88] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, 2015.

[89] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, 2015.

[90] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1474â1488, February 2023.

[91] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012.

[92] Gencer Sumbul, Marcela Charfuelan, BegÃŒm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904, 2019.

[93] Amirhossein Tavanaei. Embedded encoder-decoder in convolutional networks towards explainable ai, 2020.

[94] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, and HervÃ© JÃ©gou. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5314â5321, April 2023.

[95] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[96] Catherine Wah, Steve Branson, Peter Welinder, Takeshi Mita, Florian Schroff, Serge Belongie, and Pietro Perona. The caltechâucsd birdsâ200â2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, Jul 2011. Extended version of CUBâ200 with part annotations and 11,788 images.

[97] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349â3364, October 2021.

[98] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362â5383, August 2024.

[99] Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, Philip S. Yu, and Mingsheng Long. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2208â2225, February 2023.

[100] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2411–2418, 2013.

[101] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834â1848, September 2015.

[102] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 311–320, 2018.

[103] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502â518, January 2022.

[104] Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113â12132, October 2023.

[105] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5525–5533, 2016.

[106] Xue Yang, Junchi Yan, Wenlong Liao, Xiaokang Yang, Jin Tang, and Tao He. Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2384â2399, February 2023.

[107] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *2011 International Conference on Computer Vision*, pages 1331–1338, 2011.

[108] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872â2893, June 2022.

[109] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1â19, 2022.

[110] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for fast image restoration and enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1934â1948, February 2023.

[111] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *arXiv preprint arXiv:1702.05693*, 2017. Introduces high-quality pedestrian annotations on top of the Cityscapes dataset.

[112] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795â10816, September 2023.

[113] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.

[114] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452â1464, June 2018.

[115] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision*, 127(3):302–321, 2018.

[116] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1â20, 2022.

[117] Zhuangdi Zhu, Kaixiang Lin, Anil K. Jain, and Jiayu Zhou. Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13344â13362, November 2023.

[118] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1â1, 2022.

[119] Hasib Zunair, Shakib Khan, and A. Ben Hamza. Rsud20k: A dataset for road scene understanding in autonomous driving, 2024.

A Annotation Schema

The following schema was used to systematically extract and code annotation-related metadata from the selected TPAMI publications. Each attribute was categorized using standardized dropdown options, including *Yes*, *No*, *No information*, *Unsure*, and *Not applicable*, unless otherwise noted. This schema ensured consistent and reproducible assessments of dataset documentation quality.

Empty Whether any information is provided about the dataset in the paper.

Outcome The intended purpose or task associated with the dataset.

Human Labels Extent of human annotation coverage (e.g., all, some, none).

OG Labels Whether labels were created by the authors or taken from an external source.

Label Source Origin of the labels (e.g., MTurk, students, internal).

Prescreening Whether annotators were preselected based on general skills, platform performance, or project-specific criteria.

Compensation Type of compensation provided (e.g., money, authorship, volunteer).

Training Whether annotators received interactive, task-specific training.

Formal Instructions Whether annotators were given formal written or verbal instructions.

Labeller Population Rationale Whether the authors justified their choice of annotators.

Total Labellers Number of individuals who performed annotations.

Annotators per Item Reported number of annotators per item, if available.

Label Threshold Minimum number of labels required per item.

Overlap Whether multiple annotators labeled the same item.

Overlap Synthesis Whether annotation overlap was resolved via qualitative or quantitative methods.

Synthesis Type Specific technique used to resolve overlap (e.g., majority vote, discussion).

Discussion Whether disagreements were resolved through discussion.

IRR (Inter-Rater Reliability) Whether inter-rater agreement was reported.

IRR Metric Type of metric used if IRR was reported (e.g., Cohen's Kappa, F1 score).

Item Population Description of the items being annotated.

Item Population Rationale Reasoning for selecting that item population.

Item Source Where the items originated from.

A Priori Sample Size Whether the sample size was decided in advance.

Item Sample Size Rationale Justification for the size of the dataset collected.

A Priori Annotation Schema Whether the annotation schema was predefined or created ad hoc.

Annotation Schema Rationale Whether justification was provided for the schema design.

Link to Dataset Available Whether the dataset was accessible through the paper or supplementary material.

B Scopus Query

Queries were ran on April 25th, 2025.

15 Year Period Queries

```
SRCTITLE ( ieee AND transactions
           AND on AND pattern
           AND analysis AND machine
           AND intelligence )
AND PUBYEAR > 2009 AND PUBYEAR < 2025
```

5 Year Period Query

```
SRCTITLE ( ieee AND transactions
           AND on AND pattern
           AND analysis AND machine
           AND intelligence )
AND PUBYEAR > 2019 AND PUBYEAR < 2025
```

2 Year Period Query

```
SRCTITLE ( ieee AND transactions
           AND on AND pattern
           AND analysis AND machine
           AND intelligence )
AND PUBYEAR > 2022 AND PUBYEAR < 2025
```

C Statistical Tools

Cramer's V

$$V = \sqrt{\frac{\chi^2}{n \cdot (k-1)}} \tag{2}$$

where χ^2 is the chi-square statistic, n is the total number of observations, and $k = \min(r, c)$ is the smaller number of rows or columns in the contingency table.

Pearson's r

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3}$$

where x_i and y_i are the individual values of the two variables, $\bar{x}$ and $\bar{y}$ are their respective means, and n is the number of paired observations.

Spearman's ρ

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{4}$$

where d_i is the difference between the ranks of corresponding values x_i and y_i, and n is the number of observations.

D Datasets Analyzed

These are the datasets that were used to perform the final analysis.

- ADE20K [115]
- BAIR [26]
- BigEarthNet-S2 [92]
- BSDS300 [75]
- BSDS500 [5]
- Caltech101 [62]
- CamVid [10]
- CelebA-HQ [57]
- CityPersons [111]
- Cityscapes [19]
- CIFAR-10 [59]
- CIFAR-100 [59]
- CUB-200-2011 Birds [96]
- Dark Channel [43]
- FGVCAircraft [72]
- FMoW-S2 [17]
- Foot Keypoint [12]
- Got-10k [49]
- HMDB51 [60]
- ICDAR2015-Challenge-4 [56]
- ImageNet [23]
- ImageNet 2012 [84]
- ImageNet-1k [84]
- ImageNet-Real [9]
- INRIA Person [21]
- iNaturalist [95]
- KITTI [33]
- LIP [36]
- MegaDepth [64]
- MIT Indoor Scenes [80]
- MNIST [61]
- MPII [3]
- MSRC 21 [87]
- NYUDv2 [**?**]
- Online Object Tracking [100]
- Oxford Flowers 102 [79]
- Pascal Context [78]
- Pascal VOC 2007 [28]
- Pascal VOC 2010 [28]
- Pascal VOC 2012 [53]
- PETS 2009 [29]
- Places [113]
- RedWeb [102]
- RoadScene [119]
- SIFT Flow [68]
- Soda-A [15]
- Soda-D [15]
- Stanford 40 Actions [107]
- StanfordCars [58]
- Sun RGB-D [89]
- Tiny ImageNet [93]
- UCF101 [91]
- WiderFace [105]

E TPAMI Papers Analyzed

Paper title	Cited by
Transfer Learning in Deep Reinforcement Learning: A Survey [117]	309
UniFormer: Unifying Convolution and Self-Attention for Visual Recognition [63]	238
Towards Large-Scale Small Object Detection: Survey and Benchmarks [16]	293
KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D [66]	218
SCRDet++: Detecting Small, Cluttered and Rotated Objects via Instance-Level Feature Denoising and Rotation Loss Smoothing [106]	216
Human Action Recognition From Various Data Modalities: A Review [81]	355
A Survey on Vision Transformer [40]	2135
Image Super-Resolution via Iterative Refinement [85]	831
Learning Enriched Features for Fast Image Restoration and Enhancement [110]	254
ResMLP: Feedforward Networks for Image Classification with Data-Efficient Training [94]	302
Deep Long-Tailed Learning: A Survey [112]	275
A Comprehensive Survey of Continual Learning: Theory, Method and Application [98]	212
Domain Generalization: A Survey [116]	480
Real-Time Scene Text Detection with Differentiable Binarization and Adaptive Scale Fusion [66]	268
Constructing Stronger and Faster Baselines for Skeleton-Based Action Recognition [90]	267
Beyond Self-Attention: External Attention Using Two Linear Layers for Visual Tasks [38]	314
SpectralGPT: Spectral Remote Sensing Foundation Model [46]	353
Explainability in Graph Neural Networks: A Taxonomic Survey [109]	278
Diffusion Models in Vision: A Survey [20]	717
PredRNN: A Recurrent Neural Network for Spatiotemporal Predictive Learning [99]	287
Multimodal Learning With Transformers: A Survey [104]	361
CCNet: Criss-Cross Attention for Semantic Segmentation [51]	257
Salient Object Detection via Integrity Learning [118]	242
Class-Incremental Learning: Survey and Performance Evaluation on Image Classification [76]	319
Contextual Transformer Networks for Visual Recognition [63]	406
Image Segmentation Using Deep Learning: A Survey [77]	2641
Cascade R-CNN: High quality object detection and instance segmentation [11]	1100
Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs [73]	784
Self-Supervised Visual Feature Learning with Deep Neural Networks: A Survey [54]	1102
Meta-Learning in Neural Networks: A Survey [47]	1079
U2Fusion: A Unified Unsupervised Image Fusion Network [103]	1233
Focal Loss for Dense Object Detection [67]	4877

Got-10k: A large high-diversity benchmark for generic object tracking in the wild [50]	1080
Event-Based Vision: A Survey [30]	1238
Deep High-Resolution Representation Learning for Visual Recognition [97]	2899
Deep Learning for 3D Point Clouds: A Survey [39]	1341
Res2Net: A New Multi-Scale Backbone Architecture [31]	2319
Deep Learning for Person Re-Identification: A Survey and Outlook [108]	1317
Mask R-CNN [41]	2628
ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning [27]	809
Deep Learning for Image Super-Resolution: A Survey [97]	1097
NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding [70]	1109
Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer [82]	825
Squeeze-and-Excitation Networks [48]	4886
A Continual Learning Survey: Defying Forgetting in Classification Tasks [22]	1039
Tensor Robust Principal Component Analysis with a New Tensor Nuclear Norm [71]	750
OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields [13]	2666
Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition [44]	9854
Fully Convolutional Networks for Semantic Segmentation [86]	8132
High-speed tracking with kernelized correlation filters [45]	5702
Image Super-Resolution Using Deep Convolutional Networks [25]	8040
DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs [14]	16385
Contour detection and hierarchical image segmentation [4]	4698
Object tracking benchmark [101]	3295
Representation learning: A review and new perspectives [7]	9997
SLIC superpixels compared to state-of-the-art superpixel methods [1]	8295
Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [83]	26841
Pedestrian detection: An evaluation of the state of the art [24]	2793
Learning without Forgetting [65]	2803
3D Convolutional neural networks for human action recognition [52]	5290
Single image haze removal using dark channel prior [42]	6443
Tracking-learning-detection [55]	3262
Robust recovery of subspace structures by low-rank representation [69]	3187
Places: A 10 Million Image Database for Scene Recognition [114]	2656

Table 1: TPAMI Papers and Their Citation Counts (Scopus)