Annotation Practices in Societally Impactful Machine Learning Applications

What are these automated systems actually trained on?

Damjan Košutić[1]

Supervisors: 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: Damjan Košutić
Final project course: CSE3000 Research Project
Thesis committee: dr. Cynthia Liem, Andrew M. Demetriou, dr. Jie Yang

Abstract

The output of machine learning (ML) models can be only as good as the data that is fed into them. Because of this, when making datasets for creating ML models, it is important to ensure the quality of the data. This is especially true of human labeled data, which can be hard to standardize and assess the quality of. To assess the annotation practices of human labeled data in the field of machine learning, this paper investigates the datasets used in the highest cited papers in the AAAI Conference on Artificial Intelligence, an influential machine learning conference. After extracting the datasets from 75 papers in three overlapping publication periods, the top 20 datasets were evaluated from each period. The results showed that the majority of datasets do not use or underreport significant annotation practices, specifically about the annotators and the annotation process. This raises concern for the conference and the field more broadly, as the most influential papers build their machine learning algorithms on quite possibly low quality data. However, there is some hope for the field in this regard as the more recent papers use datasets with better quality annotation practices.

1 Introduction

When training machine learning (ML) models, one often forgotten part of the process is assessing the quality of the input data [29]. The concern of most papers is on the quality of the model that is to be trained, while the input data is assumed to be of sufficient quality. However, if the input data used for training, testing, and/or validation of models is itself of poor quality, the output of the model cannot perform at a sufficient level of quality. This principle in Computer Science is often called "garbage in, garbage out", which states that bad input data will necessarily lead to bad output data [4; 64]. Thus, to ensure the quality of the output of ML models, the data used to create and evaluate the model must itself be of high quality.

An important facet of the quality of ML data is its annotation. Annotation refers to the process of labeling raw data items such as images, text, sound clips, etc. according to a schema that seeks to give desired information about the items. This annotation is often done by humans, so as to give a reliable "ground-truth" that can then be used to train an ML model to predict these labels. However, as human perceptions are often variable [2], quality annotation practices must be used to ensure the quality of the annotations themselves [67]. Additionally, providing poor documentation about the annotation process makes it hard for that dataset to be validated by potential users of the dataset, harming responsible research goals [28].

To provide an assessment of the current annotation practices of the datasets used in the field of ML, this paper assesses the quality of the datasets used in the top cited papers in the AAAI Conference on Artificial Intelligence. The research question this work seeks to answer then is *What are the annotation practices of the datasets used in the highest cited papers in the AAAI Conference on Artificial Intelligence?* These papers will be grouped into three overlapping periods – those published in the last 15, 5, and 2 years – and the datasets extracted from these papers will be categorized in a similar way. The assessment of the datasets will be done through a set of criteria applied to each dataset, evaluating 1) the general documentation about the datasets, 2) the documentation about the annotators and the annotation process, and 3) the documentation about the selection of the items and the schema with which they were annotated.

The rest of this work is structured as follows. Section 2 will focus on the related work for this paper. In Section 3, the procedure taken to gather, extract, and evaluate the datasets will be explained. Section 4 will show the results of the evaluation of the datasets and Section 5 will draw the implications and make interpretations of the findings. Section 6 will demonstrate the ethical implications of this work as well as the measures taken for its reproducibility. Lastly, Section 7 will summarize the main conclusions of the work and provide a discussion of its limitations.

2 Related Work

This work seeks to build upon the contributions of Geiger et al. in [29]. In this work, the researchers investigated how a wide sample of supervised machine learning papers reported on the annotation practices for the data used in the training of their classifiers, with a focus on human annotated data. They did so by taking a representative sample of papers from a diverse range of academic fields and evaluated each paper according to a set of 18 criteria. The first 6 criteria were about the general information of the paper such as did the paper introduce an original ML classification task and did the paper use original human labeling (i.e., whether the dataset was created in the paper itself). With these criteria, they excluded papers that did not use original human annotated data from the remaining 12 criteria. The remaining criteria were about who the annotators were, about the compensation, training, and instructions given to the annotators, and about the annotation process, like whether overlap was used, was inter-rater reliability measured, were multiple labels given per item, etc. Their results showed that only around one quarter of the sampled papers used original human annotated data and, among those, two thirds provided no information on the remaining 12 criteria. This, they concluded, raised concern as "high-quality training data is essential to the validity of ML classifiers and human judgment is notoriously difficult to standardize" [29, p. 818].

The current work takes a different approach in several ways. It focuses on the dataset papers and sources as opposed to the machine learning papers themselves, looking at how these dataset papers and sources report on their annotation practices. By doing so, this paper avoids losing most of the sample and provides an option for the datasets to originate not only from academic papers but from any internet accessible source. Additionally, this work provides a different perspective as it takes the datasets used in the highest cited papers

found on the AAAI Conference on Artificial Intelligence, as opposed to a representative sample of papers from multiple academic fields, thus focusing only on ML papers and only on the most impactful ones published in the venue. The criteria used to evaluate the datasets in this work use most of the 18 criteria as the ones in Geiger et al. work but expand them to 27, as explained in Section 3.2. Finally, the analysis done in this work is categorized into three overlapping periods, as opposed to a single general one.

As for the other related work, three papers will be described. The first paper is [30] by Geiger et al., a paper on which [29] is based upon. Similarly to [29], this paper took a sample of supervised ML papers and evaluated them according to a set of criteria, which sought to evaluate the data the classifiers introduced in those papers were trained upon. The methodology also excluded the papers that did not introduce an original ML classifier and use human annotated data. However, the sample was only made out of classifiers that classified Tweets for diverse purposes and the evaluation criteria were smaller in extent than in [29]. The results showed a "wide range in levels of documentation" [30, p. 333], indicating that there is a large portion of datasets that provide poor documentation but also some that give hope for the field. The current work seeks to investigate if the same is true for the annotation practices of the highest cited paper in the AAAI Conference for Artificial Intelligence and seeks to give an assessment not only of the datasets used in papers for Tweet classification but for the datasets used in any type of ML task.

The second paper is [16] by Daneshjou et al. This paper sought to evaluate the documentation of the datasets used to train ML models for skin disease diagnosis. It did so through evaluating a sample of 70 dermatological papers with a set of evaluation questions, specifically aimed at evaluating images for skin disease diagnosis. Its results showed that only 64.3% of the analyzed papers met the "gold standard criteria" [16, p. 4] for disease labeling and the paper concluded that there was a "sparsity of data set characterization and lack of transparency" [16, p. 2] and the use of "nonstandard and unverified disease labels" [16, p. 2] among the papers surveyed. The current work seeks to build upon the evaluation of datasets used for the creation of ML models but focusing on the field of ML itself, and not specifically the field of dermatology. Thus, the current work will not evaluate only the datasets used in dermatological papers nor use evaluation criteria geared toward those but evaluate the datasets used in papers from the field of ML and utilize a more general set of evaluation criteria.

The third related paper that will be described is [76] by Sav et al. This paper investigated the standards of documenting data collection and annotation processes used for recommender systems papers. It did so by surveying the 100 most highly cited recommender systems papers from the most impactful venues in the area of Computing and Information Technology and evaluating them using a set of criteria similar to the ones used in [29]. The results of the analysis showed that a large portion of the papers used bad documentation practices with severe reproducibility issues and that there was a lack of a robust reporting framework. Interestingly, the work also found that the guidelines for the annotation process were usually given but that "little information was provided regarding the population from which the annotators were drawn" [76, p. 12], raising issues of generalization onto real world populations. The current work seeks to use a similar approach by evaluating the datasets used in a set of the most highly cited papers in the AAAI Conference for Artificial Intelligence but not across multiple venues in the field of ML. Furthermore, it seeks to investigate if the similar reproducibility and generalizability issues will be present and if the annotation practices will be frequently given as this work has found.

3 Methodology

For the purposes of assessing the annotation practices in the field of ML, the choice to focus on only one conference was made due to the limited time frame of the project, and the decision to use the AAAI Conference on Artificial Intelligence specifically was made due to it having one of the highest h5-index values within the field of Artificial Intelligence at the time of writing[1]. The h5-index was utilized as it is a widely used indicator of a journal/conference's academic relevance, based upon recent citations of its papers[2], meaning that the selected conference has a high relevance in the field of ML. It was also decided to extract the datasets used only in the highest cited papers, and not a representative sample, as these papers have been the ones that have shaped the research field the most and thus their datasets serve as relevant objects of study. The impact of a paper was measured in terms of its citation count, as that is a standard and widely accepted metric for impact. It is important to note that the citation counts do not necessarily represent the quality of the paper but do provide a measure of impact, which this work is interested in.

To examine the annotation practices of the datasets used in the highest cited papers in the AAAI Conference on Artificial Intelligence the following methodology was used. The process was split into three main steps: 1) gathering the highest cited papers from the conference, 2) extracting the datasets used in those papers, and 3) assessing each dataset according to a set of criteria. The assessment criteria used in the third step were split into three categories: 1) assessing the general information about the dataset, 2) assessing the information about the annotators and the annotation process, and 3) assessing the information about the items and the annotation schema. This analysis was done in Google Sheets[3], and since there was overlapping work among the project team, this table was shared.

3.1 Steps 1 & 2 – Collecting the Datasets

In step 1 of the process, the highest cited papers from the conference were gathered. This collection was done by query-

[1] A list of the conferences with the highest h-5 indexes in Artificial Intelligence can be found here: https://scholar.google.nl/citations?hl=en&vq=eng_artificialintelligence&view_op=list_hcore&venue=PV9sQN5dnPsJ.2024

[2] The definition given on Wikipedia can be found here: http://en.wikipedia.org/w/index.php?title=H-index&oldid=1289491161

[3] Link to the spreadsheet: https://docs.google.com/spreadsheets/d/16MkuS-upEQxkAj-poZO5ggPqmu_UIDbwi7HWS3-21HE/edit?usp=sharing. Each of the three steps are placed in a separate tab.

ing the top 25 cited papers from the AAAI Conference on Artificial Intelligence on Scopus[4], in three overlapping time ranges. These ranges were: 15 years (papers from 2010 to 2024, inclusive), 5 years (papers from 2020 to 2024, inclusive), and 2 years (papers from 2023 to 2024, inclusive). The search queries can be found in Appendix A. The analysis, therefore, considered a total of 75 papers, which can also be found in Appendix A.

There were several design decisions that were made for Step 1. The querying was done on Scopus, as Scopus is regarded as a reliable search database for academic papers by many researchers [3]. To obtain a copy of the papers, the DOIs of the papers were used, which were retrieved with the queries. For some papers, the DOIs were unavailable, so they were manually retrieved from the conference's official proceedings website[5]. The querying was done in the three aforementioned ranges, to give three differently sized perspectives on the datasets. The 15 year range gave a long perspective, the 5 year range a medium one, and the 2 year range a short, recent one. The ranges are taken from 2024, as the current year, 2025, is not finished yet. The number of 25 was chosen in a discussion with the project supervisor, so as to get a considerable and realistic number of datasets to analyze in step 2.

In step 2, datasets were manually extracted and documented from the the papers collected in the previous step. The datasets that were extracted must have been used in the papers for (pre-)training, validating, and/or testing a machine learning model, as all of these steps contribute to the creation of the model. Survey papers were excluded from the analysis, since they only summarize the work of other papers. The datasets were sourced from the original papers themselves, from external papers, and/or from external internet sources, like GitHub repositories, dedicated websites, and similar. If the datasets were externally sourced, there was a preference for using the official paper describing the dataset, but external internet sources were also investigated for thoroughness. At times, datasets were improperly referenced by providing the wrong paper or source, and so an effort was made for each of these datasets to find the original paper/source (which was not always possible). Since the papers found in step 1 were queried from overlapping periods, certain papers appeared in multiple periods. These papers were only analyzed once. After excluding the papers that were duplicates, had no datasets, or were surveys, there were 68 distinct papers that were analyzed.

3.2 Step 3 – Evaluating the Datasets

In step 3, the following procedure was taken to evaluate the datasets. Firstly, the datasets were ranked on the basis of their "citation sums" within the time range in which they were used. A citation sum was defined as the sum of the citation counts of all the papers that use a given dataset in the papers found in step 1 in a particular time range. This provided a metric for the impact of that dataset in the specified time range. The formula is defined as:

$$CitationSum_{d,t} = \sum_{p \in P_{d,t}} Citations(p) \quad (1)$$

where:

- $CitationSum_{d,t}$ is the citation sum for a dataset d in a time range t
- $P_{d,t}$ is the set of papers that use the dataset d in the time range t
- $Citations(p)$ is the number of citations of paper p

After ranking the datasets, the top 20 datasets from each period were evaluated based upon a set of predetermined criteria. The amount of 20 datasets per period were chosen with respect to the time restrictions of the project, in agreement with the supervisor. Considering that some datasets were in the top 20 for multiple periods, the total number of distinct datasets that were analyzed was 49. Additionally, since the work of the other students on the project team also included evaluating datasets with the same criteria, some datasets were overlapping with the ones chosen for this work. These datasets were not evaluated multiple times but were evaluated only once by a single project team member. This design decision was made on the basis of the time restrictions of the project. Of the 49 datasets included in this analysis, 45 were analyzed by the author of this work, and the remaining 4 by other team members. The a list of the datasets analyzed per period can be found in Appendix B.

There were 27 criteria that were used to evaluate the annotation practices of a given dataset. Each criterion either had a set of answer options or an open format answer. The criteria were divided into three categories:

1. **General information about the datasets.** This category included criteria on whether any information was available about the dataset, what was the dataset made for, did humans label the dataset, did the dataset collect its own annotations or not, and whether the dataset included a link to the data. This constituted a total of 5 criteria. This set of criteria tried to cover the general properties of the datasets that were deemed useful for the analysis.
2. **Information about the annotators.** This category included criteria on who the annotators were, were they trained, prescreened, how many annotators were there, were multiple annotators used for each item, etc. This constituted a total of 15 criteria. These criteria try to assess the quality of the annotators and their annotation process.
3. **Information about the items and the annotation schema.** "Item" here refers to the data item that is to be annotated with a label. This category included criteria on what the item population and sample were, was the rationale given for using them, where were the items sourced from, was the sample size determined prior to sampling, and criteria about the "annotation schema" – the labeling system. This constituted a total of 7 criteria. These criteria try to give an understanding of how, why, and from where were the items taken as well as the rationale for using the annotation schema.

[4]https://www.scopus.com

[5]https://ojs.aaai.org/index.php/AAAI

These criteria were largely based upon the 18 criteria used in the methodology of [29]. These criteria were proposed by the supervisor and adjusted in agreement with the student team. During the process of annotation, if it was not clear how to answer a criterion for a dataset, the team members discussed these datasets mutually and with the supervisor. A shared evaluation schema was also used to standardize the answers among the student team which included explanations for each criterion and its answers, with examples. This schema can be found in Appendix C.

For the findings section, the criteria and their answers will be transposed to a question/answer format with equal answer options for all so that they may be more easily compared and visualized. The answer options that will be available are "Yes", "Partially", "No information", "No", and "Not applicable", with "Partially" and "No information" being applicable to only some criteria. The questions will be framed in the way that the answers of "Yes" will be the positive answers (a higher percentage is better), while "No information" and "No" will be the opposite. "Partially" will denote a partial coverage of the question, and "Not applicable" will denote that the question does not apply for that particular dataset. To provide an example of this simplification, for the criterion of "Formal Instructions", which describes if the annotators were given formal instructions during the annotation process, the question will be "Were formal instructions given?" and the answers will map as follows: "Formal instructions" = "Yes", "Some instructions" = "Partially", "No information" = "No information", "No instructions" = "No", "Not applicable" = "Not applicable".

As for the description of these results, the percentage of the number of datasets that provided a particular answer or set of answers to a question will be presented. This percentage will take the number of datasets that had a particular answer and divide by the total number of datasets – either 49 or 20 depending on if the datasets are taken overall or per period. The "Not applicable" datasets may be excluded from the total number of datasets which will be explicitly mentioned. For the computational analysis and visualization of the data, a Python project was created[6] and shared among the team, which utilized several libraries, including NumPy, Matplotlib, scikit-learn, pandas, and others.

4 Findings

This section will discuss the results of the research conducted as described in Section 3. The results will focus on the data gathered in step 3 and will be documented in four sections: one for each category of the criteria and one for a comparison between the time periods. As described in Section 3.2, the analysis took the 20 datasets out of each period with the largest citation sums, and the total number of datasets that were analyzed was 49. The findings in all sections take the results from all datasets into account but only the last section takes into account the categorization by time period. The main figure that is introduced in this section is Figure 1, which shows a summary of the documentation statuses for all the datasets analyzed by the criteria used in Section 3.2, regardless of the period. Some criteria were omitted from the figure as they were auxiliary – they only helped describe another criterion – and would not contribute to the overall figure. An example of this is the criterion "IRR Metric", which describes the inter-rater reliability metric if inter-rater reliability was reported for the dataset in a previous criterion. The raw data for the evaluation of the datasets in the question/answer format per each period and overall can be found in Appendix D and the raw data that was used during the evaluation process can be found in the previously linked Google Sheets spreadsheet.

4.1 Findings for the General Information About the Datasets

The criteria for the general information about the datasets indicate mixed results. These criteria are denoted by questions 1-4 and 25 in Figure 1. The criteria for the availability of information about the dataset (question 1), the outcome or purpose of the dataset (question 2), and the criterion for if the labels were original (question 4) show high to very high results, with the first two criteria having 98.0% and 95.9% datasets giving a "Yes" response, respectively, and the last criterion having 87.9% datasets giving a "Yes" or "Partially" response, when the "Not applicable" datasets are excluded. However, the criterion for if humans labeled the data (question 3), and if the dataset provided a link to the data (question 25) show less high results, with the former having 59.2% of the datasets giving a "Yes" or "Partially" response, and the latter having 44.9% of the datasets responding with a "Yes" response. The "Partially" responses were not included in the percentage for the criterion about the presence of a link to the data, as those responses denote the presence of a broken link. These responses did, however, constitute a large percentage of the datasets – 20.4%.

4.2 Findings for the Information About the Annotators

Overall, the results for the criteria for the information about the annotators indicate a low level of documentation about the quality of the annotators and their annotation process in the datasets. These criteria are denoted by questions 5-17 in Figure 1. The highest scoring criterion was the one describing the label source (question 5), which described who the annotators were (i.e., volunteers, students, professionals, etc.) and had only 54.5% of the datasets provided a label source (i.e., provided a "Yes" response), when excluding the datasets that were not applicable. This category did not have any "Partially" responses. On the other hand, the lowest scoring criterion was the one describing the discussion (question 16), which documented if the annotators were able to discuss items among each other if there was confusion or difference in opinion. Only 7.4% of the datasets provided a "Yes" response to the question, when excluding the "Not applicable" answers, and there were no "Partially" responses. Furthermore, the averages for the criteria were also quite low, with the average number of datasets that provided a "Yes" or "Partially" response being 35.1%, when excluding the "Not applicable" responses. A table of the averages can be found in

[6]The codebase for the project can be found in this GitHub repository: https://github.com/Gargant0373/DatasetAnalysis

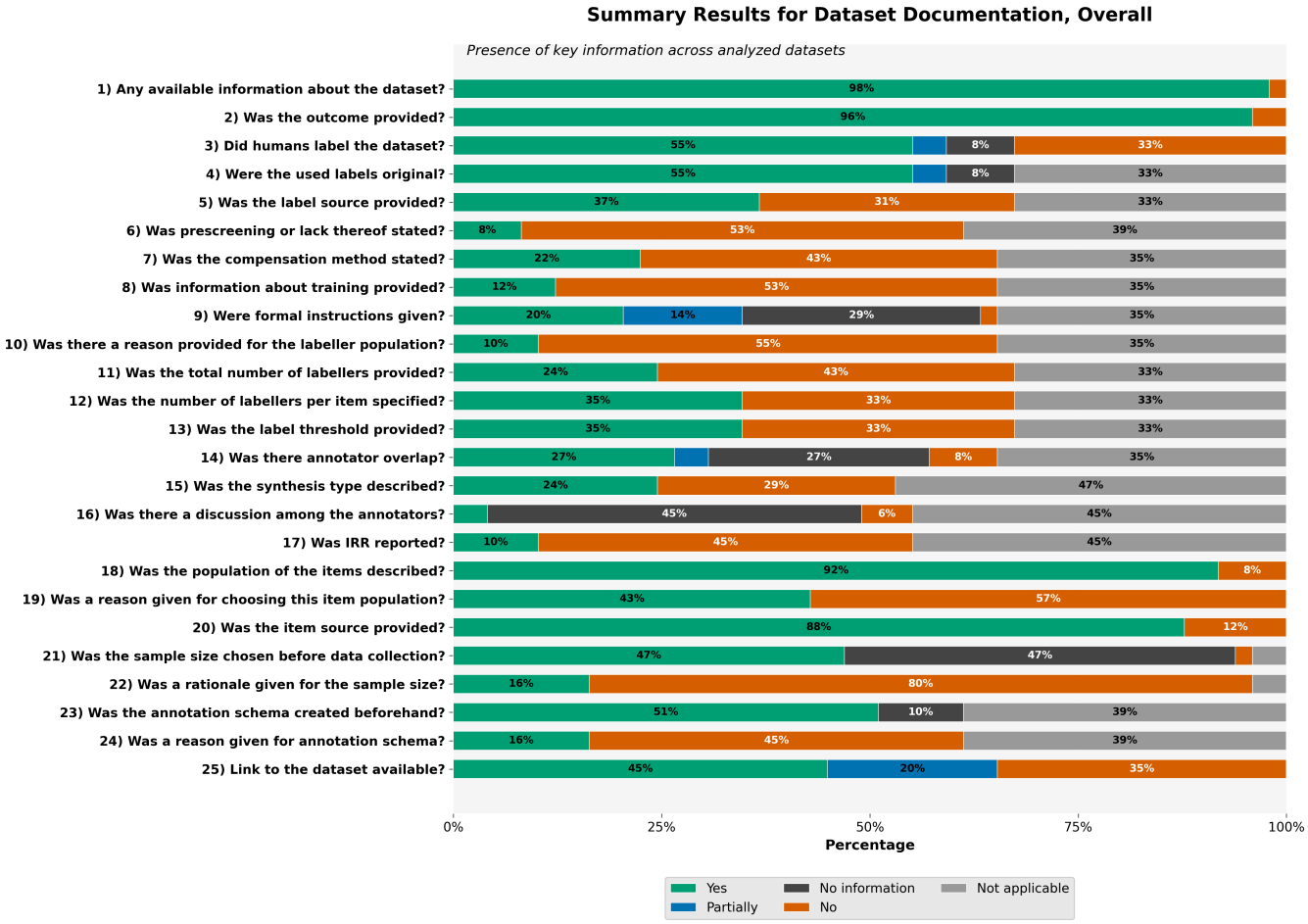

Figure 1: Results of the evaluation of all datasets. Each question on the left represents a criterion from Section 3.2. The available answers of the criteria are mapped to the options of "Yes", "Partially", "No information", "No", and "Not applicable", as described in the same section. The bottom axis represents the number of datasets that fall into one of the answer options. The maximum of the axis is 49, which is the total number of datasets analyzed.

Response type	Average Percentage
Yes	20.72%
Partially	1.41%
No information	7.69%
No	33.28%
Not applicable	36.89%

Table 1: Average responses per question for the annotator criteria. The "Average Percentage" column shows the percent of datasets that gave a particular answer averaged across all the annotator criteria.

Table 1. It is important to note that 32.7% of the datasets were deemed not applicable for any of these criteria as they were either not labeled or did not use human-made annotations (check question 3). Some criteria may have more "Not applicable" datasets because of the nature of the individual question.

4.3 Findings for the Information About the Items and the Annotation Schema

The findings for the criteria for the information about the items and the annotation schema show large variation between the criteria. These criteria are denoted by questions 18-24 in Figure 1. The criteria of the description of the item population (question 18), the description of the item source (question 20), and whether the annotation schema was chosen prior to annotating items (question 23) show high positive responses, having 91.8%, 87.8%, and 83.3% of the datasets providing a "Yes" response, respectfully, when excluding the "Not applicable" datasets. There were no "Partially" answers for these criteria. However, the criteria for the rationale of the item population (question 19), the choosing of the sample size prior to sampling (question 21), the rationale for the sample size (question 22), and the rationale for the annotation schema (question 24), all mid to low positive responses, having 42.9%, 48.9%, 17.0%, and 26.7% of the datasets giving a "Yes" response, respectfully, when excluding the "Not applicable" datasets. There were no "Partially" responses.

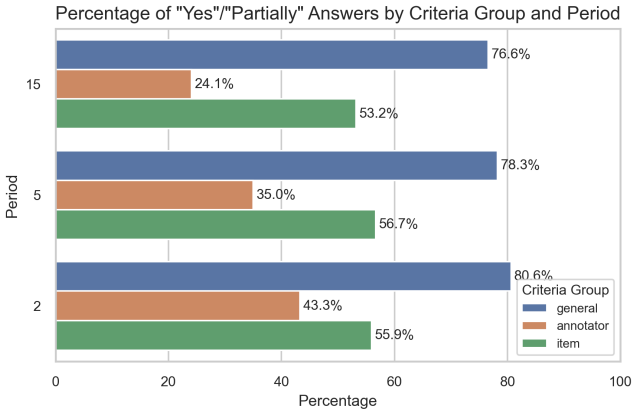

Figure 2: Percentage of positive responses per criteria group per period. The scores are calculated by adding up the "Yes" and "Partially" answers in each of the three categories of criteria and dividing by the number of datasets analyzed per period, excluding the number of datasets that had "Not applicable" as the answer.

4.4 Differences in the Findings for the Three Periods

The findings indicate significant differences between the three periods in terms of annotation practices. As can be seen in Figure 2, the percentage of positive responses to the criteria overall increase as the period is shorter, especially in the case of the criteria evaluating the information about the annotators, with the latter being 24.1% in the 15 year period, 35.0% in the 5 year period, and 43.3% in the 2 year period. As for the other two criteria categories, the criteria for the general information about the datasets shows a slight increase with the shortening of the period, from 76.6% in the 15 year period, 78.3% in the 5 year period, to 80.6% in the 2 year period. Similarly, the criteria for the information about the items and the annotation schema shows a slight increase from the 15 year period to the 5 year period (from 53.2% to 56.7%), but then a slight decrease in the 2 year period (55.9%).

5 Discussion

This section will discuss the results presented in Section 4. The section will cover what the results from each of the different categories of the criteria imply for the AAAI Conference on Artificial Intelligence. Also, it will provide an analysis of the differences in the findings among the periods.

For the criteria for the general information about the datasets, the results indicate the following. Almost all dataset sources provide some information about their datasets and the outcome or purpose for which they were made. This was to be expected, as it is improbable that the highest cited papers in the conference use datasets that either cannot be found or are unclear as to what they represent. A major portion (around 33%) of the datasets were machine made, meaning that they were excluded (deemed "Not applicable") for the criteria for the information about the annotators. This means that a third of the datasets was not available for the analysis of the major part of the criteria. Lastly, working links to the data itself were available only in 45% of the datasets, indicate a worrying fact that most datasets do not provide the official links to their datasets or provide them but do not upkeep them. This implies that if someone is interested in using those datasets, they will have to find the source elsewhere, potentially getting flawed data.

The low results for the criteria for the information about the annotators are a cause for concern. Since most datasets do not provide information about their label source, prescreening of annotators, means of compensation, information about training, and how they were chosen out of the population as well as information about the number of annotators in total, the number of annotators per item, the number of labels per item, the use of overlap, and, if overlap was used, the means of synthesis between labels, the use of discussion, and the use of an inter-rater reliability (IRR) metric, this implies that there is a poor reproducibility of the data and quite possibly its quality, as these processes that involve human judgment are in no way standardized or documented. On a broader level, this implies that the most impactful papers from the conference are significantly dependent on datasets that have a very questionable quality which means that the models trained in these papers themselves should also be questioned, by the principle of *garbage in, garbage out*. These models may have quite bad generalizability onto real-world applications, thus the results for the models reported in those papers may be severely inaccurate.

As for the results of the criteria for the information about the items and the annotation schema, the mixed results may indicate that some of the easier-to-report and more "obvious" criteria tend to be documented more than the other harder-to-report and more "obscure" ones. The criteria for the description of the item population, item source, and whether the annotation schema was chosen prior to annotation can be taken to score highly because they are more easily described and are more "obvious", since they describe the basic properties of the dataset (like what is it made out of, where were the items taken from, and what was the annotation schema). It is also possible that other potential users would be suspicious of a dataset that did not specify these basic things, and thus these highly impactful datasets have these properties specified. On the other hand, the criteria for the rationale of the item population, sample size, and annotation schema, as well as whether the sample size was determined prior to sampling may take more effort to specify and are more "obscure" as their value may not be immediately obvious. This line of reasoning would then imply that generally more convoluted and obscure properties will be more rarely described, which points to the need for standardization of the documentation of these properties, thus ensuring that less obvious properties get documented as well.

The increase in the positive responses with the shortening of the period of analysis indicates a larger focus on the annotation practices and their documentation in recent years in the conference and provides a source of hope. This is especially the case with the information provided about the annotators, with this category having the highest increase over the periods (as documented in Section 4.4). This is crucial as this is the category with the least positive responses and the most

important category from the perspective of annotation practices. These results may be explained as a consequence of the increase in the awareness of the importance of the quality of annotation practices and their documentation over the years in the conference itself and the Machine Learning field at large. Thus, this provides a source of hope for the conference, as the newer papers that are released tend to pay more attention to the annotation practices and their documentation, especially in the category that matters the most – the information about the annotators and the process of annotation. However, there is still a long ways to go, since the average of the positive responses is still relatively low in the 2 years period - only 43.3%.

6 Responsible Research

To ensure that the research made is reproducible, several steps were taken in the writing of this report. For the work done in step 1, Section 3.1 documents where the papers were collected from, how many papers were collected, what time ranges were used, and provides a reference to Appendix A which contains the exact queries which were used, on what date, and the results of the queries themselves. The issue that could not be practically assessed is that the citation counts for the papers queried may change if the same queries are run again, thus resulting in a different set of papers to extract the datasets from. This should not influence the results drastically as the most cited papers should remain mostly the same. For step 2, Section 3.1 also provides information on how the datasets were extracted from the papers collected in step 1 and which papers were excluded from the analysis. Finally, for step 3, Section 3.2 details how the datasets were ranked and chosen to be analyzed and provides a description of what evaluation criteria were used. The section also references Appendix C which contains the evaluation schema used by the project team in the evaluation process which describes what each criterion assessed and what were the available answers. Furthermore, Section 4 references Appendix D which documents the raw results of the evaluation of the datasets, and references the Google Sheets spreadsheet used by the project team in the evaluation process of the datasets. Also, a link is provided to the GitHub repository containing the code that was used to create the figures in this paper, thus allowing for their reproducibility. Therefore, as this work documents the steps made in the process of collecting the top cited papers from the AAAI Conference on Artificial Intelligence, extracting the datasets, evaluating them, and documenting the results, this work ensures the reproducibility of the research made.

The ethical implications that were considered for this work are the following. Since this work tried to give an assessment of the annotation practices of the datasets used in the highest cited papers in the AAAI Conference for Artificial Intelligence, and since this assessment showed negative results, if the analysis was done poorly, then this work might unnecessarily tarnish the reputation of the conference. Therefore, effort was put into standardizing the evaluation process, by creating an evaluation schema and being in constant communication with the supervisor to ensure the accuracy of the results and minimize the chances of this negative implication. On the other hand, the negative results of this work might contribute to helping the conference and the field of ML at large realize the bad quality of the annotation processes used for the creation of the datasets and thus help in improving the annotation practices. This would then constitute a positive ethical implication.

7 Conclusion

To conclude, while the annotation practices of the datasets used in the highest cited papers in the AAAI Conference for Artificial Intelligence show improvements in recent years, there is still a lot of worrying bad practices. As the results from the analysis of the top 20 impactful datasets from past 15, 5, and 2 years have shown, the documentation for the annotation practices is improving but even in the last 2 years, on average, only 43.3% of the analyzed datasets gave positive responses to the criteria relating to the annotators and the annotation process. As for the other criteria that were used, the criteria relating to the general documentation of a dataset and the criteria relating to the documentation about the items and the annotation schema, show mixed results, with some scoring highly and others quite low, indicating that certain information about the annotation process may be more "obscure" or hard and some more "obvious" or easy to document. Overall, this means that a lot of the datasets that are used in the highest cited papers in AAAI Conference for Artificial Intelligence, and possibly in the field of ML at large, have doubtable validity, and that the ML models that were trained with those datasets in those papers have questionable quality, by the principle of "garbage in, garbage out".

7.1 Limitations

As for the limitations of this work, there are several to consider. Firstly, because of the focus on the most highly impactful papers and datasets, there is a reduced generalization of the results onto other papers and datasets. Less highly cited papers may use other datasets which may have annotation practices of a different quality. However, it would be expected that the most highly impactful papers would use more quality datasets than the average (considering their high citation counts) but this is no certainty.

Secondly, the work only considered 75 papers and 49 datasets extracted from them. This limits generalization as this sample was not large and therefore may not map well to the exact practices in all highly cited papers from the conference.

Thirdly, since the datasets used in the papers of only one conference were analyzed, there is a reduced generalization onto the ML field at large. The highly cited papers published in other venues/conferences other than the AAAI Conference for Artificial Intelligence may use different datasets with different quality annotations. However, as the selected conference is among the highest cited ones in the field of ML, its papers will probably be in the highest cited papers in the field overall, thus providing a good approximation of the datasets used in the highest cited papers in the field overall. Of course, this is not a certainty, as other highly cited conferences might

have better/worse standards for the datasets that they use in their papers.

Finally, there is a reduced validity of the work because the extraction of the datasets and their evaluation was done only by the author, and not by multiple people. This means that some datasets may have been left out of the analysis because they were not properly extracted in step 2. This also means that some of the results may be skewed due to biased responses to the criteria. However, the use of a structured evaluation schema and the constant communication with the supervisor of the project helped standardize the evaluation process.

References

[1] Sercan Ö. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):6679–6687, May 2021.

[2] Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, Mar. 2015.

[3] Jeroen Baas, Michiel Schotten, Andrew Plume, Grégoire Côté, and Reza Karimi. Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, 1(1):377–386, 02 2020.

[4] Charles Babbage. *Passages from the life of a philosopher*. Longman, Green, Longman, Roberts, & Green, London, 1864.

[5] Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. arXiv preprint arXiv:1207.4708, 2012. Submitted on 19 Jul 2012, last revised 21 Jun 2013.

[6] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, Mar. 2024.

[7] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. Rumor detection on social media with bi-directional graph convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):549–556, Apr. 2020.

[8] Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439, Apr. 2020.

[9] Guillaume Bouchard, Sameer Singh, and Théo Trouillon. On Approximate Reasoning Capabilities of Low-Rank Vector Spaces. In *AAAI Spring Symposium Series*. AAAI Press, 2015.

[10] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Fredo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *CVPR 2011*, pages 97–104, 2011.

[11] 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, 2020.

[12] Centers for Disease Control and Prevention (CDC). National, regional, and state level outpatient illness and viral surveillance dashboard. https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html, 2025. Accessed June 20, 2025.

[13] Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. Nhits: Neural hierarchical interpolation for time series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6989–6997, Jun. 2023.

[14] Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3438–3445, Apr. 2020.

[15] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762, Mar. 2024.

[16] Roxana Daneshjou, Mary P Smith, Mary D Sun, Veronica Rotemberg, and James Zou. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review. *JAMA Dermatol*, 157(11):1362–1369, November 2021.

[17] Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4027–4035, May 2021.

[18] 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, June 2009.

[19] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1201–1209, May 2021.

[20] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge

graph embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[21] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 2009.

[22] Yahya H. Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A. Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7494–7502, Jun. 2023.

[23] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5369–5378, 2019.

[24] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3674–3683, 2020.

[25] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3558–3565, Jul. 2019.

[26] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[27] Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. On the effectiveness of parameter-efficient fine-tuning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):12799–12807, Jun. 2023.

[28] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, November 2021.

[29] R. Stuart 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? *Quantitative Science Studies*, 2(3):795–827, 11 2021.

[30] 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.

[31] Deepti Ghadiyaram and Alan C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2016.

[32] Chunle Guo, Ruiqi Wu, Xin Jin, Linghao Han, Weidong Zhang, Zhi Chai, and Chongyi Li. Underwater ranker: Learn which is better and how to be better. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):702–709, Jun. 2023.

[33] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):922–929, Jul. 2019.

[34] Matteo Hessel, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[35] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

[36] Xiaobin Hu, Shuo Wang, Xuebin Qin, Hang Dai, Wenqi Ren, Donghao Luo, Ying Tai, and Ling Shao. High-resolution iterative feedback network for camouflaged object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):881–889, Jun. 2023.

[37] 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.

[38] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, Jul. 2019.

[39] Jiahao Ji, Jingyuan Wang, Chao Huang, Junjie Wu, Boren Xu, Zhenhe Wu, Junbo Zhang, and Yu Zheng. Spatio-temporal self-supervised learning for traffic flow prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):4356–4364, Jun. 2023.

[40] Jiawei Jiang, Chengkai Han, Wayne Xin Zhao, and Jingyuan Wang. Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):4365–4373, Jun. 2023.

[41] Renhe Jiang, Zhaonan Wang, Jiawei Yong, Puneet Jeph, Quanjun Chen, Yasumasa Kobayashi, Xuan Song, Shintaro Fukushima, and Toyotaro Suzumura. Spatio-temporal meta-graph learning for traffic forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):8078–8086, Jun. 2023.

[42] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):1042–1050, Jun. 2023.

[43] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8018–8025, Apr. 2020.

[44] 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.

[45] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017. Submitted 19 May 2017.

[46] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 04 2009.

[47] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. arXiv preprint arXiv:1703.07015, 2018. Submitted on 21 Mar 2017, last revised 18 Apr 2018.

[48] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11336–11344, Apr. 2020.

[49] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5455–5463, 2015.

[50] Mengzhang Li and Zhanxing Zhu. Spatial-temporal fusion graph neural networks for traffic flow forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4189–4196, May 2021.

[51] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13094–13102, Jun. 2023.

[52] Qimai Li, Zhichao Han, and Xiao-ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[53] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):1405–1413, Jun. 2023.

[54] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[55] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1486–1494, Jun. 2023.

[56] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1477–1485, Jun. 2023.

[57] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1504–1512, Jun. 2023.

[58] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11474–11481, Apr. 2020.

[59] 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.

[60] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):2901–2908, Apr. 2020.

[61] Hu Lu, Xuezhang Zou, and Pingping Zhang. Learning progressive modality-shared transformers for effective visible-infrared person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1835–1843, Jun. 2023.

[62] Xintian Mao, Yiming Liu, Fengze Liu, Qingli Li, Wei Shen, and Yan Wang. Intriguing findings of

frequency selection for image deblurring. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1905–1913, Jun. 2023.

[63] Max-Planck-Institute for Biogeochemistry. Max-planck-institute for biogeochemistry jena weather dashboard. https://www.bgc-jena.mpg.de/wetter/, 2025. Accessed June 20, 2025.

[64] William Mellin. Work with new electronic 'brains' opens field for army math experts. *The Hammond Times*, 1957.

[65] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5191–5198, Apr. 2020.

[66] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4296–4304, Mar. 2024.

[67] Joseph Nassar, Viveca Pavon-Harr, Marc Bosch, and Ian McCulloh. Assessing data quality of annotations with krippendorff alpha for applications in computer vision. *CoRR*, abs/1912.10107, 2019.

[68] National Centers for Environmental Information (NCEI), NOAA. Local climatological data (lcd). https://www.ncei.noaa.gov/data/local-climatological-data/, 2025. Accessed June 20, 2025.

[69] Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, Tao Schardl, and Charles Leiserson. Evolvegcn: Evolving graph convolutional networks for dynamic graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5363–5370, Apr. 2020.

[70] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[71] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11908–11915, Apr. 2020.

[72] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. Regularized evolution for image classifier architecture search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4780–4789, Jul. 2019.

[73] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[74] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. arXiv preprint arXiv:1609.01775, 2016. Submitted on 6 Sep 2016, last revised 19 Sep 2016.

[75] 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*, 115(3):211–252, Dec 2015.

[76] Andra Georgiana Sav, Andrew M. Demetriou, and Cynthia C.S. Liem. Annotation practices in societally impactful machine learning applications: What are popular recommender systems models actually trained on? *CEUR Workshop Proceedings*, 3476, 2023. 3rd Workshop Perspectives on the Evaluation of Recommender Systems, PERSPECTIVES 2023 ; Conference date: 19-09-2023.

[77] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

[78] Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar. 2016.

[79] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. arXiv preprint arXiv:1604.02808, 2016. Submitted 11 April 2016.

[80] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[81] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):717–729, 2016.

[82] Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):914–921, Apr. 2020.

[83] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar. 2016.

[84] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet

and the impact of residual connections on learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017.

[85] Yue Tan, Yixin Liu, Guodong Long, Jing Jiang, Qinghua Lu, and Chengqi Zhang. Federated learning on non-iid graphs via structural knowledge sharing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9953–9961, Jun. 2023.

[86] Artur Trindade. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C58C86.

[87] Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar. 2016.

[88] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R. Zaiane. Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2441–2449, Jun. 2022.

[89] Jianyi Wang, Kelvin C.K. Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2555–2563, Jun. 2023.

[90] Jingyuan Wang, Jiawei Jiang, Wenjun Jiang, Chao Li, and Wayne Xin Zhao. Libcity: An open library for traffic prediction. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '21, page 145–148, New York, NY, USA, 2021. Association for Computing Machinery.

[91] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3796–3805, 2017.

[92] Tao Wang, Kaihao Zhang, Tianrun Shen, Wenhan Luo, Bjorn Stenger, and Tong Lu. Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3):2654–2662, Jun. 2023.

[93] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. arXiv preprint arXiv:1808.04560, 2018. Submitted on 14 Aug 2018.

[94] Jun Wei, Shuhui Wang, and Qingming Huang. F^{3}net: Fusion, feedback and focus for salient object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12321–12328, Apr. 2020.

[95] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. Session-based recommendation with graph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):346–353, Jul. 2019.

[96] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017. Submitted on 25 Aug 2017, last revised 15 Sep 2017.

[97] Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12549–12556, Apr. 2020.

[98] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[99] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3166–3173, 2013.

[100] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J. Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, Yuqiang Zheng, Yanyun Qu, Yuhong Xie, Liang Chen, Zhonghao Li, Chen Hong, Hao Jiang, Siyuan Yang, Yan Liu, Xiaochao Qu, Pengfei Wan, Shuai Zheng, Minhui Zhong, Taiyi Su, Lingzhi He, Yandong Guo, Yao Zhao, Zhenfeng Zhu, Jinxiu Liang, Jingwen Wang, Tianyi Chen, Yuhui Quan, Yong Xu, Bo Liu, Xin Liu, Qi Sun, Tingyu Lin, Xiaochuan Li, Feng Lu, Lin Gu, Shengdi Zhou, Cong Cao, Shifeng Zhang, Cheng Chi, Chubing Zhuang, Zhen Lei, Stan Z. Li, Shizheng Wang, Ruizhe Liu, Dong Yi, Zheming Zuo, Jianning Chi, Huan Wang, Kai Wang, Yixiu Liu, Xingyu Gao, Zhenyu Chen, Chang Guo, Yongzhou Li, Huicai Zhong, Jing Huang, Heng Guo, Jianfei Yang, Wenjuan Liao, Jiangang Yang, Liguo Zhou, Mingyue Feng, and Likun Qin. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020.

[101] Xiaocheng Yang, Mingyu Yan, Shirui Pan, Xiaochun Ye, and Dongrui Fan. Simple and efficient heterogeneous graph neural network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):10816–10824, Jun. 2023.

[102] Xue Yang, Junchi Yan, Ziming Feng, and Tao He. R3det: Refined single-stage detector with feature refinement for rotating object. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3163–3171, May 2021.

[103] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7370–7377, Jul. 2019.

[104] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference

over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 02 2014.

[105] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017.

[106] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11121–11128, Jun. 2023.

[107] Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Fedala: Adaptive local aggregation for personalized federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11237–11244, Jun. 2023.

[108] Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017.

[109] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. An end-to-end deep learning architecture for graph classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[110] Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1234–1241, Apr. 2020.

[111] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015.

[112] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12993–13000, Apr. 2020.

[113] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13001–13008, Apr. 2020.

[114] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115, May 2021.

[115] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13041–13049, Apr. 2020.

A Search Queries and Results

The following queries were executed on 23 April 2025.

- 15 year range:

```
SRCTITLE ( aaai AND conference AND on AND
    artificial AND intelligence ) AND
    PUBYEAR > 2009 AND PUBYEAR < 2025
```

- 5 year range:

```
SRCTITLE ( aaai AND conference AND on AND
    artificial AND intelligence ) AND
    PUBYEAR > 2019 AND PUBYEAR < 2025
```

- 2 year range:

```
SRCTITLE ( aaai AND conference AND on AND
    artificial AND intelligence ) AND
    PUBYEAR > 2022 AND PUBYEAR < 2025
```

The results of the 15 year range query can be found in Table 2, of the 5 year range query in Table 3, and of the 2 year range query in Table 4.

B Datasets Analyzed per Period

A list of the analyzed datasets from the 15 year period can be found in Table 5, from the 5 year period in Table 6, and from the 2 year period in Table 7.

C Dataset Evaluation Criteria

The following are the evaluation criteria used to evaluate the datasets in step 3. These descriptions were taken from a shared annotation schema from the team.

Each criterion has an "Unsure", "No information" and "Not applicable" option, unless otherwise stated. "Unsure" signifies the entry is marked for discussion. "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 that the answer to the question is no (e.g. it was stated that no prescreening was done).

- **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
- **Notes** – are there any additional notes to be worth mentioning about the dataset?
- **Additional links** – are there any additional relevant links the dataset provides?

D Raw Dataset Evaluation Data

The raw data of the evaluation of the datasets regardless of period is shown in Table 8. The raw data for the evaluation of the datasets in the 15 year period is shown in Table 9, for the 5 year period in Table 10, and for the 2 year period in Table 11.

Paper title	Year	Citation count
Attention based spatial-temporal graph convolutional networks for traffic flow forecasting [33]	2019	2246
Regularized evolution for image classifier architecture search [72]	2019	2043
Building end-To-end dialogue systems using generative hierarchical neural network models [78]	2016	1258
FFA-Net: Feature fusion attention network for single image dehazing [71]	2020	1272
Deep spatio-temporal residual networks for citywide crowd flows prediction [108]	2017	1787
Distance-IoU loss: Faster and better learning for bounding box regression [112]	2020	3534
Convolutional 2D knowledge graph embeddings [20]	2018	2228
CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison [38]	2019	1651
Inception-v4, inception-ResNet and the impact of residual connections on learning [84]	2017	9648
Spatial temporal graph convolutional networks for skeleton-based action recognition [98]	2018	3458
Counterfactual multi-agent policy gradients [26]	2018	1428
SeqGAN: Sequence generative adversarial nets with policy gradient [105]	2017	1695
Session-based recommendation with graph neural networks [95]	2019	1487
Deeper insights into graph convolutional networks for semi-supervised learning [52]	2018	2136
Anchors: High-precision model-agnostic explanations [73]	2018	1527
Hypergraph neural networks [25]	2019	1180
An end-to-end deep learning architecture for graph classification [109]	2018	1264
Rainbow: Combining improvements in deep reinforcement learning [34]	2018	1298
Graph convolutional networks for text classification [103]	2019	1737
Return of frustratingly easy domain adaptation [83]	2016	1392
GMAN: A graph multi-attention network for traffic prediction [110]	2020	1235
Random erasing data augmentation [113]	2020	2311
Deep reinforcement learning with double Q-Learning [87]	2016	5484
Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting [114]	2021	3335
FiLM: Visual reasoning with a general conditioning layer [70]	2018	1238

Table 2: Resulting papers of the 15 year period query. The "Year" column denotes the year of publication.

Paper title	Year	Citation count
SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines [97]	2020	835
Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training [48]	2020	587
FFA-Net: Feature fusion attention network for single image dehazing [71]	2020	1272
PIQA: Reasoning about physical commonsense in natural language [8]	2020	653
Graph Neural Network-Based Anomaly Detection in Multivariate Time Series [17]	2021	815
R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object [102]	2021	723
EvolveGCN: Evolving graph convolutional networks for dynamic graphs [69]	2020	802
Distance-IoU loss: Faster and better learning for bounding box regression [112]	2020	3534
Are Transformers Effective for Time Series Forecasting? [106]	2023	1137
Spatial-Temporal Fusion Graph Neural Networks for Traffic Flow Forecasting [50]	2021	656
K-BERT: Enabling language representation with knowledge graph [60]	2020	614
Voxel R-CNN: Towards High Performance Voxel-based 3D Object Detection [19]	2021	651
Is BERT really robust? A strong baseline for natural language attack on text classification and entailment [43]	2020	746
Real-time scene text detection with differentiable binarization [58]	2020	659
Rumor detection on social media with bi-directional graph convolutional networks [7]	2020	579
Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting [82]	2020	1130
Unified vision-language pre-training for image captioning and VQA [115]	2020	661
F3Net: Fusion, feedback and focus for salient object detection [94]	2020	899
UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-Wise Perspective with Transformer [88]	2022	578
Measuring and relieving the over-smoothing problem for graph neural networks from the topological view [14]	2020	820
Improved knowledge distillation via teacher assistant [65]	2020	815
GMAN: A graph multi-attention network for traffic prediction [110]	2020	1235
Random erasing data augmentation [113]	2020	2311
TabNet: Attentive Interpretable Tabular Learning [1]	2021	812
Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting [114]	2021	3335

Table 3: Resulting papers of the 5 year period query. The "Year" column denotes the year of publication.

Paper title	Year	Citation count
Federated Learning on Non-IID Graphs via Structural Knowledge Sharing [85]	2023	92
BEVDepth: Acquisition of Reliable Depth for Multi-View 3D Object Detection [56]	2023	287
Underwater Ranker: Learn Which Is Better and How to Be Better [32]	2023	87
On the Effectiveness of Parameter-Efficient Fine-Tuning [27]	2023	89
NHITS: Neural Hierarchical Interpolation for Time Series Forecasting [13]	2023	212
BEVStereo: Enhancing Depth Estimation in Multi-View 3D Object Detection with Temporal Stereo [55]	2023	96
T2I-Adapter: Learning Adapters to Dig Out More Controllable Ability for Text-to-Image Diffusion Models [66]	2024	118
Are Transformers Effective for Time Series Forecasting? [106]	2023	1137
CLIP-ReID: Exploiting Vision-Language Model for Image Re-identification without Concrete Text Labels [53]	2023	101
PolarFormer: Multi-Camera 3D Object Detection with Polar Transformer [42]	2023	89
TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models [51]	2023	152
Learning Progressive Modality-Shared Transformers for Effective Visible-Infrared Person Re-identification [61]	2023	105
High-Resolution Iterative Feedback Network for Camouflaged Object Detection [36]	2023	86
Curriculum Temperature for Knowledge Distillation [57]	2023	103
PDFormer: Propagation Delay-Aware Dynamic Long-Range Transformer for Traffic Flow Prediction [40]	2023	192
Exploring CLIP for Assessing the Look and Feel of Images [89]	2023	222
Intriguing Findings of Frequency Selection for Image Deblurring [62]	2023	108
Simple and Efficient Heterogeneous Graph Neural Network [101]	2023	117
Ultra-High-Definition Low-Light Image Enhancement: A Benchmark and Transformer-Based Method [92]	2023	200
Benchmarking Large Language Models in Retrieval-Augmented Generation [15]	2024	125
Graph of Thoughts: Solving Elaborate Problems with Large Language Models [6]	2024	132
Spatio-Temporal Meta-Graph Learning for Traffic Forecasting [41]	2023	134
FedALA: Adaptive Local Aggregation for Personalized Federated Learning [107]	2023	172
Spatio-Temporal Self-Supervised Learning for Traffic Flow Prediction [39]	2023	131
FairFed: Enabling Group Fairness in Federated Learning [22]	2023	86

Table 4: Resulting papers of the 2 year period query. The "Year" column denotes the year of publication.

Dataset name	Citation sum
ImageNet [18]	11691
Arcade Learning Environment [5]	6782
PASCAL VOC 2007 [21]	5845
CIFAR-10 [46]	4354
COCO [59]	3534
Kinetics [45]	3458
NTU RGB+D [79]	3458
ECL / Electricity [86]	3335
ETT [114]	3335
US Weather [68]	3335
Cora [77]	3316
PubMed [77]	3316
CIFAR-100 [46]	2311
CUHK03 [54]	2311
DukeMTMC-reID [74]	2311
Fashion-MNIST [96]	2311
Market-1501 [111]	2311
PEMSD4 [33]	2246
PEMSD8 [33]	2246
Countries [9]	2228

Table 5: Datasets analyzed in the 15 year period. The "Citation sum" column refers to the citation sum value for each dataset with reference to the 15 year period.

Dataset name	Citation sum
PEMS03 [82]	1786
PEMS04 [82]	1786
PEMS07 [82]	1786
PEMS08 [82]	1786
ICDAR2015-Challenge-4 [44]	1382
Conceptual Captions [80]	1248
Flickr30k [104]	1248
ECL / Electricity [86]	1137
ETT [114]	1137
Exchange-Rate [47]	1137
ILI [12]	1137
Jena Weather [63]	1137
Traffic	1137
DUT-OMRON [99]	899
DUTS [91]	899
ECSSD [81]	899
HKU-IS [49]	899
GOT-10k [37]	835
ILSVRC 2014 [75]	835
LaSOT [23]	835

Table 6: Datasets analyzed in the 5 year period. The "Citation sum" column refers to the citation sum value for each dataset with reference to the 5 year period. The "Traffic" dataset does not have a reference paper or source.

Dataset name	Citation sum
nuScenes [11]	472
LOL [93]	422
MIT-Adobe FiveK [10]	422
CIFAR-100 [46]	275
CLIP User Study [89]	222
KonIQ-10k [35]	222
LIVE In the Wild [31]	222
SPAQ [24]	222
COCO [59]	221
ECL / Electricity [86]	212
ETT [114]	212
Exchange-Rate [47]	212
ILI [12]	212
Jena Weather [63]	212
Traffic	212
DARK FACE [100]	200
UHD-LOL [92]	200
BIKECHI [90]	192
NYTaxi [90]	192
PEMS04 [82]	192

Table 7: Datasets analyzed in the 2 year period. The "Citation sum" column refers to the citation sum value for each dataset with reference to the 2 year period. The "Traffic" dataset does not have a reference paper or source.

Question	Yes	Partially	No information	No	Not applicable
1) Any available information about the dataset?	48	0	0	1	0
2) Was the outcome provided?	47	0	0	2	0
3) Did humans label the dataset?	27	2	4	16	0
4) Were the used labels original?	27	2	4	0	16
5) Was the label source provided?	18	0	0	15	16
6) Was prescreening or lack thereof stated?	4	0	0	26	19
7) Was the compensation method stated?	11	0	0	21	17
8) Was information about training provided?	6	0	0	26	17
9) Were formal instructions given?	10	7	14	1	17
10) Was there a reason provided for the labeller population?	5	0	0	27	17
11) Was the total number of labellers provided?	12	0	0	21	16
12) Was the number of labellers per item specified?	17	0	0	16	16
13) Was the label threshold provided?	17	0	0	16	16
14) Was there annotator overlap?	13	2	13	4	17
15) Was the synthesis type described?	12	0	0	14	23
16) Was there a discussion among the annotators?	2	0	22	3	22
17) Was IRR reported?	5	0	0	22	22
18) Was the population of the items described?	45	0	0	4	0
19) Was a reason given for choosing this item population?	21	0	0	28	0
20) Was the item source provided?	43	0	0	6	0
21) Was the sample size chosen before data collection?	23	0	23	1	2
22) Was a rationale given for the sample size?	8	0	0	39	2
23) Was the annotation schema created beforehand?	25	0	5	0	19
24) Was a reason given for annotation schema?	8	0	0	22	19
25) Link to the dataset available?	22	10	0	17	0

Table 8: Raw data for the results of the evaluation of all datasets. The first column represents the question that was asked for each dataset and the rest of the columns represent the number of datasets that answered the question with that answer. The total number of datasets evaluated is 49.

Question	Yes	Partially	No information	No	Not applicable
1) Any available information about the dataset?	19	0	0	1	0
2) Was the outcome provided?	18	0	0	2	0
3) Did humans label the dataset?	10	1	3	6	0
4) Were the used labels original?	11	0	3	0	6
5) Was the label source provided?	6	0	0	8	6
6) Was prescreening or lack thereof stated?	1	0	0	13	6
7) Was the compensation method stated?	5	0	0	9	6
8) Was information about training provided?	2	0	0	12	6
9) Were formal instructions given?	5	2	7	0	6
10) Was there a reason provided for the labeller population?	1	0	0	13	6
11) Was the total number of labellers provided?	2	0	0	12	6
12) Was the number of labellers per item specified?	5	0	0	9	6
13) Was the label threshold provided?	5	0	0	9	6
14) Was there annotator overlap?	3	1	7	3	6
15) Was the synthesis type described?	3	0	0	8	9
16) Was there a discussion among the annotators?	1	0	9	1	9
17) Was IRR reported?	0	0	0	11	9
18) Was the population of the items described?	17	0	0	3	0
19) Was a reason given for choosing this item population?	7	0	0	13	0
20) Was the item source provided?	17	0	0	3	0
21) Was the sample size chosen before data collection?	9	0	11	0	0
22) Was a rationale given for the sample size?	4	0	0	16	0
23) Was the annotation schema created beforehand?	9	0	4	0	7
24) Was a reason given for annotation schema?	4	0	0	9	7
25) Link to the dataset available?	8	5	0	7	0

Table 9: Raw data for the results of the evaluation of the datasets from the 15 year period. The first column represents the question that was asked for each dataset and the rest of the columns represent the number of datasets that answered the question with that answer. The total number of datasets evaluated is 20.

Question	Yes	Partially	No information	No	Not applicable
1) Any available information about the dataset?	20	0	0	0	0
2) Was the outcome provided?	19	0	0	1	0
3) Did humans label the dataset?	10	0	2	8	0
4) Were the used labels original?	8	2	2	0	8
5) Was the label source provided?	7	0	0	5	8
6) Was prescreening or lack thereof stated?	1	0	0	9	10
7) Was the compensation method stated?	3	0	0	9	8
8) Was information about training provided?	2	0	0	10	8
9) Were formal instructions given?	2	3	7	0	8
10) Was there a reason provided for the labeller population?	2	0	0	10	8
11) Was the total number of labellers provided?	6	0	0	6	8
12) Was the number of labellers per item specified?	7	0	0	5	8
13) Was the label threshold provided?	7	0	0	5	8
14) Was there annotator overlap?	5	1	5	1	8
15) Was the synthesis type described?	5	0	0	5	10
16) Was there a discussion among the annotators?	1	0	9	1	9
17) Was IRR reported?	2	0	0	9	9
18) Was the population of the items described?	18	0	0	2	0
19) Was a reason given for choosing this item population?	7	0	0	13	0
20) Was the item source provided?	17	0	0	3	0
21) Was the sample size chosen before data collection?	9	0	8	1	2
22) Was a rationale given for the sample size?	3	0	0	15	2
23) Was the annotation schema created beforehand?	10	0	2	0	8
24) Was a reason given for annotation schema?	4	0	0	8	8
25) Link to the dataset available?	9	4	0	7	0

Table 10: Raw data for the results of the evaluation of the datasets from the 5 year period. The first column represents the question that was asked for each dataset and the rest of the columns represent the number of datasets that answered the question with that answer. The total number of datasets evaluated is 20.

Question	Yes	Partially	No information	No	Not applicable
1) Any available information about the dataset?	20	0	0	0	0
2) Was the outcome provided?	19	0	0	1	0
3) Did humans label the dataset?	10	1	2	7	0
4) Were the used labels original?	11	0	2	0	7
5) Was the label source provided?	8	0	0	5	7
6) Was prescreening or lack thereof stated?	3	0	0	8	9
7) Was the compensation method stated?	6	0	0	6	8
8) Was information about training provided?	4	0	0	8	8
9) Were formal instructions given?	6	2	3	1	8
10) Was there a reason provided for the labeller population?	3	0	0	9	8
11) Was the total number of labellers provided?	5	0	0	8	7
12) Was the number of labellers per item specified?	8	0	0	5	7
13) Was the label threshold provided?	7	0	0	6	7
14) Was there annotator overlap?	6	0	4	2	8
15) Was the synthesis type described?	5	0	0	4	11
16) Was there a discussion among the annotators?	0	0	7	2	11
17) Was IRR reported?	3	0	0	6	11
18) Was the population of the items described?	19	0	0	1	0
19) Was a reason given for choosing this item population?	11	0	0	9	0
20) Was the item source provided?	17	0	0	3	0
21) Was the sample size chosen before data collection?	7	0	11	0	2
22) Was a rationale given for the sample size?	1	0	0	17	2
23) Was the annotation schema created beforehand?	9	0	2	0	9
24) Was a reason given for annotation schema?	2	0	0	9	9
25) Link to the dataset available?	13	1	0	6	0

Table 11: Raw data for the results of the evaluation of the datasets from the 2 year period. The first column represents the question that was asked for each dataset and the rest of the columns represent the number of datasets that answered the question with that answer. The total number of datasets evaluated is 20.