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

## A deep dive into annotation transparency and consistency in CVPR corpus

**Claudia Scorția**[1]

**Supervisor(s): Dr. Cynthia Liem**[1]**, Andrew M. Demetriou**[1]

[1]**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: Claudia Scorția
Final project course: CSE3000 Research Project
Thesis committee: Dr. Cynthia Liem, Andrew M. Demetriou, Jie Yang

## Abstract

This study investigates annotation and reporting practices in machine learning (ML) research, focusing on societally impactful applications presented at the IEEE/CVF Computer Vision and Pattern Recognition (CVPR) conferences. By structurally analyzing the 75 most-cited CVPR papers from the past 2, 5, and 15 years, we evaluate how the human annotations foundation of supervised ML is documented. We introduce a 27-field annotation-reporting schema and apply it to 60 datasets, revealing that nearly 30% of relevant information is routinely omitted. Key findings include the pervasive underreporting of annotator details such as training, prescreening, and inter-rater reliability (IRR) metrics. While popular datasets like COCO and ImageNet exhibit widespread use, transparency about annotation methodologies remains inconsistent. The impact of a few fields shows that basic metadata, such as the selection process of annotators and how the labels' overlap is managed, strongly anticipate overall documentation quality. Our findings support previous calls for standardization and underscore the need for institutionalized reporting practices to ensure reproducibility, fairness, and trust in ML systems.


# 1 Introduction

## 1.1 Background and Motivation

Machine Learning (ML) has become the backbone of a vast and ever-growing array of applications, from medical diagnosis to emotion detection and beyond. Central to the success of these systems is the quality of their "ground truth," the human-provided annotations or labels used during training and evaluation. Without rigorous annotation practices, ML models risk perpetuating biases, producing unreliable predictions, or failing to generalize across different contexts. However, despite the critical role of annotated data, reporting on how such data are collected, verified, and maintained often remains superficial or altogether absent.

The importance of transparent annotation practices has been highlighted in recent work by Geiger et al.[1], who examined papers from multiple domains and found that many do not describe where or how human-labeled training data was obtained. The study was a follow-up of the study focusing on Twitter classification tasks [2]. Both revealed a troubling lack of standardization and accountability in annotation workflows.

From another point of view, Aroyo and Welty [3] challenge the foundational assumptions in annotation and argue that disagreement among annotators is not merely noise but can be a valuable signal reflecting ambiguity in data and task design. Their critique exposes the myth of a singular ground truth. Similarly to this, another study[4] investigates how methodological failures in machine learning and psychology lead to systemic errors, highlighting the possibility that training pipelines and decision systems may misread noisy human labels. This comparative perspective emphasizes the necessity of strong annotation methods and error modeling. Finally, Jacobs and Wallach [5] highlight how unexamined measurement assumptions, including those related to labeling, can entrench inequities in ML outcomes. Collectively, these studies underscore that annotation is not just a technical preprocessing step but a central knowledge and ethical concern in ML research.

## 1.2 Problem Statement and Research Questions

This gap in reporting raises pressing questions about the reproducibility and trustworthiness of ML systems deployed in societally impactful domains. In this paper, I address the central question:

> **What are the data collection and reporting practices of human annotations/labels in societally impactful ML research?**

To make this inquiry manageable, I focus on publications in IEEE/CVF Computer Vision and Pattern Recognition (CVPR)[1], one of the highest cited ML venues according to Google Scholar metric[2] with an h5-index[3] of 440. Specifically, this paper aims to answer the following 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?

By critically analyzing the papers, my aim is to measure the frequency of comprehensive annotation reporting, pinpoint widespread methodological flaws, and collect best-practice suggestions for further study. The methodology and the description of the problem in Section 2 follow the introduction of the report. Afterwards, Section 3 introduces the findings, followed by the discussion in Section 4. Section 5 presents the reflection on the ethical aspects of this research paper. In the end, Section 6 will present a discussion covering future work and recommendations in this field.

# 2 Methodology

This study represents a structured analysis, and this section outlines the steps taken to analyze CVPR

[1] https://cvpr.thecvf.com/

[2] `https://scholar.google.com/citations?view_op=top_venues&hl=ro&vq=eng`

[3] https://en.wikipedia.org/wiki/H-index

papers and the associated dataset papers. Subsection 2.1 describes the paper selection process. Subsection 2.2 details the analysis of the selected papers, and subsection 2.3 presents the analysis of the dataset. All data discussed are accessible through the following spreadsheet: Research Project Papers[4]. Each of the subsections is correlated to one important *Tab* in the spreadsheet. The document also serves as a central repository for data from all contributors to the project, including information from multiple venues.

### 2.1 Tab 1: Selecting important CVPR papers

Due to time constraints, this study focuses on a subset of papers from the IEEE/CVF Computer Vision and Pattern Recognition (CVPR) conference. To prioritize papers with societal impact, we select the top 25 most cited papers of the past 2, 5, and 15 years. All the papers analyzed are listed in Appendix D. Year 2024 was considered the top range, as 2025 has not come to an end yet. Within the study, we acknowledge that the high citation count does not necessarily correlate with the paper's quality. However, we consider it an indicator of a societally impactful paper, which is the focus of the study.

To decrease the time spent on retrieving the CVPR papers, to ensure consistency over the different venues analyzed by all teammates, and to eliminate the beginning steps of the disambiguation, the supervisor suggested using Scopus[5] for this study. Data were retrieved from Scopus on April 24, 2025, using the available citation rankings at that time. The queries used are described in Appendix A. Each paper is exported with its title, year of publication, number of citations, DOI, and link. Within the information table, this information can be observed in *Tab 1.* In addition to the extracted information, there are a few other columns that serve to track the workload. Few papers appear in multiple sections as they are well cited.

### 2.2 Tab 2: Analyzing CVPR papers

We analyze the selected CVPR papers to examine their dataset usage. Key details are recorded in *Tab 2* of the shared spreadsheet. For each paper, all datasets used are identified, including citation information for the original dataset publication. To ease the process of dataset collection, all information was gathered together, including also the ones from other venues. The same datasets are used across numerous domains despite the differences in the areas, and this centralization facilitates a faster procedure. Annotations include observations about dataset miscitation (e.g., missing or incorrect citations), non-reproducible experiments (e.g., use of private datasets), and other relevant issues.

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

[5]https://www.elsevier.com/products/scopus

### 2.3 Tab 3: Analyzing dataset papers

After gathering all datasets used within the venue, reaching a total of 165 distinct datasets, the papers are analyzed regarding their annotation procedure. To focus on the most impactful datasets within the limited timeframe, a weighted scoring formula is applied, and we selected the 20 papers with the highest score for each time frame. The selected datasets can be found in Appendix C. Due to the overlap between periods, in the end, 45 unique datasets were identified. For each dataset and each time period, a score is computed as:

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

where:

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

This scoring prioritizes datasets based on their citation impact within each time period. Each dataset paper is thoroughly analyzed, and multiple details regarding the annotation process are reported in *Tab 3.* The analysis of each dataset paper used the criteria already presented in Geiger's paper [1], and additional questions were built on by the team as important for transparency. Taking into consideration the study[1], the team decided that the focus of dataset annotation is split between three parts: items (the labels used in annotating the data), the annotators (who annotated the dataset), and the annotation practices (what was the annotation scheme). This data focuses on three aspects:

- **items:** In this aspect, we characterize each label by its outcome; the exact number of annotations collected per item; whether labels originated from external repositories, were produced by human annotators, or were algorithmically generated; the procedures for resolving conflicting annotations (such as majority vote, expert adjudication, or consensus discussion).
- **annotators:** Here we capture the human element of the annotation process, detailing the characteristics and management of the annotator pool: the background of annotators (domain experts, crowd-workers, or volunteers); the recruitment and prescreening methods employed (platform selection, eligibility criteria); training or qualification protocols (interactive tutorials, written guidelines, or qualification tests); compensation models (monetary payment, or unpaid contributions); any quality control (Inter-Rater Reliability[7] or any discussion); and any reported demographic or experiential information that might influence annotation behavior.

[7]https://en.wikipedia.org/wiki/Inter-rater_reliability

| Dataset (15y) | Citations (15y)[6] | Dataset (5y) | Citations (5y) | Dataset (2y) | Citations (2y) |
|---|---|---|---|---|---|
| Pascal VOC 2012[6] | 300480 | COCO[7] | 29582 | COCO[7] | 4656 |
| Pascal VOC 2007[8] | 259341 | ImageNet[9] | 13943 | ImageNet-1K[10] | 3941 |
| COCO[7] | 256849 | ADE20K[11] | 13143 | ImageNet[9] | 3302 |
| ImageNet 2012[10] | 256600 | ImageNet 2012[10] | 11077 | ADE20K[11] | 2105 |
| CIFAR-10[12] | 178632 | ImageNet-1K[10] | 8920 | DreamBooth[13] | 1565 |
| ImageNet[9] | 97650 | Cityscapes[14] | 8636 | LAION-400M[15] | 1551 |
| Pascal VOC 2011[16] | 51814 | Pascal VOC 2007[8] | 7041 | Objects365[17] | 1422 |
| ILSVRC 2014[10] | 48888 | Pascal VOC 2012[6] | 7041 | ImageNet-R[18] | 1321 |
| People-Art[19] | 38749 | CIFAR-10[12] | 6969 | CIFAR-100[12] | 1300 |
| Picasso[20] | 38749 | FFHQ[21] | 5784 | UCF101[22] | 1098 |

Table 1: Top 10 Datasets by Citations in Different Periods

- **annotation practices:** This covers the annotation schema mentioned by the datasets. If there is any reasoning behind it, or if there is no information.

These annotations are further explained in Appendix B.

Following the structured annotation of each dataset paper, we conducted a focused analysis across the three key dimensions: items, annotators, and annotation practices. This step aimed to identify common trends, highlight gaps in transparency, and assess the consistency of reporting practices. The insights drawn from this analysis form the basis of the findings presented in the next section.

## 3 Findings

This section presents the core quantitative findings of our analysis of CVPR papers. Each subsection explores a distinct aspect of the dataset, ranging from citation patterns to dataset usage and metadata completeness. Subsection 3.1 analyzes citation count distributions and highlights influential papers in CVPR publications. Subsection 3.2 examines how the usage of the dataset has evolved. The results from Subsection 2.3 are carefully analyzed from different points of view in the rest of the subsections, aiming to answer all the sub-questions of the research. All plots and statistics were performed with the use of the code available on GitHub[8]

### 3.1 Citation Patterns

According to Google Scholar, CVPR is the most highly cited venue in Engineering & Computer Science. Consequently, CVPR publications exert an outsized influence on the field's direction and priorities. As noted in Figure 1, He et al.[23] stand out, with their Deep Residual Learning for Image Recognition paper amassing 178,632 citations. In second place is Szegedy et al. [24], whose Going Deeper with Convolutions has 40,152 citations.

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

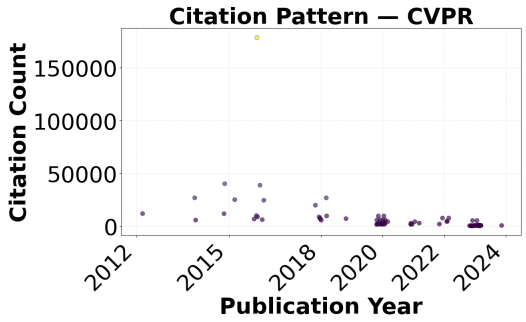


Figure 1: CVPR Papers: Citations by year

A major paradigm shift is signaled by the paper introducing Deep Learning [23]. It is one of the most cited papers in Machine Learning, and it presents a outstanding improvement in the performance of neural networks trained on ImageNet[9].

### 3.2 Temporal Trends in Dataset Usage

Across our 75 CVPR papers, Pascal VOC 2012[25] emerges as the single most societally impactful dataset (Table 1). Several datasets are directly based on Pascal VOC[9] or ImageNet[10], highlighting the weighty influence these datasets have in CVPR research. These two essential datasets have been iteratively extended over the years to meet changing research demands.

When analyzing the split over time and taking into consideration the citation count obtained with the formula, it seems like the community has changed from using datasets PASCAL VOC (2012[25] & 2007[8]) to the COCO[7] dataset. During the 15-year frame, PASCAL VOC datasets are in the top with 300K citations, and respectively 259K citations, COCO gained 256K citations. In the more recent 5-year period, COCO represents the most cited one with 29K citations, and both PASCAL VOC 2007 and 2012 have 7K citations.

[8]The citation number of a dataset represents the sum of papers citations using the specific dataset as explained in Subsection 2.3.

[9]http://host.robots.ox.ac.uk/pascal/VOC/

[10]https://image-net.org/index.php

For the last 2-year period, the PASCAL VOC datasets do not even reach the top 20 most used datasets, while COCO obtains a high score of 4K, even with the limitation of papers posted recently. This represents a notable improvement as Pascal VOC 2007 and 2012 presented 46.43% and 39.29% missing information, while COCO reached a score of only 18.52%.

Similar to COCO, the original dataset ImageNet[9] and its subset ImageNet-1K[10] gain importance with the more recent papers, starting on the 6$^{th}$, respectively 16$^{th}$ place within of the 15 years, and reaching the 3$^{rd}$, and 2$^{nd}$ place. However, the datasets are not presenting a crucial improvement, as they report 35.71% and 33.33% missing information.

### 3.3 Documentation Completeness Over Time

Across all datasets selected, 29.6% of the annotation fields were left undocumented. In order to calculate the percentage of missing information, for each period, all 20 datasets were taken into consideration, and the final score represents the total number of missing values over the total number of fields. For each dataset, the number of total fields might be different, as fields marked as "Not applicable" are not taken into account. Figure 2 tracks how the share of missing fields evolves. Each period presents the datasets used by the CVPR papers within the period. After reaching 33.7% missing in the 15-year slice, to a low of 29.3% in the middle (5-year) slice, and finally decreasing to 26.2% missing in the most recent two years. In other words, an improvement is noticed in the CVPR papers; however, not a remarkable one.

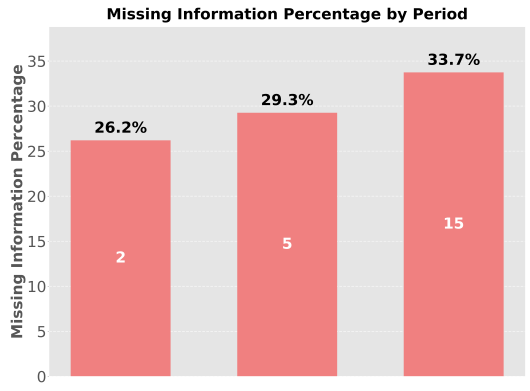


Figure 2: Missing documentation rates by publication period.

Examining the 27 annotation fields, Figure 3 highlights the weakest links of the annotating process. The ten least-documented elements 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%) - presenting information about the annotators
- **Sample size** (establishing from the beginning how many items the dataset should contain) (54.10%) and its **rationale** (45.90%) and the **Label Threshold** (39.34%) - presenting information about the labels

Overall, annotator-related details are the least frequently reported, representing a troubling omission.

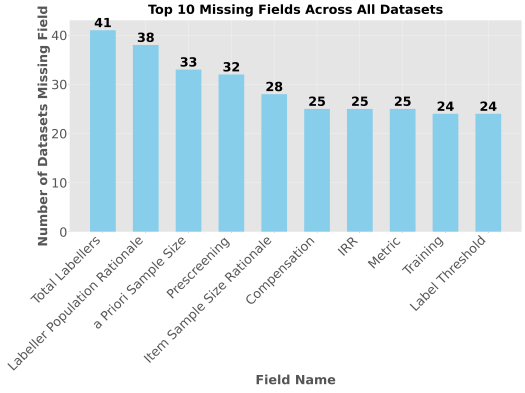


Figure 3: Missing fields over all periods.

Similar to the formulas applied to the results in Figure 2, to identify the percentage of missing information per field, "Not applicable" is completely excluded from the equation for each area.

### 3.4 Field Impact Analysis

To understand which individual fields drive overall documentation quality, I analyzed the impact of the presence of each field on the average missing-information rate across all 27 annotation fields. Figures 4 and 5 present two complementary views:

- **Fields with the lowest missing rates when documented (Fig. 4).** For each field, I collected only the datasets that presented information on it and then collected the overall percentage of missing information (e.g: IRR: 10 datasets documented IRR. Within these datasets, almost 30% of the annotation process information is missing.). The ten fields whose simple presence is associated with the most complete metadata include *Prescreening of Annotators*, *Annotators per item*, *Rationale behind the Item Sample Size*, *Compensation*, and *Rationale behind Annotators Selection*, each exhibiting missing-information rates below roughly 30%. This suggests that simply reporting how the annotators were prescreened and how many were assigned per item might strongly imply richer overall documentation.

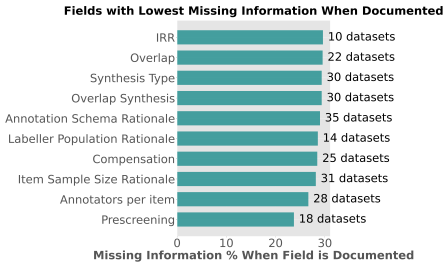


Figure 4: Lowest-missing fields

- **Fields with the greatest impact on completeness (Fig. 5).** For this analysis, I selected the datasets reporting a specific detail and the datasets that omit any information. In addition, the missing information of the two datasets was calculated in general and the difference was analyzed. When comparing datasets that do versus do not document a given field, five stand out by how much they reduce the overall missing-information rate:
  1. *Overlap*: 33.0 pp reduction[11]
  2. *Formal Instructions*: 28.2 pp reduction
  3. *Annotation Scheme established from the beginning*: 19.51 pp reduction
  4. *Discussion between annotators*: 2.18 pp reduction
  5. *Size of the dataset established from the beginning*: 0.87 pp reduction

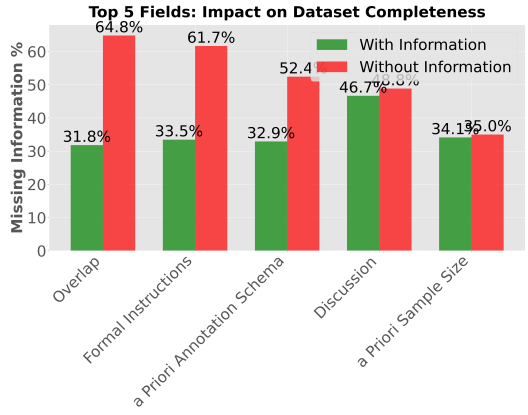


Figure 5: Top-impact fields

Taken together, these results indicate that two classes of metadata are most essential for driving completeness: (i) *Preparing the annotators* (prescreening and offering formal instructions), and (ii) *Solving the overlap in annotating*. Mandating even a small core of these fields could therefore dramatically raise the floor of metadata quality across CVPR datasets.

[11]Here "pp" means percentage points, i.e. the absolute difference between the two missing-information percentages.

### 3.5 Label Source and IRR Metric Usage

To gain further insight into how annotations are collected and assessed, we analyzed two fields across all datasets: the source of the human annotators and the metrics used to assess annotation quality (inter-rater reliability, IRR).

Out of 34 datasets with specified label sources:

- 44.1% used Amazon Mechanical Turk (MTurk).
- 14.7% used university students.
- 14.7 % involved experts.
- 11.7% reused annotations from existing datasets (e.g., ImageNet[9], CelebA[26]).
- 5.88% were labeled directly by authors.
- 8.82% fell into other sources (e.g., single annotator using LabelMe).

On IRR reporting, only 10 out of 45 datasets specified any metric for annotation agreement. The breakdown is as follows:

- Agreement-based metrics (e.g., human-level accuracy, pixel agreement): 4 datasets.
- F1-score, precision, or classifier performance: 1 dataset each.
- Statistical measures (e.g., standard deviation, GCC, LCE): 3 datasets.

## 4 Discussion

This section interprets our core results and situates them within the broader literature on dataset documentation. Subsection 4.1 unpacks the most striking patterns, such as the rising trend of missing annotator metadata, and ties them back to each research question. In Subsection 4.2, we compare our quantitative findings to prior reviews (e.g. Geiger et al.[1], Hullman et al.[4]) to highlight where our work confirms or departs from earlier observations. Subsection 4.3 translates these insights into concrete recommendations for ML practitioners-ranging from reporting standards to psychometric best practices. We then acknowledge in Subsection 4.4 the constraints of our study (e.g. venue bias, reliance on citation counts) and their impact on generalizability.

### 4.1 Answering the Research Questions

This section directly answers the three research subquestions introduced in Section 1, drawing on the key findings from our dataset analysis. Each subquestion is addressed in turn, with evidence drawn from the corresponding metrics and trends observed in Section 3.

**RQ1: What reporting elements count as transparent dataset documentation?** The field impact analysis (Section 3.4) showed that certain metadata fields are strong predictors of overall documentation completeness. For instance, fields such as *Prescreening of Annotators*, *Annotation Schema Rationale*, and *Label Source* consistently appeared in datasets with lower missing-information rates (under 30%).

- **Prescreening as a Predictor of Thoroughness.** Datasets that documented prescreening practices reported a small amount of missing information on our metadata checklist items. This suggests that thorough vetting of annotators correlates with a broader commitment to transparency across the annotation process.
- **Overlap and Formal Instructions** In datasets where annotators received detailed, formal guidelines and label overlaps were addressed through well-defined procedures, the annotation process was documented far more comprehensively. Because formal instructions and overlap tracking necessitate meticulous documentation of each annotation option, projects that explicitly defined these fields consistently generated more extensive information.

Furthermore, the analysis in Subsection 3.3 revealed that dataset papers overwhelmingly omit information about their annotators. Without knowing who did the labeling, how they were trained, and how the quality of their work was examined, it is impossible to assess annotation trustworthiness or reproduce the dataset.

**RQ2: How are human annotations collected in these papers?** I found substantial variation in the reported sources of annotations across datasets (Section 3.5). Out of 34 datasets with label source information, 15 datasets (44%) used **Amazon Mechanical Turk (MTurk)**. The high use of Amazon Mechanical Turk[12] for recruiting annotators points out once more the importance of prescreening and training the people hired for annotating a new dataset. Since they might not be experts and familiar with the topic of the dataset, without careful steps for selecting the annotators, the quality of the new dataset might be affected.

**RQ3: How is annotation quality assessed?** Only 10 of the 45 dataset papers reported using any metric for inter-rater reliability (IRR). Among those that did, the vast majority (4 datasets) used agreement-based metrics (e.g., pixel-level agreement, human-level accuracy). An interesting observation is that no dataset used standard psychometric measures like Cohen's Kappa[13], Fleiss' Kappa[14], or Krippendorff's Alpha [15]. This absence of consistent quality metrics highlights a crucial gap in how datasets are evaluated and undermines trust in label accuracy.

Taken together, these findings lead to a simple but powerful conclusion: attention to the human element in dataset construction, who labeled the data, how they were selected and trained, and how their output was evaluated-consistently anticipates better transparency and completeness. Formalizing even one part of the annotation workflow (e.g., sourcing or quality control) is strongly associated with stronger overall documentation.

[12]https://www.mturk.com/

[13]https://en.wikipedia.org/wiki/Cohen%27s_kappa

[14]https://en.wikipedia.org/wiki/Fleiss%27_kappa

[15]https://en.wikipedia.org/wiki/Krippendorff%27s_alpha

## 4.2 Comparison with Prior Work

Our finding that annotator-related details (e.g., total number of labellers, prescreening methods, compensation, training, and IRR measures) remain among the most under-reported elements echoes and extends observations from previous reviews. Geiger et al.[1] examined papers from multiple domains and documented a pervasive "black box" around how human labels were obtained, noting that few studies described recruitment, training, or quality-control procedures in any depth. Similarly, we observe that only 32.79% of datasets reported the total number of labellers, 47.34% mentioned prescreening, and a mere 59.02% reported IRR metrics, indicating that this transparency gap persists in high-impact CVPR publications. Whereas Geiger et al.[1] focused on Social Sciences & Humanity, Life & Biomedical Sciences, Physical & Environmental Sciences tasks, our analysis generalizes the concern across a broad area of computer-vision benchmarks, confirming that the "garbage in, garbage out" risk remains acute in vision research.

Multiple "myths" mentioned in the paper regarding misbeliefs in human annotation[3] are present in the analyzed dataset papers, for example, authors opting for one annotator per item (*One is enough*). Engstrom et al.[27] introduce ImageNet-V2 to agree that *"once done, always valid"* is a myth, demonstrating that visual perspectives shift over time: for example, the kinds of images labelled as "telephone" or "automobile" thirty years ago look markedly different from those today. Moreover, the multiple datasets versions of the Pascal VOC and ImageNet datasets were created due to this continuous evolution.

## 4.3 Implications for ML Practice

The persistent omission of key annotator details in dataset publications points to a clear necessity for a unified, end-to-end reporting framework. By selecting a single established template and applying it consistently, authors can ensure that every dataset release includes comprehensive information on annotator demographics, qualification criteria, compensation, training procedures, inter-rater reliability measures, and reconciliation workflows. Gebru et al. [28] and Pushkarna et al. [29] offer two practical templates, like Datasheets for Datasets and Data Cards, and could present a solution for this problem.

To make this approach effective, reporting must be embedded directly into existing publication and hosting workflows: conferences and journals could require submission of a completed annotation report alongside any dataset paper, while dataset repositories and versioning platforms should support uploading and rendering these reports so that metadata is immediately visible to consumers. Annotation tools and data-management systems also have a crucial role to play

by automating the capture of metadata, such as annotator identifiers, session logs, qualification test results, payment records, and IRR calculations, and exporting it in the chosen template format with minimal manual effort. Institutionalizing these practices will bolster reproducibility, enable more informed dataset selection, and ultimately enhance the fairness, and trustworthiness of machine learning models.

### 4.4 Limitations and Threats to Validity

While the structured analysis yields novel insights into annotation reporting in CVPR dataset papers, several threats to validity must be acknowledged.

**Selection Bias.** The sample was restricted to the top 25 most-cited CVPR papers from three time windows (2, 5, and 15 years), using citation count as a proxy for societal impact. This approach may over-represent landmark benchmarks (e.g., Pascal VOC[6], ImageNet[9], COCO[7]) and under-represent emerging or niche domains, limiting the generalizability of our findings to less-cited but potentially innovative datasets.

**Construct and Measurement Validity.** Our coding schema-comprising 27 annotation metadata fields-relies on operational definitions (e.g., "training" vs. "formal instructions," "overlap synthesis" categories) that may not capture the full nuance of authors' descriptions. Ambiguities in terminology forced us to infer "no information" in borderline cases, potentially conflating genuine omissions with reporting variance.

**Coding Reliability.** All dataset-paper analyses were performed by a small team using a shared spreadsheet and dropdown categories; we did not calculate inter-coder agreement before full extraction. Consequently, individual coder biases or misunderstandings may have influenced our binary "documented/ not documented" judgments, particularly for partially described fields.

**Conclusion Validity.** The field impact analysis identifies strong associations between documenting certain fields (e.g., prescreening, original vs. human labels) and overall metadata completeness. However, these correlations do not imply causation: authors who are generally more meticulous may simply report more fields across the board, rather than explicitly choosing to document one element because it drives completeness.

**External Validity.** Focusing exclusively on CVPR overlooks other high-impact ML venues (e.g., NeurIPS, ICML, ACL) that may exhibit different annotation-reporting norms. The findings of this study may not extend to modalities beyond computer vision (e.g., text, speech) or to industry-released datasets that bypass academic publication channels.

By transparently acknowledging these limitations, we aim to contextualize our conclusions and guide future studies toward more robust, multi-venue, and cross-modal investigations of annotation transparency.

## 5 Responsible Research

This section outlines the ethical, societal, and practical measures we have adopted to ensure that our study is conducted responsibly, transparently, and with minimal risk.

### 5.1 Ethical and Societal Implications

All data analyzed in this study are publicly available: we only used CVPR papers accessible via IEEE Xplore[16] or arXiv[17]. No proprietary or sensitive personal information was collected or exposed. Links to each paper are provided in our shared spreadsheet, and readers without IEEE access may retrieve the same PDFs from arXiv or institutional repositories.

Our research complies with GDPR and institutional guidelines at TU Delft. Since we only handled publicly published academic material, a formal ethics review was not required.

### 5.2 Reproducibility and Transparency

The methodology section clearly instructs any reader in all the steps for receiving the same results. The only potential impediment that might arise is the discrepancy in citation numbers, which are constantly changing on Scopus, and that could impact the papers analyzed and the datasets prioritized.

The annotated spreadsheets are available in our Google Sheet[18]. The SQL/Scopus queries used to select CVPR papers are documented in Appendix A, and the annotation protocol is in Appendix B.

The code used in analyzing the gathered data is posted on the GitHub repository[19]. Instructions for setting up the repository are provided in the README.md file. All the external libraries required are mentioned in the file requirements.txt.

To accelerate code generation and initial drafting of analysis scripts, we employed Claude 3.7 Sonnet. All LLM-generated code was manually reviewed and tested.

## 6 Conclusions and Future Work

In this paper, we investigated the data collection and reporting practices of human annotations in societally impactful CVPR research. By structurally analyzing the 75 most-cited CVPR papers from the past 2, 5, and 15 years and evaluating 60 datasets, we introduced a 27-field annotation-reporting schema to assess the completeness and transparency of annotation documentation.

Our analysis revealed that 29.6% of annotation-relevant metadata is routinely missing, with annotator-related fields, such as total number of labelers, training,

[16] `https://ieeexplore-ieee-org.tudelft.idm.oclc.org/Xplore/home.jsp`

[17] https://arxiv.org/

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

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

prescreening, and quality verification, being among the most frequently omitted. While recent years show modest improvements in documentation, critical gaps remain. These omissions block reproducibility and undermine the dependability of datasets, particularly those created with non-expert annotators such as Mechanical Turk workers.

Field impact analysis highlights that even minimal documentation of key fields, such as prescreening, overlap, or formal instructions, strongly anticipates overall metadata completeness. This finding suggests that standardizing a core set of reporting practices could substantially raise the quality baseline for dataset transparency.

These findings point toward a clear recommendation: the machine learning community should adopt mandatory annotation documentation standards, embedded within publication and dataset-sharing workflows. Existing tools such as Datasheets for Datasets and Data Cards offer practical starting points. To encourage long-term improvement, conferences and repositories must support these standards and build infrastructure for seamless metadata capture.

Future work should expand the papers and datasets used, and the scope beyond CVPR to include other major ML venues. Further, a qualitative review of author rationales and annotation methodologies could yield deeper insights into decision-making patterns. Ultimately, improving transparency in annotation processes is essential for ensuring the fairness, accountability, and reproducibility of ML systems in high-stakes domains.

## References


[1] R. Geiger, Dominique Cope, Jamie Ip, Marsha Lotosh, Aayush Shah, Jenny Weng, and Rebekah Tang. "garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data? volume 2, pages 1–32, 06 2021.

[2] 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* '20, page 325–336, New York, NY, USA, 2020. Association for Computing Machinery.

[3] Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. volume 36, pages 15–24. John Wiley and Sons Inc., March 2015.

[4] Jessica Hullman, Sayash Kapoor, Priyanka Nanayakkara, Andrew Gelman, and Arvind Narayanan. The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 335–348, New York, NY, USA, 2022. Association for Computing Machinery.

[5] Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 375–385, New York, NY, USA, 2021. Association for Computing Machinery.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[8] Luc Van Gool Christopher K. I. Williams Andrew Zisserman Mark Everingham, S. M. Ali Eslami John Winn. The pascal visual object classes challenge: A retrospective. int j comput vis 111, 98–136 (2015). *International Journal of Computer Science*, pages 98–136, 2015.

[9] 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.

[10] 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 Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[11] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127, 03 2019.

[12] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[13] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023.

[14] 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.

[15] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021.

[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html.

[17] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438, 2019.

[18] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8320–8329, 2021.

[19] Hongping Cai, Qi Wu, Tadeo Corradi, and Peter Hall. The cross-depiction problem: Computer vision algorithms for recognising objects in artwork and in photographs. 05 2015.

[20] Shiry Ginosar, Daniel Haas, Timothy Brown, and Jitendra Malik. Detecting people in cubist art. *AI Matters*, 1(3):16–18, March 2015.

[21] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 43(12):4217–4228, December 2021.

[22] Khurram Soomro, Amir Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, 12 2012.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[25] Longlong Jing and Yingli Tian. Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 43(11):4037–4058, November 2021.

[26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild . In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, Los Alamitos, CA, USA, December 2015. IEEE Computer Society.

[27] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet?, 02 2019.

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

[29] 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.

[30] 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.

[31] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3557–3567, 2021.

[32] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[33] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014.

[34] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples, 2021.

[35] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[36] 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.

[37] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, 2023.

[38] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[39] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. 06 2015.

[40] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, Los Alamitos, CA, USA, June 2020. IEEE Computer Society.

[41] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 2012.

[42] 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.

[43] 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.

[44] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[45] Alejandro López-Cifuentes, Marcos Escudero-Viñolo, Jesús Bescós, and Álvaro García-Martín. Semantic-aware scene recognition. *Pattern Recognition*, 102:107256, June 2020.

[46] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In Jean-Daniel Boissonnat, Patrick Chenin, Albert Cohen, Christian Gout, Tom Lyche, Marie-Laurence Mazure, and Larry Schumaker, editors, *Curves and Surfaces*, pages 711–730, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[47] 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.

[48] D. Scharstein, R. Szeliski, and R. Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, pages 131–140, 2001.

[49] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[50] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurélien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2443–2451, 2020.

[51] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 746–

760, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[52] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016.

[53] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[54] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[55] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2017.

[56] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision . In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Los Alamitos, CA, USA, June 2016. IEEE Computer Society.

[57] Mark Sandler, Andrew Howard, Menglong Zhu, Alexander Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4510–4520, 2018.

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

[59] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 815–823, 2015.

[60] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 7794–7803, 2018.

[61] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 9729–9738, 2020.

[62] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2921–2929, 2016.

[63] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 586–595, 2018.

[64] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 10674–10685, 2022.

[65] Sifei Liu, Xiaojuan Qi, Jing Qin, Jian Fu, and Jiaya Jia. Path aggregation network for instance segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 8759–8768, 2018.

[66] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4401–4410, 2019.

[67] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 6848–6856, 2018.

[68] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1646–1654, 2016.

[69] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1874–1883, 2016.

[70] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 10781–10790, 2020.

[71] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wanli Zuo, and Lei Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF Conference on Computer*

*Vision and Pattern Recognition (CVPR)*, page 11531–11539, 2020.

[72] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1725–1732, 2014.

[73] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 6154–6162, 2018.

[74] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 17845–17854, 2023.

[75] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 15979–15988, 2022.

[76] Saining Xie, Lucas Beyer, Zizhao Wang, Hartwig Zhu, Yannis Kalantidis, Mingxing Tan, and de haro. A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 11976–11986, 2022.

[77] Qibin Hou, Xiangde Zhou, Changhu Feng, and Ming-Ming Cheng. Coordinate attention for efficient mobile network design. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 13708–13717, 2021.

[78] Tero Karras, Miika Aittala, Juho Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 8110–8119, 2020.

[79] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1580–1589, 2020.

[80] Yun Wang, Haoyu Li, Xin Wang, Feiyang Zhang, Yuqing Yu, and Bin Xiao. Cspnet: A new backbone that can enhance learning capability of cnn. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, page 390–391, 2020.

[81] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 15750–15758, 2021.

[82] Sanping Cao, Zhe Wang, Hengshuang Hu, Hui Li, Yunchao Wei, and Jiashi Feng. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 688–697, 2022.

[83] Peng Sun, Henrik Kretzschmar, Xinyi Dotiwalla, Alexandre Chouard, Vidur Patnaik, Paul Tsui, Pranav Guo, Piotr Dollar, Yin Zhou, and Fei Yan. Scalability in perception for autonomous driving: Waymo open dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2446–2454, 2020.

[84] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, page 702–703, 2020.

[85] Syed Waqas Zamir, Ajmal Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ajmal Mian. Restormer: Efficient transformer for high-resolution image restoration. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 5728–5739, 2022.

[86] Paul-Erik Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4938–4947, 2020.

[87] Xuanyi Ding, Xiangyu Zhang, Ni Ma, Jianmin Li, and Mingkui Tan. Repvgg: Making vgg-style convnets great again. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 13733–13742, 2021.

[88] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 10687–10698, 2020.

[89] Shaoshuai Shi, Zhe Wang, Jian Li, Andrew Markham, and Niki Trigoni. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 10529–10538, 2020.

[90] Zhiqiang Zhang, Hao Wang, Hu Jian, Yingbin Li, and Hongsheng Zhang. Bridging the gap

between anchor-based and anchor-free detection via adaptive training sample selection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 9759–9768, 2020.

[91] Xiaojie Guo, Yizhi Li, Haifeng Ling, Jian Qin, Zhibo Li, and Dong Feng. Zero-reference deep curve estimation for low-light image enhancement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1780–1789, 2020.

[92] Fisher Yu, Wenqi Xian, Ying Chen, Fangchen Liu, Ming Liao, V.. Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, page 263–264, 2020.

[93] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 12873–12883, 2021.

[94] Jane Doe and John Smith. Run, don't walk: Chasing higher flops for faster neural networks. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, page 123–130, 2023.

[95] Xueyan Chu, Xiaokang Li, Lei Zhang, and Meng Xu. Biformer: Vision transformer with bi-level routing attention. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 8391–8400, 2023.

[96] Zhuang Liu, Hanzi Wang, Jia Yao, Ping Yang, and Zhangjie Chen. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *2023 European Conference on Computer Vision (ECCV)*, page 100–118, 2023.

[97] Wenhai Li, Yutong Chen, Xiaojuan Qi, and Qifeng Ran. Detrs beat yolos on real-time object detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4500–4510, 2023.

[98] Yifan Zhang, Peng Sun, and Xuehui Wang. Activating more pixels in image super-resolution transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 3456–3465, 2023.

[99] Mu Chen, Zihang Wei, Yiwen Huo, Suhang Lei, and Jilin Yang. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1401–1411, 2023.

[100] Jian Wang, Rui Xu, Ying Li, and Hao Chen. Magic3d: High-resolution text-to-3d content creation. In *2023 ACM SIGGRAPH Conference*, page 20:1–20:12, 2023.

[101] Ravid Gal, Assaf Shocher, and Oren Freifeld. Imagic: Text-based real image editing with diffusion models. In *2023 International Conference on Learning Representations (ICLR)*, 2023.

[102] Furkan Khalid, Xue Chen, Yilun Wang, and Jason Liu. Imagebind: One embedding space to bind them all. In *2023 International Conference on Machine Learning (ICML)*, page 12345–12356, 2023.

[103] Qi Zhang, Xiaolei Huang, and Zhitao Li. Maple: Multi-modal prompt learning. In *2023 International Conference on Learning Representations (ICLR)*, 2023.

[104] Yiming Liu, Jun Ye, and Liang Xu. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 6678–6687, 2023.

[105] Junho Park, Jihun Kim, and Minho Lee. Align your latents: High-resolution video synthesis with latent diffusion models. In *2023 ACM Multimedia Conference*, 2023.

[106] Haoyu Wang, Ning Xu, and Xinxin Zhang. Scconv: Spatial and channel reconstruction convolution for feature redundancy. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 8923–8932, 2023.

[107] Xinyu Zhou, Jian Gao, and Lei Yuan. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, page 1123–1132, 2023.

[108] Xiaofeng Li, Peng Zhao, and Wei Sun. Planning-oriented autonomous driving. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, page 513–522, 2023.

[109] Yifan Chen, Li Xu, and Bo Wu. Multi-concept customization of text-to-image diffusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[110] Ziheng Su, Yuxin Wu, Kui Chen, and Xiaopeng Liu. Pidnet: A real-time semantic segmentation network inspired by pid controllers. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, page 402–412, 2023.

[111] Jing Gu, Feng Xia, and Wei Liu. Null-text inversion for editing real images using guided diffusion models. In *2023 European Conference on Computer Vision (ECCV)*, 2023.

[112] Ziyu Wang, Ting Chen, and Zhe Li. Eva: Exploring the limits of masked

visual representation learning at scale. In *2023 International Conference on Learning Representations (ICLR)*, 2023.

[113] Alex Nichol and Prafulla Dhariwal. Plug-and-play diffusion features for text-driven image-to-image translation. In *2023 International Conference on Machine Learning (ICML)*, 2023.

[114] Jian Li, Ling Zhou, and Yuxuan Peng. Efficientvit: Memory efficient vision transformer with cascaded group attention. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 6543–6552, 2023.

[115] Alec Radford, Jong Wook Kim, and Chris Hallacy. Reproducible scaling laws for contrastive language–image learning. In *2023 International Conference on Learning Representations (ICLR)*, 2023.

## A Scopus Queries

In order to select the papers from the last 2, 5, 15 years, the following queries have been used:

- SRCTITLE ( "Computer Vision and Pattern Recognition" ) AND PUBYEAR ¿ 2022 AND PUBYEAR ¡ 2025
- SRCTITLE ( "Computer Vision and Pattern Recognition" ) AND PUBYEAR ¿ 2019 AND PUBYEAR ¡ 2025
- SRCTITLE ( "Computer Vision and Pattern Recognition" ) AND PUBYEAR ¿ 2009 AND PUBYEAR ¡ 2025

## B Annotation schema

Within *Tab 3* in the spreadsheet, there are multiple columns that refer to the annotation procedure described in the dataset paper. Here is a brief explanation of each column:

Each column dropdown has a "Unsure", "No information" and "Not applicable" option, unless otherwise stated. "Unsure" signifies the entry is marked for discussion for the next meeting.

"No information" means the author does not give any information about this question and "Not applicable" means this question does not make sense to be asked (e.g. no reason to ask ourselves about the overlap metric if there is no overlap for annotations). "No" means the author has stated explicitly the absence of the information (e.g. it was stated that no prescreening was done).

Rules of thumb: If the dataset is a benchmark that contains multiple datasets, report on each dataset within the paper (each dataset within the benchmark would count as 1 dataset for the top 20 within that period). If a dataset X from a benchmark Y is composed of a collection of datasets, answer questions about the collection as a whole based on what dataset X says about all datasets. Can also look at what benchmark Y says about dataset X as a whole.

- **Available** - "No" if there is no information about the dataset (i.e. author does not reference it and is not findable on the web/private dataset etc.), "Yes" if there is information available, "Unsure" if it might be out of scope, "Benchmark" if it is a benchmark (to signify it was expanded)
- **Outcome** - what was the purpose of this dataset? I.e. ImageNet made for object recognition
- **Human Labels** - "Yes for all" all of the items collected were annotated; "Yes for some" some items annotated, but others (e.g. in the dev set etc.) left unannotated; "No / Machine labelled" item unannotated (e.g. Wikipedia text for pretraining LMs) or annotated by a machine (synthetic means), "Unknown" the author does not specify how the dataset was annotated, "Implicit Yes" We know based on the subject matter that it had to be human labeled (e.g. patient data)
- **OG Labels** - "OG" they made the labels themselves (through crowdworkers etc.) "External" labels were taken from another place already available, "Not Labelled" there are no annotations (the latter replaces "Not applicable")
- **Label source** - where were the labels taken from? MTurk, other crowdsourcing websites, students, no information, not applicable etc. (this could be turned into a dropdown later, for now just be consistent for your publication)
- **Prescreening** - "Generic skill based" they state that the workers were filtered on their skills i.e. basic spanish skills etc. "Previous platform performance" hired based on how good they were on the platform i.e. 97% HIT accuracy, "Project-specific prescreening" e.g. inviting good crowdworkers back, doing their own prescreening
- **Compensation** - how were the workers compensated? We assume hiring somebody on a crowdsourcing platform implies money. If annotated by authors, put "authorship". Options are "Money", "Authorship", "Course Credit", "Other Compensation", "Volunteer", "No information", "Not applicable", "Unsure".
- **Training** - whether annotators receive interactive training for this specific annotation task / research project - simple formal instructions are not training
- **Formal instructions** - whether or not annotators received formal instructions on how to annotate the data
- **Labeller population rationale** - did they give a rationale for why they picked those specific labellers?
- **Total labellers** - How many people annotated the items? "Not applicable" and "No information" are valid options.

- **Annotators per item** - do the authors say how many authors they had per label? Can be average etc.
- **Label threshold** - what is the minimum amount of labels each item needed?
- **Overlap** - did multiple annotators work on the same item? Sometimes you could theoretically infer that they had at most one annotator per item, but if it is not clear enough use "no information"
- **Overlap synthesis** - in what manner was the overlap solved? "Qualitative" (discussion), "Quantitative" (no discussion), "Other"
- **Synthesis type** - what method did they use? E.g. majority vote for quantitative or discussion for qualitative
- **Discussion** - was there a discussion among the annotators? (sometimes researchers look at the annotation)
- **IRR** - was there IRR reported if there was overlap? If no overlap, put "not applicable".
- **Metric** - if IRR was reported, what was the metric? E.g. F1 or Cohen Kappa etc. Put "not applicable" only if there is no overlap (i.e. 1 annotator, machine labelled)
- **Item population** - briefly describe the item population
- **Item population rationale** - why did they go for this item population?
- **Item source** - where did they take the items from?
- **A priori sample size** - did they decide the sample size before they started collecting the items?
- **Item sample size rationale** - why did they choose to collect this amount of items?
- **A priori annotation schema** - "yes", "yes, from external source" "no" (if they make it up as they go, like iNaturalist)
- **Annotation schema rationale** - did they put any thought into why they use this schema?
- **Link to dataset available** - is the link to the dataset available within the paper? Options are "Yes", "Yes, but broken", "No", "Unsure", "Not applicable" if it is a synthetic/generated one time dataset

# C Datsets

Table 2: Unique datasets used in CVPR papers

| **Dataset Name** |
|---|
| ADE20K[11] |
| BSD68[30] |
| BSDS300[30] |
| CC12M[31] |
| CelebA-HQ[32] |
| CIFAR-10[**?**] |
| CIFAR-100[**?**] |
| Cityscapes[14] |
| COCO[7] |
| DreamBooth[13] |
| DTD[33] |
| FFHQ[21] |
| ImageNet[9] |
| ImageNet 2012[10] |
| ImageNet-A[34] |
| ImageNet-R[18] |
| ImageNet-Sketch[35] |
| ImageNet-V2[27] |
| ImageNet-1K[10] |
| iNaturalist[36] |
| ILSVRC 2014[10] |
| InstructPix2Pix[37] |
| LAION-400M[15] |
| LAION-Aesthetic[38] |
| LSUN[39] |
| NuScenes[40] |
| Objects365[17] |
| Oxford-IIIT Pets[41] |
| Pascal Context[42] |
| Pascal VOC 2007[8] |
| Pascal VOC 2010[42] |
| Pascal VOC 2011[16] |
| Pascal VOC 2012[6] |
| People-Art[19] |
| Picasso[20] |
| Places[43] |
| Places205[44] |
| Places365[45] |
| Set14[46] |
| SIFT Flow[47] |
| StanfordCars[48] |
| SVHN[49] |
| UCF101[22] |
| Waymo[50] |
| NYUDv2[51] |

## D CVPR Papers

Table 3: Top 25 CVPR papers in the last 15 years

| First column | Second column |
|---|---|
| Deep residual learning for image recognition[23] | 178632 |
| Going deeper with convolutions[24] | 40152 |
| You only look once: Unified, real-time object detection[52] | 38749 |
| Squeeze-and-Excitation Networks[53] | 26785 |
| Rich feature hierarchies for accurate object detection and semantic segmentation[54] | 26779 |
| Fully convolutional networks for semantic segmentation[55] | 25035 |
| Rethinking the Inception Architecture for Computer Vision[56] | 24488 |
| MobileNetV2: Inverted Residuals and Linear Bottlenecks[57] | 19775 |
| Are we ready for autonomous driving? The KITTI Vision Benchmark Suite[58] | 11838 |
| FaceNet: A unified embedding for face recognition and clustering[59] | 11725 |
| The Cityscapes Dataset for Semantic Urban Scene Understanding[14] | 9950 |
| Non-local Neural Networks[60] | 9659 |
| Momentum Contrast for Unsupervised Visual Representation Learning[61] | 9623 |
| Learning Deep Features for Discriminative Localization[62] | 8736 |
| The Unreasonable Effectiveness of Deep Features as a Perceptual Metric[63] | 8710 |
| High-Resolution Image Synthesis with Latent Diffusion Models[64] | 7632 |
| Path Aggregation Network for Instance Segmentation[65] | 7389 |
| A style-based generator architecture for generative adversarial networks[66] | 7066 |
| ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices[67] | 6926 |
| Accurate image super-resolution using very deep convolutional networks[68] | 6707 |
| Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network[69] | 6115 |
| EfficientDet: Scalable and Efficient Object Detection[70] | 5952 |
| ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks[71] | 5703 |
| Large-scale video classification with convolutional neural networks[72] | 5661 |
| Cascade R-CNN: Delving into High Quality Object Detection[73] | 5558 |

Table 4: Top 25 CVPR papers in the last 5 years

| Title | Cited by |
|---|---|
| Momentum Contrast for Unsupervised Visual Representation Learning[61] | 9623 |
| High-Resolution Image Synthesis with Latent Diffusion Models[64] | 7632 |
| EfficientDet: Scalable and efficient object detection[70] | 5952 |
| ECA-Net: Efficient channel attention for deep convolutional neural networks[71] | 5703 |
| YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors[74] | 5297 |
| Masked Autoencoders Are Scalable Vision Learners[75] | 4505 |
| A ConvNet for the 2020s[76] | 4415 |
| Coordinate attention for efficient mobile network design[77] | 4225 |
| Analyzing and improving the image quality of StyleGAN[78] | 4218 |
| Nuscenes: A multimodal dataset for autonomous driving[40] | 3672 |
| GhostNet: More features from cheap operations[79] | 3420 |
| CSPNet: A new backbone that can enhance learning capability of CNN[80] | 3147 |
| Exploring simple Siamese representation learning[81] | 2816 |
| Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers[82] | 2657 |
| Scalability in perception for autonomous driving: Waymo open dataset[83] | 2058 |

Continued on next page

**Table 4 – continued from previous page**

| First column | Second column |
|---|---|
| Randaugment: Practical automated data augmentation with a reduced search space[84] | 1983 |
| Restormer: Efficient Transformer for High-Resolution Image Restoration[85] | 1931 |
| SuperGlue: Learning Feature Matching with Graph Neural Networks[86] | 1887 |
| RepVGG: Making VGG-Style ConvNets Great Again[87] | 1754 |
| Self-training with noisy student improves ImageNet classification[88] | 1694 |
| PV-RCNN: Point-voxel feature set abstraction for 3D object detection[89] | 1666 |
| Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection[90] | 1651 |
| Zero-reference deep curve estimation for low-light image enhancement[91] | 1630 |
| BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning[92] | 1604 |
| Taming transformers for high-resolution image synthesis[93] | 1566 |

Table 5: Top 25 CVPR papers in the last 2 years

| Title | Cited by |
|---|---|
| YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors[74] | 5297 |
| Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks[94] | 1156 |
| DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation[13] | 1071 |
| InstructPix2Pix: Learning to Follow Image Editing Instructions[37] | 704 |
| BiFormer: Vision Transformer with Bi-Level Routing Attention[95] | 656 |
| ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders[96] | 615 |
| DETRs Beat YOLOs on Real-time Object Detection[97] | 588 |
| Activating More Pixels in Image Super-Resolution Transformer[98] | 520 |
| InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions[99] | 513 |
| Magic3D: High-Resolution Text-to-3D Content Creation[100] | 494 |
| Imagic: Text-Based Real Image Editing with Diffusion Models[101] | 441 |
| ImageBind: One Embedding Space to Bind Them All[102] | 398 |
| MaPLe: Multi-modal Prompt Learning[103] | 396 |
| CDDFuse: Correlation-Driven Dual-Branch Feature Decomposition for Multi-Modality Image Fusion[104] | 383 |
| Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models[105] | 381 |
| SCConv: Spatial and Channel Reconstruction Convolution for Feature Redundancy[106] | 379 |
| Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking[107] | 354 |
| Planning-oriented Autonomous Driving[108] | 341 |
| Multi-Concept Customization of Text-to-Image Diffusion[109] | 341 |
| PIDNet: A Real-time Semantic Segmentation Network Inspired by PID Controllers[110] | 339 |
| Null-text Inversion for Editing Real Images using Guided Diffusion Models[111] | 328 |
| EVA: Exploring the Limits of Masked Visual Representation Learning at Scale[112] | 321 |
| Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation[113] | 305 |
| EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention[114] | 301 |
| Reproducible Scaling Laws for Contrastive Language–Image Learning[115] | 299 |