UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
**ICT International Doctoral School**

# A SEMANTICS-DRIVEN METHODOLOGY FOR HIGH-QUALITY IMAGE ANNOTATION

Xiaolei Diao

Ph.D. Degree Thesis

Academic Year 2020–2024

Advisor

Prof. Fausto Giunchiglia

Università degli Studi di Trento

January 2024

# Acknowledgements

First of all, I wish to extend my appreciation and thanks to my advisor, Prof. Fausto Giunchiglia. Thanks for providing me with such an exceptional opportunity to study and work closely with him. I learned a lot from him, including the methodology for doing research, writing papers, and constructing social networks, which inspired me a lot at the beginning of being a young researcher. Meanwhile, he encouraged me and trusted me a lot during the past four years, I could not finish my Ph.D. study without his help and encouragement. I would say working with him really broadened my eyes and constructed the foundation of my future research career. Also thanks to "DELPhi - DiscovEring Life Patterns" project and European Union's Horizon 2020 FET Proactive project "WeNet – The Internet of us" that underpinned our research.

Additionally, my deepest thanks to all the other professors and my colleagues for their invaluable help and support. I am immensely thankful for the opportunity to be involved in the Knowlive group, which allows me to connect with passionate colleagues from diverse backgrounds. My colleagues also help me a lot. For example, Mayuhk Bagchi introduces the canons in the knowledge area, and Andrea Bontempelli helps me understand the "towards vision semantic" project. I also need to thank Prof. Massimo Poesio. During my visiting period, he helps me with both research and life. He also gave me many valuable suggestions for my thesis. Last but certainly not least, please allow me to convey my sincerest gratitude and love to my husband Daqian Shi, my mother Zhihong Lyu, and my father Zengyong Diao for their companionship and support throughout my doctoral journey.

# Abstract

In recent years, the field of computer vision has achieved significant advancements, largely driven by high-quality datasets. However, current models struggle to provide the responses that humans expect, especially when addressing complex computer vision tasks. One primary reason for this limitation lies in the systemic flaws inside ground truth object recognition benchmark datasets. Our basic tenet is that these flaws are rooted in the many-to-many mappings which exist between the visual information encoded in images and the intended semantics of the labels annotating them. The net consequence is that the current annotation process is largely under-specified, thus leaving too much freedom to the subjective judgment of annotators. In this study, we propose an integrated methodology for constructing image datasets, referred to as `vTelos`, which integrates natural language processing, knowledge representation, and computer vision techniques. The primary objective of `vTelos` is to make explicit the intended annotation semantics, thus minimizing the number and role of subjective choices. The methodology introduces two key roles, the Classificationist and the Classifier. The Classificationist ensures semantic consistency within the dataset by defining a classification hierarchy and visual properties. The Classifier, on the other hand, performs image annotation tasks based on the hierarchy outlined by the Classificationist and iterative refinement processes. For large-scale datasets, the methodology incorporates crowdsourcing techniques to enhance annotation efficiency while maintaining high quality. To validate the feasibility and effectiveness of the proposed methodology, we constructed an image dataset named vTelos-img, under the guidance of `vTelos`. This dataset was subjected to a multidimensional evaluation framework, including quality assessment, annotation efficiency analysis, model performance comparison, and ablation studies of core components. The results comprehensively demonstrate

the scientific rigor and practical utility of the vTelos methodology. Experimental findings reveal that vTelos significantly improves dataset quality, optimizes annotation workflows, and enhances the performance of machine learning models. This study provides a robust foundation for future data-driven computer vision tasks and offers a novel perspective on high-quality dataset construction and annotation methodologies.

**Keywords:** Computer Vision, Object Recognition, Image Dataset Construction, Visual Properties

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 The Context

In recent years, computer vision has made significant advancements, improving everyday life through numerous practical applications. Machine learning (ML) models are inherently data-driven, with their performance highly dependent on the quality of the datasets used for training and evaluation. The quality and diversity of data directly determine a model's generalization capability and robustness [111, 159]. In modern ML systems, data is not only a core component of the infrastructure but also a key determinant of system effectiveness and applicability [57]. For instance, large-scale datasets like ImageNet [26] have become fundamental to advancing deep learning, especially in computer vision tasks, where they have played a critical role in optimizing convolutional neural network (CNN) architectures.

The degradation of data quality can lead to what is known as the "data cascade" effect [124]. This phenomenon refers to systemic issues caused by data quality defects that progressively amplify errors in ML models and undermine their performance. For example, in autonomous driving systems, inaccurate annotations of traffic signs in training datasets may lead to misclassification by vehicles, posing significant safety risks [160]. Data quality can be influenced by various factors, such as imbalanced class distributions, annotation errors, ge-

ographic or cultural biases, and distribution mismatches between training and testing datasets. These issues not only reduce the reliability of models but also limit their applicability in real-world scenarios. Sambasivan et al. [124] explored the downstream impact of data quality issues and found that data cascades frequently occur in tasks where data quality is underestimated. Early intervention, particularly during data collection and labeling, is crucial for mitigating these issues and improving performance across various tasks.

The growing emphasis on data-centric AI [73, 134] has heightened awareness of the critical role of data quality in determining model performance across diverse tasks, such as object recognition. Recent studies have reported various systemic flaws in the development of benchmark datasets for object recognition. Critical analyses [144, 145, 27, 48] of the construction process of benchmark image datasets, exemplified by ImageNet [26], have revealed flaws that could be a crucial challenge in impeding the continued progress of object recognition. For example, while ImageNet is widely used as a benchmark for object recognition tasks, it exhibits significant geographic and cultural biases. Shreya et al. [126] critically analyzed the geographic diversity of ImageNet [26] and OpenImages [82], revealing their Amerocentric and Eurocentric representational biases, where images from Western regions dominate, leaving data from other regions underrepresented.

In addition, annotation errors are another prevalent issue. Studies have shown that in ImageNet, certain categories have annotation error rates exceeding 5%, introducing substantial biases in model evaluation. Dimitris et al. conducted an extensive analysis of annotation errors in ImageNet [26], identifying that noisy data collection pipelines contribute to systemic misalignments between benchmark datasets and real-world tasks. Lucas et al. [6] described inherent biases in the annotation pipeline used to construct ImageNet. Terrance et. al. [24] report significant discrepancies in the classification accuracy of six *'in production'* object recognition systems, and ties the discrepancies to diversity

in *socio-economic status*, *culture* and *language* from where the images were sourced. These challenges are particularly pronounced in cross-lingual and cross-cultural contexts. For instance, specific cultural symbols may have distinct semantic interpretations in different regions, further complicating model development.

As modern AI systems demand increasingly high performance, constructing high-quality datasets by leveraging existing knowledge has become an attractive research direction. For example, integrating natural language processing (NLP) techniques with knowledge graphs can provide richer contextual information for dataset annotations, reducing annotation ambiguity. Our research focuses on addressing challenges in image datasets. The goal is to develop a systematic approach to constructing high-quality image datasets by integrating theories from computer vision, NLP, and knowledge representation. Through these efforts, we aim not only to improve the performance of current ML models but also to provide a robust data foundation for future, more complex, and diverse application scenarios.

## 1.2 The Problem

As a data-driven discipline, the fields of Machine Learning (ML) and Computer Vision (CV) increasingly emphasize the critical importance of data quality. Numerous studies have begun investigating the systemic flaws inherent in the development of image classification benchmark datasets. These datasets, such as ImageNet [26] and OpenImages [82], have long been regarded as the standard for training and evaluating ML models, particularly in object recognition and classification tasks. However, recent analyses have highlighted several significant issues, which were introduced in the previous section. These flaws not only reduce the quality of the datasets but also limit their generalizability across diverse application scenarios.

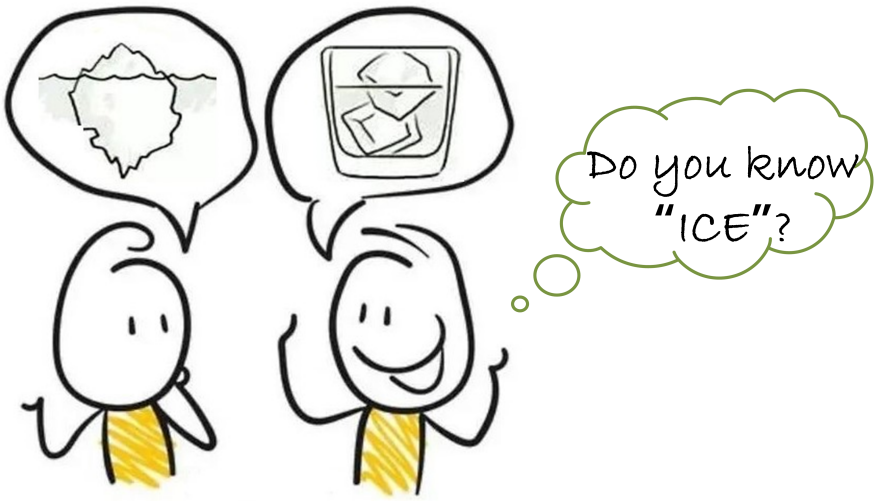


Figure 1.1: An interesting case of the semantic gap problem.

Our core principle is that these flaws originate from the interaction between perception and language. An early version of this problem was identified in [135] as the Semantic Gap Problem (SGP), described as "the lack of consistency between the information extracted from visual data and the interpretation of the same data by a user in a given context". The SGP manifests in multiple ways. For instance, individuals from different cultural or educational backgrounds may interpret the same visual data differently due to variations in their prior knowledge or contextual understanding. In real-world scenarios, this can lead to significant discrepancies in how visual information is described linguistically. This problem remains unresolved in the process of constructing image datasets and was formalized in [44] as the existence of a '"many-to-many mapping" between the *visual information* encoded in an image and the *intended semantics* of its linguistic description. For example, as shown in Figure 1(a), an image might be labeled as "guitar", but depending on its visual appearance, it could also be interpreted as a "ukulele" or "bass". Similarly, in Figures 1(b) and 1(c), the label "apple" could ambiguously refer to either the fruit or the brand. The existence of such many-to-many mappings introduces considerable ambiguity, making it challenging to establish a consistent relationship between

visual representations and linguistic labels in datasets.

The presence of the semantic gap renders the current annotation process largely under-specified, granting annotators excessive freedom in subjective judgments. This lack of standardization often results in annotators selecting a single mapping from the numerous possible SGP mappings based on their personal preferences or biases. As a result, annotation consistency is compromised, the accuracy of collected labels is reduced, and the overall quality of the dataset is negatively impacted. For example, two annotators might assign different labels to the same object due to differing interpretations of its features or context, leading to inconsistencies within the dataset. This issue is particularly critical for datasets intended for large-scale ML tasks, where even minor inconsistencies can propagate errors and significantly degrade the performance of downstream models.

Based on the above observations, the main research problems we identify in this thesis are as follows:

- Comprehensive evaluation and analysis of existing visual datasets: Current visual datasets exhibit numerous quality issues, including annotation errors, semantic ambiguities, and geographic or cultural biases. These problems are closely related to the semantic gap (SGP) and require a thorough and systematic evaluation to identify their root causes and impacts.

- Proposing a suitable dataset construction methodology: To address the semantic gap, it is essential to develop a systematic methodology that ensures semantic consistency between visual information and linguistic descriptions. This methodology must minimize ambiguities caused by many-to-many mappings and improve the overall accuracy of annotations.

- Developing standardized annotation guidelines: The lack of standardization in the current annotation process leads to significant discrepancies among annotators, which undermines annotation consistency and dataset

quality. Establishing a set of standardized annotation guidelines is crucial to improving the accuracy, consistency, and reliability of the annotation process.

- Designing effective evaluation schemes: A comprehensive evaluation framework is needed to validate the proposed dataset construction methodology. This includes experimentally assessing the quality of the dataset, annotation accuracy, and its applicability to various tasks. Such an evaluation framework will provide scientific evidence for optimizing dataset design.

## 1.3 The Solution

Conducting a comprehensive evaluation and analysis of existing visual datasets, our goal is to propose a methodology that integrates the technologies in the field of Natural Language Processing (NLP), Knowledge Representation (KR), and Computer Vision (CV), referred to as `vTelos`[1]. This methodology aims to generate high-quality ground truth datasets and serves as a first step toward a general solution to the Semantic Gap Problem (SGP). The intuition is that the current annotation process is largely under-specified, granting annotators excessive freedom to subjectively select a mapping among many possible SGP interpretations. Our idea is to leverage KR to provide guidelines that standardize the annotation process, thereby indirectly embedding intended semantics into ML models, encoded in an appropriate natural language format.

Our objective is to propose a general method for building an integrated hierarchical structure of classification and material concepts, where classification concepts are annotated using their corresponding material concepts' visual representations (e.g., images or videos). Unlike previous approaches that build hierarchical structures based on linguistic definitions, this methodology organizes

[1] `vTelos` originates from the Greek word `Telos`, meaning goal or purpose, where `v` stands for visual, indicating visual purpose.

concepts by their visual attributes. The properties of classification concepts are utilized to describe the core visual characteristics of material concepts. The key advantage of this approach lies in defining a clear and standardized set of rules for constructing high-quality classification hierarchies. These rules ensure that visual properties and linguistic definitions are effectively encoded and consistently expressed within the hierarchy. Based on this methodology, we align the classification process of visual datasets with the core visual properties of substance concepts, enabling the creation of high-quality image datasets that are intuitively aligned with human visual perception.

In constructing large-scale visual datasets, crowdsourcing is commonly adopted as a cost-effective means for annotation. However, based on the existence of the SGP, the crowdsourcing process often faces challenges, such as inconsistencies in annotators' understanding and subjective judgments, which directly impact dataset quality and annotation uniformity. To address these challenges and effectively apply the proposed `vTelos` methodology to real-world annotation tasks, we developed a standardized annotation guideline. This guideline clearly defines annotation rules and procedures, ensuring the standardization and consistency of the annotation process while reducing costs and improving data collection efficiency. Additionally, we conducted a series of targeted crowdsourcing experiments leveraging crowdsourcing platforms to evaluate the guideline's effectiveness in ensuring annotation consistency and efficiency during practical operations.

To thoroughly verify the quality and validity of the constructed dataset, we designed a multidimensional evaluation framework that includes both quantitative and qualitative analyses. First, we evaluated annotation accuracy and consistency through intra-class consistency tests, ensuring that samples within the same category maintain uniform labels. Second, we employed deep learning models to train and test the dataset, validating its effectiveness and performance in practical tasks. Finally, we applied label noise detection techniques to iden-

tify and eliminate annotation errors that might occur during the process, further enhancing the reliability of the dataset. Through this comprehensive evaluation framework, we demonstrated the practical applicability and effectiveness of the proposed method for building high-quality visual datasets.

Corresponding to the problems, we conclude the main contributions of this thesis as follows:

- We propose a systematic dataset construction methodology (`vTelos`), which integrates theories from natural language processing, knowledge representation, and computer vision, to address the Semantic Gap Problem (SGP) and enable the creation of high-quality datasets.
- A set of standardized annotation guidelines is constructed, including definitions of image dataset categories and their visual properties, along with an annotation process based on the defined visual properties. This aims to improve annotation consistency and dataset quality, especially for addressing challenges in annotation tasks caused by subjective biases and inconsistencies.
- We design and implement a comprehensive evaluation of the dataset constructed with the proposed methodology. The evaluation includes intra-class consistency tests, label noise detection, and deep learning model validation to systematically verify the effectiveness and quality of the constructed datasets.

## 1.4 Structure of the Thesis

In Chapter 2, we provide a comprehensive review of four key topics related to this research. First, we discuss the current state of image benchmarks and datasets, including methods for constructing image datasets and the challenges and limitations of existing datasets. Next, we present advancements in image

classification tasks and their corresponding methods, covering areas such as object recognition, zero-shot image classification, and fine-grained image classification. We also delve into categorization methodologies, starting with faceted classification and exploring classical classification taxonomy. Finally, we examine the role of crowdsourcing in dataset construction and annotation, including an introduction to crowdsourcing, the design of effective annotation tasks, and evaluation metrics for assessing the quality of crowdsourced contributions. This chapter lays the foundation for understanding the fundamental methods and challenges addressed in this research.

Chapter 3 focuses on the analysis of image datasets and presents the core problem statement of this research. We begin by outlining the importance of understanding dataset-related issues in machine learning and computer vision tasks. Next, we explore the primary sources of annotation mistakes, including mislabeled images, challenges associated with multi-object images, and issues specific to single-object images. We then introduce the semantic gap problem, discussing the disconnect between visual data and their semantic interpretations. This chapter establishes the foundation for the proposed image construction methodology in the following chapters.

In Chapter 4, we introduce `vTelos`, the proposed image dataset construction methodology, and provide a detailed explanation of its components and implementation. The chapter begins with an introduction to the methodology and its significance in addressing the identified challenges. We discuss the four key annotation design choices that guide the construction process, ensuring consistency and accuracy. Besides, we explore the role of annotators in the methodology, differentiating between the responsibilities of the "Classificationist", who defines the label space and its semantic meaning, and the "Classifier", who applies these labels to images. Then, we present the overall methodology, integrating the design choices and annotator roles into a cohesive framework. This chapter is based on the following publication:

- Fausto Giunchiglia, Mayukh Bagchi, **Xiaolei Diao***. A semantics-driven methodology for high-quality image annotation[C]. Proceedings of the 26th European Conference on Artificial Intelligence (ECAI), 2023.

- **Xiaolei Diao**, Fausto Giunchiglia. *Building a Visual Semantics Aware Object Hierarchy*[C]. Doctoral Consortium, Proceedings of the 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence (IJCAI), 2022.

Chapter 5 centers on the role of the *Classificationist* in defining and organizing categories for constructing high-quality image datasets. The chapter begins by introducing the importance of precise and systematic category definitions, which are essential for ensuring semantic consistency across the dataset. It discusses the use of *Cannons* as foundational principles to guide the standardization of category definitions. The chapter further explores *Glosses*, which provides clear linguistic descriptions that encapsulate the intended semantics of each category. Additionally, it examines the visual properties of categories, focusing on the concepts of *Genus*, which represents shared features among categories, and *Differentia*, which highlights distinguishing attributes. These properties are then utilized to construct a hierarchical structure that systematically organizes categories and their relationships. The chapter concludes by summarizing the pivotal role of the "Classificationist" in establishing a robust framework for category definition, which is vital for creating datasets with high semantic integrity and applicability. Chapter 5 is based on the following publications:

- Fausto Giunchiglia, Mayukh Bagchi, **Xiaolei Diao**. *Aligning Visual and Lexical Semantics*[C]. Normality, Virtuality, Physicality and Inclusivity: 18th International Conference, iConference 2023.

- **Xiaolei Diao**, Fausto Giunchiglia. *A Visual Semantics Aware Object Hierarchy for Image Dataset Construction*[J]. Artificial Intelligence. (Under-

going)

Chapter 6 focuses on the role of the *Classifier* and highlights the contributions of crowdsourcing to the annotation process and its integration into the overall methodology for constructing robust image datasets. The chapter begins by introducing the importance of crowdsourcing as an effective method for scaling annotation efforts while maintaining quality and consistency. It then discusses the design of crowdsourcing tasks, including the preparation of labels to ensure clarity and semantic accuracy, and the preparation of images to optimize their suitability for annotation. The chapter also presents a series of crowdsourcing experiments, designed to evaluate the effectiveness of the proposed task designs in achieving high-quality annotations efficiently. This chapter is based on the publication:

- Fausto Giunchiglia, **Xiaolei Diao***, Mayukh Bagchi. *Incremental Image Labeling via Iterative Refinement*[C]. IWCIM 2023: The Eleventh International Workshop on Computational Intelligence for Multimedia Understanding, Satellite Workshop of ICASSP 2023.

- **Xiaolei Diao**, Fausto Giunchiglia. *Crowdsourcing Image Annotation by Visual Properties*[C]. Annual Meeting of the Cognitive Science Society, CogSci 2025. (Under Reviewing)

Chapter 7 focuses on the evaluation and application of the proposed *vTelos* methodology. It begins with an introduction to the evaluation framework and its significance in validating the effectiveness of the methodology. The chapter then delves into quality evaluation, examining key metrics such as inter-annotator agreement to measure consistency, annotation cost to assess efficiency, machine learning accuracy to evaluate dataset performance, and an ablation study to analyze the contributions of individual components. Following this, the chapter discusses the automated construction and evaluation of hierarchies built using machine learning techniques. The practical applications of *vTelos* in

computer vision are also explored, including its impact on tasks such as object recognition, fine-grained image classification, and zero-shot image classification. The content of this chapter is grounded in research from the following publications:

- Zhancheng Ren, Qiang Shen, **Xiaolei Diao**, Hao Xu. *A sentiment-aware deep learning approach for personality detection from text*[J]. Information Processing and Management, 2021, 58(3): 102532.
- Yang Chi, Fausto Giunchiglia, Daqian Shi, **Xiaolei Diao**, Chuntao Li, Hao Xu, 2022. *ZiNet: Linking Chinese Characters Spanning Three Thousand Years [C].* Findings of the Association for Computational Linguistics: ACL 2022.
- Jian Li, Ziyao Meng, Daqian Shi, Rui Song, Xiaolei Diao, Jingwen Wang, and Hao Xu, 2023. *FCC: Feature Clusters Compression for Long-Tailed Visual Recognition [C].* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- Luca Erculiani, Andrea Bontempelli, Andrea Passerini, **Xiaolei Diao**, Fausto Giunchiglia. *Egocentric Hierarchical Visual Semantics*[J]. Journal of Artificial Intelligence Research (JAIR). (Under Reviewing)

Finally, Chapter 8 presents the conclusions and future work respectively.

# Chapter 2

# State of the Art

In the computer vision area, research on image processing and classification has made rapid advancements, particularly with the rise of deep learning, which has significantly enhanced the performance of computer vision systems. Image processing and classification are fundamental research problems in computer vision, with widespread applications in autonomous driving, medical image analysis, intelligent surveillance, and many other areas. As an important task of artificial intelligence systems to understand the visual world, the development of image classification has not only driven academic progress but also had a profound impact on industrial applications. Therefore, understanding image classification and its related research progress is crucial in the field of computer vision. This chapter will focus on four main aspects: image datasets, image classification, categorization methods, and crowdsourced annotation, discussing current research achievements, existing issues, and relevant discussions. By discussing these aspects, we aim to reveal the challenges and future directions in the field of image processing and classification, providing a comprehensive theoretical foundation and reference.

## 2.1 Image Datasets

Image datasets are a fundamental cornerstone in the fields of computer vision and deep learning. Many classic computer vision tasks, such as image classification and object recognition, rely heavily on the construction of high-quality image datasets. With the advancement of deep learning, data-driven approaches have become one of the primary drivers of progress in computer vision, and high-quality datasets are crucial for both model training and evaluation. Image datasets are not only foundational resources for training models but also serve as benchmarks for model evaluation. Through these datasets, researchers can objectively and systematically assess model performance. The release and application of datasets such as ImageNet [26], COCO [93], and PASCAL VOC [32] have significantly advanced research on tasks like image classification, object detection, and semantic segmentation. The construction of such datasets often requires substantial resources, including manual annotation, data collection, data cleaning, and validation [123]. The rigor of these processes directly determines the quality of the dataset, which in turn affects the performance of models trained on them. Therefore, the process of constructing image datasets and ensuring their quality control is a core issue in computer vision research. Building representative, high-quality datasets remains a key focus for both academia and industry.

### 2.1.1 Methods of image dataset construction

The construction of image datasets is a fundamental aspect of computer vision research. Common methods include manual annotation, automated data collection, and crowdsourcing. These methods are described in detail below.

**Manual Annotation.**

Manual annotation by humans is one of the most common and critical methods for constructing image datasets. This approach ensures high-quality annotations but requires significant human effort and time, especially for tasks demanding high precision and fine-grained classifications, such as large-scale image classification, object detection, and semantic segmentation. Manual annotation is widely used to build complex and fine-grained datasets. In the construction of the object detection dataset, by combining automated tools with manual annotation, the quality of annotations for complex images is ensured. Annotators are tasked with accurately outlining object boundaries within images and verifying annotation consistency [93]. To ensure quality, multi-stage validation is typically employed, where multiple annotators annotate the same image to achieve accuracy and consistency [32, 93, 123, 1, 82, 98, 81].

For general image classification tasks, the primary advantages of manual annotation are its high precision and flexibility, as it can be adapted to different research objectives and category definitions. However, manual annotation also faces challenges, including high costs and extended time requirements, particularly for large-scale image datasets [140]. Such annotation tasks are often carried out on crowdsourcing platforms [26]. In some cases, the annotation process requires domain-specific expertise, such as in fine-grained classification tasks [66, 150, 83, 100] and the construction of medical imaging datasets [106, 133, 92]. These tasks often demand a high level of expertise from annotators, and annotations are typically completed by domain experts to ensure accuracy. For instance, in medical imaging datasets, annotators with medical backgrounds are required to correctly label disease features or anatomical structures [88, 153, 62, 97].

**Automated Data Collection.**

Automated data collection involves using tools such as web crawlers to gather images from the internet, combined with techniques like natural language processing and text analysis to generate initial labels. One major advantage of automated data collection is its efficiency. It allows the rapid acquisition of large volumes of images from the internet, significantly reducing the need for manual intervention. However, a core challenge of this method lies in ensuring the diversity of images and the consistency of annotations [5]. Image data from different sources may contain substantial noise and redundant information, necessitating data cleaning and preprocessing steps to remove invalid data and ensure that models are trained on high-quality datasets [83, 41].

Additionally, automated annotation using NLP to generate image descriptions can effectively reduce labeling costs. However, the accuracy of such annotations largely depends on the quality of the descriptive text. If the textual information is ambiguous or noisy, the generated labels may be inaccurate, thereby affecting model performance [5]. Therefore, while automated data collection offers significant efficiency advantages, its strict requirements for data quality present considerable challenges.

To improve the quality of automated annotations, researchers have proposed various methods. For instance, semi-supervised learning techniques can effectively leverage noisy data and reduce reliance on precise labels [137, 157]. Online hard example mining [129] can help models focus on more challenging samples, improving robustness and generalization. Furthermore, techniques like compact bilinear pooling [41] enhance the efficiency of visual feature extraction, which is crucial for handling large-scale datasets.

Automated data collection also faces challenges related to insufficient data diversity. This is particularly problematic in specific domains, where obtaining representative, high-quality images can be difficult. To address this, researchers

have explored combining manual and automated annotation methods to improve overall dataset quality. For example, the COCO dataset [93] employed this approach to annotate complex objects and scenes, ensuring annotation accuracy and consistency. In the future, one of the important research directions will be how to integrate automated tools with human verification to further enhance the efficiency and quality of data annotation.

### 2.1.2 Discussions in current image datasets

Issues related to dataset quality have been widely recognized. In recent years, numerous studies have focused on addressing problems associated with dataset quality [144, 111]. We summarize the key dataset quality issues below.

**Label Disagreement.** Label disagreement [148], where different annotators assign inconsistent labels to the same image, is a common issue in image datasets. This problem is particularly prevalent in crowdsourced annotation scenarios. Research has shown that factors such as annotator subjectivity, the complexity of image content, and the ambiguity of task descriptions can contribute to label disagreement [110]. Label disagreement affects dataset quality by introducing inconsistent information into the training process, ultimately leading to reduced model generalization performance. To mitigate label disagreement, some studies have introduced multi-annotator strategies, which aggregate annotations from multiple annotators to improve label consistency and reliability [136]. Additionally, consistency metrics like Krippendorff's alpha [84] have been employed to evaluate annotation quality, helping to identify and reduce label disagreements [107, 71].

**Label Uncertainty.** Label uncertainty occurs when annotators are unable to confidently determine the most accurate label for certain images. This typically arises in cases where image content is ambiguous or open to multiple interpretations [64]. Label uncertainty lowers dataset quality and negatively impacts model training outcomes [37, 166]. To address label uncertainty, researchers

have proposed various methods, such as quantifying uncertainty and incorporating it into model training through weighted processing [112]. Other studies have explored leveraging model-predicted uncertainty to assist the annotation process, thereby reducing subjective bias from annotators [76]. Active learning [12] has proven to be an effective approach for minimizing label uncertainty. By allowing models to actively select samples with higher uncertainty for annotation, active learning can significantly reduce uncertainty and improve dataset quality [168].

Torralba and Efros [144] highlighted in early research that biases, and annotation errors in datasets are major factors affecting model performance. The work by [135], which identified the Semantic Gap Problem (SGP) as a key issue, is considered one of the primary reasons for dataset quality problems. This work broadly encompasses issues including label disagreement and label uncertainty [40, 138, 75]. To address these challenges, researchers have proposed various improvement methods. Paullada et al. [111] advocate more careful practices in the development of datasets that are attentive to their limitations and impact. Yun et al. [159] introduced a method that involves training classifiers on high-quality datasets to enhance the accuracy of ImageNet annotations. This approach leverages high-quality datasets to improve labeling precision. Yang et al. [156] proposed a fairness-oriented method aimed at balancing specific categories within the ImageNet dataset, emphasizing fairness and diversity in the dataset.

## 2.2 Categorization

For image datasets designed for tasks such as image classification and object recognition, categorization[102, 55] is a core task in the dataset construction process. Its goal is to divide data into distinct categories to facilitate model training and evaluation. Categorization tasks typically rely on clear classifi-

cation standards and high-quality annotated data. However, existing research highlights several challenges in the categorization process, such as the ambiguity in category definitions, semantic overlap between categories, and imbalanced data distributions [144, 125]. To address these issues, researchers have proposed various improvement methods. For instance, faceted classification introduces multiple independent classification dimensions, e.g., time, location, and topic, to provide a flexible structure for organizing complex data [60, 147]. Hierarchical categorization leverages the hierarchical relationships between categories to effectively manage multi-level labels and dependencies among categories [26, 130]. Furthermore, multimodal data fusion techniques, which integrate multiple modalities such as images and text, have been utilized to enhance classification accuracy and robustness [163, 3]. These methods provide a diverse set of tools for categorization tasks, significantly improving the quality and applicability of constructed datasets.

### 2.2.1 Faceted classification

Faceted Classification was first introduced in the field of library and information science, aiming to provide a flexible framework for organizing and retrieving complex information [118]. This method supports multidimensional descriptions of data by incorporating multiple independent classification dimensions. Faceted classification enables more precise descriptions of image content [61] and offers richer information compared to traditional single-dimensional classification when handling complex, multi-attribute objects [46].

In recent years, faceted classification has been widely adopted in computer science, particularly in dataset construction and management tasks that require organizing high-dimensional, multimodal data. Gnoli et al. [52] analyzed faceted classifications as linked data, proposing a framework for representing syntactic facets with RDF properties, enabling seamless integration with systems like SKOS. Bianchini and Bargioni [8] developed the CCLitBox tool,

which uses faceted classification to organize linked open data from Wikidata to automate the classification of literary works. Researchers have combined machine learning with faceted classification methods to further enhance classification accuracy and flexibility [14]. Kovashka et al. [80] proposed an interactive faceted classification system that allows users to optimize search results by providing preferences across multiple dimensions. Faceted classification is also applied in video representation learning. A recent study introduced the MUlti-Faceted Integration (MUFI) framework, which leverages facets from multiple datasets to improve video representation by embedding semantic information across facets, achieving significant performance gains in video recognition and captioning tasks [113].

Faceted classification is also applied in multiple area. It has been used to enhance stock trading models. Ensembles of faceted classification models trained on subsets of financial data (e.g., price trends, news sentiment) have shown improved robustness and profitability in generating trading signals [14]. In medical image retrieval, a faceted classification approach linked low-level image features with high-level textual features, significantly improving the accuracy of content-based image retrieval systems [149].

Another domain of application is image annotation. A robust multi-facet and structural knowledge approach exploited both labeled and unlabeled data to improve automatic Web image annotation, leveraging correlated and complementary facets to enhance annotation consistency and accuracy [69]. This integration of multiple facets within robust semi-supervised frameworks demonstrates the adaptability and effectiveness of faceted classification in handling noisy, multimodal data. These applications highlight the versatility of faceted classification in addressing diverse challenges across video, financial, medical, and multimedia domains, making it an indispensable tool for modern data-driven systems.

### 2.2.2 Hierarchical classification

Broadly speaking, methods that use the tree-like hierarchy of categories (category tree) to assist classification can be called hierarchical classification. Essentially, the biggest difference between the hierarchical classification method and the traditional classification method is: relying on the category tree, hierarchical classification resolves a global and large-scale classification problem into a series of local and small-scale classification problems. According to the way the category tree is generated, hierarchical classification can be divided into two categories, hierarchical classification based on predefined category trees and hierarchical classification based on automatically generated category trees. The former is mainly used for text classification, and the latter is mainly used for image classification.

Hierarchical classification based on predefined category trees is mainly applied to text classification. According to the learning strategy of the classifier, it can be divided into flat hierarchical classification (also known as the big-bang method) and top-down hierarchical classification. The most commonly used flat hierarchical classification method is Margin classifiers, which takes Support Vector Machine (SVM) [20] as representative. Dekel et al. [25] combined Bayesian analysis with Margin classifiers to make the classification hyperplanes of nodes with the same ancestor have a certain similarity. Cai et al. [15] combined the category tree with multi-class SVM, learned a classification hyperplane for each path from the root node to the leaf node, and proposed the concept of the hierarchical loss function (H-loss). Ramaswamy et al. [115] defined a loss function based on the distance of the category tree, that is, the loss is calculated based on the distance between the predicted value and the true value on the category tree.

Compared with flat hierarchical classification, top-down hierarchical classification is less applied in text classification, but more common in image classifi-

cation. Dumais et al. [29] compared the effects of hierarchical classification in the same flat through experiments, and trained several ”one vs all” SVM classifiers for each node. Fan et al. [33, 34] used hierarchical classification methods to solve the problems of image annotation and image semantic mining. Liu et al.[95] use the maximum likelihood estimation method based on statistical learning to synchronize the generation of the category tree and the learning of the classifier. Lu et al. [99] use CNN features to greatly improve the classification accuracy. Fan et al. [35] proposed a method of training the classifier layer by layer from the bottom up, and the feature subset is selected while learning the classifier.

## 2.3 Crowd-sourcing

### 2.3.1 Introduction

Crowdsourcing is a method for rapidly collecting and processing large-scale data by distributing tasks to a vast number of online participants. As a common annotation approach, crowdsourcing has been widely used in image dataset labeling, enabling the construction of large-scale datasets [67]. In recent years, crowdsourcing has demonstrated significant advantages in flexibility and large-scale processing capabilities for data annotation. For example, Rahman et al. [114] investigated how optimizing task design can reduce participant fatigue, thereby improving annotation quality and efficiency. Liu et al. [96] explored combining crowdsourcing with automated annotation, significantly reducing annotation time and cost by integrating human and machine-assisted labeling. Crowdsourcing techniques have attracted widespread attention from academia and industry, giving rise to many successful platforms such as Amazon Me-

chanical Turk[1], Prolific[2], Upwork[3], and Crowdflower[4].

### 2.3.2 Quality Control of the Crowdsourcing Annotation

Ensuring annotation quality is a critical challenge in crowdsourcing. This subsection discusses key strategies to address this challenge, including effective task design to improve clarity and reduce errors, task allocation strategies that optimize assignments based on participant skills and task complexity, multi-annotator strategies to enhance annotation reliability through redundancy, and incentive mechanisms that motivate annotators and ensure consistent, high-quality results. Each of these approaches contributes to improving the overall effectiveness and efficiency of crowdsourced annotation systems.

**Task Design.** Ensuring annotation quality is a critical challenge in crowdsourcing. Task design is crucial for maintaining high annotation quality, particularly for complex datasets [78]. Clear task instructions and well-structured allocation mechanisms can reduce errors and improve consistency [139]. Studies highlight that reasonable task design significantly enhances data annotation quality and participant engagement [94, 63].

**Task Allocation Strategies.** Task allocation strategies are equally important. Yu et al. [158] proposed a dynamic allocation method based on task difficulty, assigning tasks appropriate to participants' skill levels, thus improving annotation efficiency and quality. Jiang et al. [74] introduced a reinforcement learning-based task optimization method that dynamically adjusts task difficulty and types based on participant performance, maximizing annotation outcomes. Similarly, crowd diversity-aware allocation methods have been suggested to leverage the varied expertise of participants for more accurate results [7]. Moreover, Ho et al. [65] emphasized that integrating real-time feedback mechanisms

[1] http://www.mturk.com/
[2] https://www.prolific.com/
[3] http://www.upwork.com/
[4] http://crowdflower.com/

can further enhance task allocation by allowing participants to refine their submissions during the process. Zhao et al. [165] proposed combining task allocation strategies with predictive models to estimate annotator performance and dynamically assign tasks, yielding a significant boost in annotation consistency and reliability.

**Multi-Annotator Strategies.** Adopting a multi-annotator strategy is another effective method to ensure annotation consistency and accuracy [154]. Allowing multiple annotators to label the same data item significantly reduces noise and increases reliability [136]. Sheng et al. [127] proposed a method that focuses selectively on data points requiring higher-quality annotations, effectively managing noisy labels and improving data quality through repeated annotations. Welinder et al. [154] leveraged crowdsourced wisdom for multidimensional data annotation, enhancing annotation quality and consistency. These strategies are widely adopted in crowdsourcing platforms, especially for complex or highly subjective tasks, as they mitigate individual biases by aggregating judgments from multiple annotators.

**Incentive Mechanisms.** Incentive mechanisms are key to increasing annotator engagement and ensuring annotation quality [164]. Appropriate incentives attract high-quality annotators and encourage careful annotations through reward systems, reducing errors [72]. Studies show that performance-based reward mechanisms are more effective than fixed payments in motivating participants, ultimately improving overall annotation quality [101]. Recent research by Niu et al. [108] introduced a machine learning-based dynamic quality monitoring method, which automatically detects and corrects low-quality annotations during the annotation process.

## 2.4 Dataset Quality Evaluation Metrics

Evaluating the quality of image datasets is essential for ensuring their reliability and utility in machine-learning applications. Key metrics for dataset quality assessment include annotation accuracy, which measures the correctness of labeled data, and consistency, which evaluates the agreement among multiple annotators [136]. Metrics such as Krippendorff's alpha [84] are widely used to assess inter-annotator reliability, particularly for identifying low-quality annotations and annotators, ensuring the overall dataset quality [2]. For multi-annotator tasks, Fleiss' kappa [36] provides a robust measure of agreement among annotators, quantifying their collaborative consistency. In simpler scenarios involving two annotators, Cohen's kappa [19] is a valuable metric to evaluate agreement, especially for small-scale annotation tasks.

As datasets grow in scale and complexity, efficient quality evaluation methods are increasingly critical. Recent advancements include automated evaluation techniques. For instance, Lease et al. [91] proposed a machine learning-based method to predict annotation accuracy using classification models, significantly reducing manual verification efforts. Similarly, Burmania et al. [13] introduced a dynamic evaluation framework with real-time feedback mechanisms, enabling instant correction of low-quality annotations during the annotation process. Kuan et al. [87] introduced a framework for scoring label quality in large datasets, helping prioritize samples for review based on scoring metrics that incorporate agreement and label confidence. The IQAGPT system further illustrates the potential of leveraging large vision-language models and language models like ChatGPT for image quality assessment, combining cross-modal attention and semantic descriptions to outperform traditional methods in medical imaging tasks [17].

The challenge of detecting errors in labels across various tasks has also garnered significant attention. Token Classification Label Errors [152] developed

methods to identify label inconsistencies in token classification tasks, emphasizing their impact on downstream performance. Similarly, ObjectLab [143] proposed techniques to detect mislabeled images in object detection datasets, while Softmin [89] was proposed to focus on uncovering errors specific to segmentation annotations. These advancements highlight the growing need for domain-specific error detection methods tailored to different types of datasets.

Another critical aspect of dataset quality is addressing annotation uncertainty, which measures the variability in annotations for the same data item. Sheng et al. [127] demonstrated that analyzing discrepancies in multiple annotations for the same data can effectively identify challenging items for special handling, such as expert review. Chen et al. [16] further advanced this concept by introducing a Bayesian inference-based approach to quantify annotators' confidence levels for different categories, offering a precise measurement of annotation uncertainty. Additionally, ActiveLab [53] combined active learning with data re-labeling to iteratively refine annotations, addressing uncertainties in datasets while reducing labeling costs.

Tools like MetricDoc have also emerged to facilitate interactive quality metric customization and visualization, enabling analysts to profile, diagnose, and address quality issues more effectively during data preprocessing [9]. These systems allow users to apply domain-specific metrics with immediate visual feedback, ensuring high-quality data for downstream tasks. Similarly, the CROWDLAB framework provides a supervised model approach for inferring consensus labels, annotator reliability, and label confidence, outperforming traditional majority-vote methods in multi-annotator settings [54].

Detecting broader dataset issues such as drift and sampling bias has also gained traction. Cummings et al. [21] proposed methods to identify shifts in data distributions and non-independent identically distributed (non-IID) patterns, both of which significantly affect model performance. Zhou et al. [167] extended this effort by identifying mislabeled or erroneous entries in datasets

containing structured numerical data, further enhancing dataset reliability across domains.

Overall, these evaluation metrics and methodologies provide a robust foundation for ensuring high-quality image datasets, accommodating the growing diversity and complexity of modern annotation tasks. By integrating traditional metrics with advanced machine learning techniques, researchers can achieve more efficient and accurate assessments tailored to specific dataset requirements.

# Chapter 3

# Image Dataset Analysis and Problem Statement

## 3.1 Introduction

Datasets are the foundation of machine learning models, serving as the essential input for training and validation [162]. Their accuracy and consistency are pivotal in determining the generalizability and performance of these models. In computer vision, image datasets play a critical role in tasks such as object recognition and classification. The quality of these datasets directly impacts model performance and their applicability in real-world scenarios, making high-quality datasets a cornerstone of successful machine learning applications [155].

Several studies emphasize that high-quality image datasets must meet a series of stringent standards, including accurate labeling, consistent semantic annotations, and comprehensive representation of visual features [48]. These standards are essential for minimizing biases, ensuring reproducibility, and enhancing the reliability of trained models. Despite these criteria, current datasets often fall short, revealing significant gaps in quality [144, 111, 24, 109]. Common issues include semantic inconsistencies, labeling errors, and incomplete or imbalanced visual feature representation, which collectively undermine the

dataset's utility for robust machine learning.

This chapter aims to analyze the current state of image datasets, identify key challenges, and explore methods to address these issues. A particular focus is given to the Semantic Gap Problem (SGP) [135] and the origins of annotation errors, two critical factors that significantly degrade dataset quality. The SGP refers to the disconnect between visual data and their semantic interpretations, a challenge that exacerbates annotation inconsistencies and impairs a model's ability to generalize effectively. By understanding this problem, we can better design datasets that bridge this gap.

In this chapter, we will first delve into the concept and definition of the Semantic Gap Problem, exploring its prevalence in existing datasets and providing illustrative examples of how it leads to inconsistent annotations and affects model training. For instance, datasets like ImageNet have been criticized for ambiguous labeling and lack of semantic clarity, which result in a mismatch between visual features and their associated labels. This discrepancy not only hampers model performance but also limits the dataset's cross-cultural and cross-context applicability. Following this, we will systematically analyze the primary types of annotation errors, such as those related to high-quality images, mislabeled images, multi-object images, and single-object images. Each type will be examined in detail, discussing their root causes and their implications for dataset quality. For instance, multi-object images often introduce ambiguity in object identification and classification, while mislabeled images directly misguide model learning. We will highlight how these errors affect dataset usability and propose strategies to mitigate their impact.

Through this exploration, we aim to provide a comprehensive understanding of the challenges in constructing high-quality image datasets and offer insights into overcoming these issues. The ultimate goal is to create datasets that not only adhere to the highest standards of accuracy and consistency but also align semantic and visual properties, thereby laying a robust foundation for machine

learning models to achieve superior performance and broader applicability.

## 3.2 Semantic gap problems

The Semantic Gap Problem (SGP) was first introduced by Smeulders et al. [135] and defined as the *lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation*. Our intuition is that the SGP represents a core challenge in image datasets, manifesting as a misalignment between the labels and the visual data. The primary cause of this issue lies in the disconnect between visual representations and semantic interpretations. This problem not only arises in ambiguous label selection but also in the mismatch between visual feature extraction and label semantics.

The Semantic Gap Problem refers to the phenomenon where the visual features of the same image may correspond to multiple semantic labels. These discrepancies often stem from annotators' diverse backgrounds, cognitive differences, and the limitations of annotation tools. In other words, variations in context and cultural background can result in the same visual data being assigned entirely different semantic interpretations. As we show in Figure 1.1, objects such as an iceberg and an ice cube might both be labeled as "Ice". The SGP is not only observed across different individuals but also within the same individual under varying contexts. This many-to-many mapping highlights the complexity of the SGP, which manifests not only in label ambiguity but also in the subjectivity of semantic understanding among annotators.

The semantic gap issues are widely observed across various image datasets. We provide several examples in Figure 3.1. A "Brass Band" image with multiple objects is labeled as "wind instruments" in ImageNet, highlighting the discrepancy between annotators focusing on the overall scene versus specific objects, as shown in Figure 3.1(a). A similar case is illustrated in Figure 3.1(b),

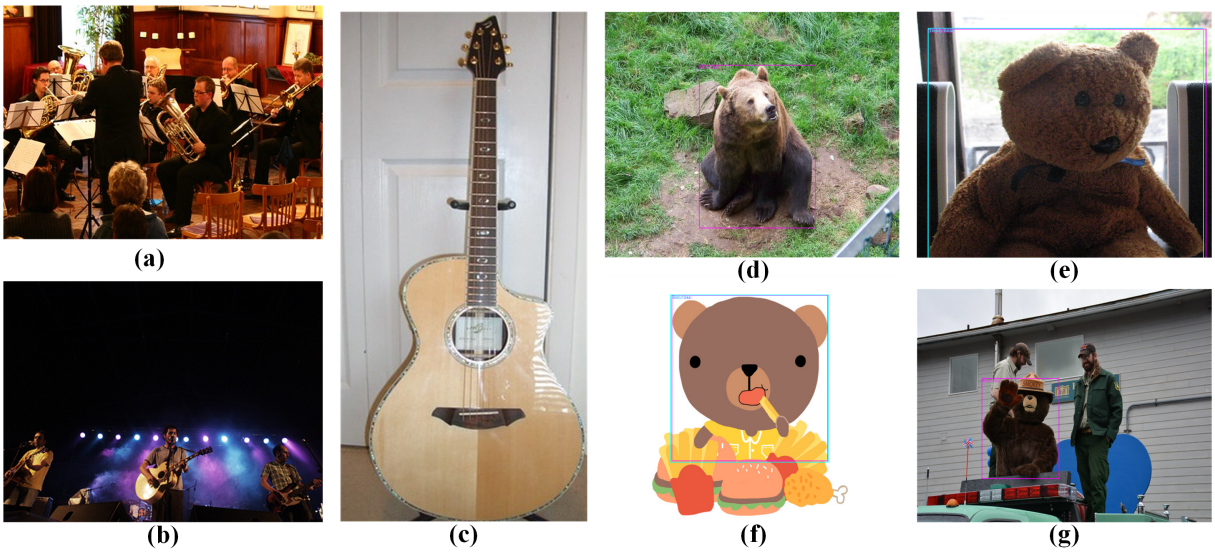


Figure 3.1: Images and their labels in existing image datasets. (a)-(b) These two images are examples of controversial labels in ImageNet. The seemingly valid labels "Brass Band" and "Stage" do not match the ImageNet labels, which are "Wind Instrument" and "Acoustic Guitar," respectively. [145] (c) This image appears in three categories of ImageNet, i.e., "Musical Instrument", "Guitar", and "Acoustic Guitar". (b)-(e) These four images are all labelled as the "Brown Bear" category in the Open Image Dataset.

where a multi-object image (e.g., a stage scene) is labeled based on a secondary object (e.g., "Electric Guitar") rather than the primary object (e.g., "Stage"). Such complex scenes and overlapping objects exacerbate the semantic gap problem (SGP). We found that Figure 3.1(c) in ImageNet falls under three categories, being labelled as "Acoustic Guitar", "Guitar", and "Musical Instrument". In our opinion, this situation occurs due to the level of categorization granularity provided subjectively by the annotators. This inconsistency appears to stem from the subjective granularity level chosen by annotators. Some image labels are overly generic, some image labels are overly generic, e.g., "tool", which makes it challenging for annotators to differentiate specific tool categories, thereby reducing the dataset's precision. While we cannot declare that these labels are incorrect, it is undeniable that the fine-grained inconsistency in image categories may negatively impact object recognition results. In addition, in the Open Images Dataset [81], we discovered that Figure 3.1 (d), (e), (f) and

(g) are all labelled under the category "Brown Bear", despite significant differences: Figure 3.1 (d) is an *animal*, Figure 3.1 (e) is a *toy* and also labelled as "Teddy Bear", Figure 3.1 (f) is a stick figure of a *cartoon* bear, and Figure 3.1 (g) looks like a person playing the role of a brown bear doll. The reason these four images are categorized under the same label is due to the lack of constraints in upper-layer domain categories [47], allowing annotators to subjectively classify at their discretion. These examples clearly illustrate the prevalence of the semantic gap issue across various datasets, which poses significant challenges for model training and performance evaluation.

## 3.3 Sources of annotation mistakes

Annotation mistakes are a common phenomenon in image datasets, directly impacting dataset quality and the performance of machine learning models. These mistakes primarily arise from subjectivity in the annotation process and the limitations of the tools and methods employed, representing a concrete manifestation of the SGP in image datasets.

We follow the framework outlined in [145], which characterizes three recurring *design flaws* associated with specific types of images in ImageNet. Using ImageNet as an example, we categorize its images into three types of annotation mistakes and an additional general category, where the fourth category includes images considered to have been correctly annotated. We have the following:

1. **Good Images**, for which there is wide inter-annotator agreement.
2. **Multi-Object Images**, where the flaw arises from the occurrence of multiple objects in the same image.
3. **Single-Object Ambiguous Images**, where the flaw arises from the fact that there is ambiguity in how to label the single object in the image.
4. **Mislabelled Images**, where the flaw arises due to labelling mistakes.

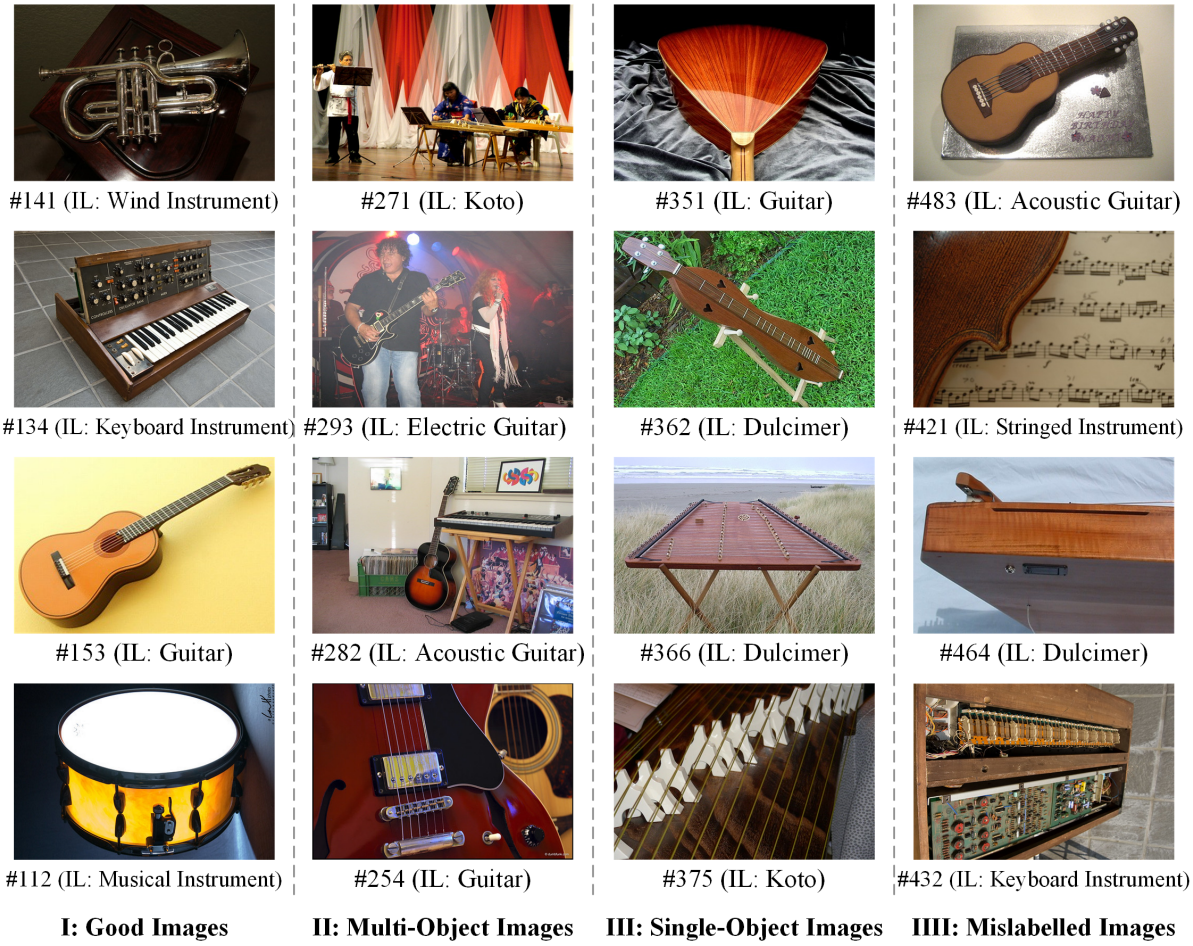


Figure 3.2: Samples from ImageNet Musical Instrument Sub-Hierarchy (IL stands for *ImageNet Label)*.

Evidence that the last three categories suffer from the SGP, comes from the fact that the ImageNet creators and the authors of [145] have contradictory opinions about them. However, also the Good Images category suffers from the SGP. We have detailed discussions on the characteristics of correctly annotated images. Then, we analyze the primary types of annotation mistakes in image datasets, providing detailed examples to illustrate their manifestations and their impact on dataset quality.

### 3.3.1 Good Images

High-quality images typically have a high level of annotator agreement. We can notice at least three characteristics definitive of this class of images. Firstly, these images are almost always those images containing a single object (the *"main object"*, as called in [145]). Secondly, these images are less noisy, in the sense that they have minimum influence of confounding variables such as occlusion and clutter distorting them. Thirdly, all of these images are captured from an optimal viewpoint leading to clear visibility of their defining visual characteristic; see, e.g., the image-label pairs for (I) Good Images in Figure 3.2. For example, an image that clearly displays a piano with a simple background and a well-chosen angle showing the keyboard in its entirety would generally be considered high quality.

Based on the good images, we have three observations. Firstly, though it is a fact that confounding variables are ubiquitous in open world settings [4], the training corpora for object recognition models should be majorly, but not only, comprised of good images. This is crucial given the fact that it is *only* the good images which can allow the model to extract and learn about its distinguishing visual characteristics. Secondly, and again ubiquitous in open world settings, is the issue of *parts* of an object in an image. Studies such as [145] have considered different labels for objects and its parts (e.g., "Car" vs "Car Wheel") treating them as distinct objects. The present work being focused on methodological visual classification postpones the specific issue of detection of parts of an object to a future version of the methodology. Thirdly, the SGP still persists for any of these images, even if agreed upon by multiple annotators. Specifically, different visual features of a single object may lead annotators to assign varying labels [50]. For instance, as shown in Figure 3.2, the image #153 is a close-up image of a wooden guitar might be labeled as either "Guitar" or "Acoustic Guitar," depending on the annotator's interpretation of the label semantics, and in

the ImageNet it is labelled as “Guitar”. Additionally, some high-quality images may exhibit ambiguity in labeling object components. For example, an image of a car might be labeled as either “Car” or “Car Wheel”, depending on whether the annotator focuses on the whole object or its part. Such ambiguities persist even among annotators with high agreement levels, underscoring the inherent challenges in achieving consistent labeling. To realize this, it is sufficient to think of the many other labels we could subjectively use to describe any of the good images.

### 3.3.2 Multi-Object Images

Multi-object images (MOI) refer to images containing multiple objects, where annotation mistakes often arise from the subjectivity in selecting the primary object. The flaw stems from the *systematic incongruence* between the ImageNet label and the label of the most likely main object, as deemed by humans. For instance, the example shown in Figure 3.1(b) is the image of a “Stage” labelled by humans as such, having the ImageNet label as “Acoustic Guitar”. We can also find more examples in (II) Multi-Object Images of Figure 3.2). However, an image of a stage scene may include multiple objects such as an electric guitar, a microphone, and speakers. While humans typically rely on cue validity to identify and annotate the most visually salient primary object, such as labeling the image as “Stage,” the actual annotation may depend on the annotator’s subjective choice. This can result in the image being labeled as “Acoustic Guitar”, “Electric Guitar” or “Microphone”, leading to semantic inconsistencies. According to our statistics, these images constitute more than one-fifth of ImageNet’s total images.

Multi-object images assume central importance due to two pivotal observations. Firstly, empirical evidence in cognitive psychology [122], makes clear that the main object chosen by humans possesses the highest *‘cue validity’*, in other words, carries the most information via perceptual attributes (*stimuli*) and

thus is visually the most salient. Secondly, in violation of the first observation, ImageNet exhibits an established bias for many multi-object images wherein, for such an image, its label corresponds to a very *distinctive object* instead of the main object in that image, thus exploiting features that don't generalize to object recognition in the wild [145]. Such errors not only hinder the model's ability to learn the primary object's features but can also cause the model to overfit to irrelevant features.

### 3.3.3 Single-Object Ambiguous Images

Single-Object Images (SOI), as shown in Figure3.2-(II)), are by far the most important source of annotation mistakes. The main reason is in the underlying interaction between *visual polysemy* [135] and *linguistic polysemy* [103]. Single-Object images present three types of ambiguity.

**Single-Object Image – Ambiguity Caused by by Itself (SOII).** In the first case, we named SOII, *the source of subjectivity is intrinsic to the Image itself.* It occurs when an object in an image is *visually polysemic*[1], namely when its visual *"semantics is described only partially"* [135] and, consequently, its linguistic description is not unique. An example is Image #351 representing a dreadnought-shaped instrument, which is labelled *Guitar* in Figure3.2-(III), but could also be labelled, e.g., *Bass* or *Ukelele*. Notice that images of this type usually encode *a typical partial view* with no possibility of discrimination among different objects. [145] observes that not only do humans assign an alternative label for 40% of the images, but they also assign up to 10 different labels.

**Single-Object Images – Ambiguity Caused by Lables (SOIL).** The second source of subjectivity in single-object images *is intrinsic to the meaning of Labels*, which we named SOIL. There are two cases. In the case of *polysemic* labels (i.e., of one-to-many mapping between labels and images), different objects, each corresponding to one of the possible meanings, are associated with

[1] See the definition in https://gradientscience.org/benchmarks/

the same label[49]. One such example is the two images labelled as *Dulcimer* (Image #362, #366) in Figure 3.2-(II)) where the two objects are essentially two different musical instruments, i.e., the first *Dulcimer* being native to the American and the second to the Chinese culture, and are visually very different. The study in [145] highlights how these ambiguous classes generate overlaps in the image distributions with a negative impact on performance. The second case labels which are *synonyms* (i.e., of many-to-one mapping between labels and images) wherein a single set of images, will wrongly be split into two sets. One such example is that of *Koto* (Image #375) which can also be synonymously labelled as, for instance, a *Kin* or a *Jusangen*[42].

**Single-Object Images – Ambiguity Caused by Annotators (SOIA).** The third source of subjectivity in single-object images *relates to the annotator's incomplete background knowledge*, we called SOIA. Thus, for instance, Image #357 is labelled as *Guitar* in Figure 3.2-(II), while the image depicts an *Acoustic Guitar*. This type of generic label appears whenever correct labelling requires deep domain specific knowledge. [24] discusses the bias which arises because of this problem in the case of minority *languages*, *cultures*, and also with images about *low income* people. Notice that this is not mislabelling. The issue is that the label captures some but not all of the visual properties of the object.

Overall, the sources of annotation mistakes can be summarized by saying that MOI images generate a one-to-many mapping from images to labels, those of type SOIA generate a many-to-one mapping, while those of type MI, SOII, and SOIL generate both types of mappings. These four types of mistakes are independent of one another and can, therefore, occur together in the same image, thus generating multiple occurrences of the SGP many-to-many mapping problem.

### 3.3.4 Mislabelled Images

Mislabeled images (MI) refer to instances where the assigned label does not match the actual content of the image. Such errors typically result from carelessness during annotation, a lack of knowledge, or excessive annotation speed, that is, all well-known crowdsourcing problems, see for instance [22]. A concrete example is in Figure 3.2-(III), in the ImageNet dataset, an image labeled as "Acoustic guitar" might actually depict a birthday cake shaped like an acoustic guitar. These labeling errors directly impact the accuracy of model training and can lead to the model learning incorrect features.

Another common example involves mislabeling in the "Tools" category. An image of a screw might be mistakenly labeled as a "Nail" due to shared background elements or similar usage contexts. This type of error is particularly prevalent in crowdsourced annotations, especially when annotators lack sufficient domain expertise.

Furthermore, such mistakes may also reflect inherent limitations in dataset design. For instance, the definitions of certain categories may be overly ambiguous or unintuitive, causing annotators to frequently make incorrect label selections. The cumulative effect of these issues can significantly undermine the generalization ability of models in downstream tasks.

## 3.4 Statistical Analysis of Annotation Mistakes

We conducted a detailed statistical analysis of annotation mistakes in image datasets, with ImageNet serving as a representative example. The objective of this experiment is to examine the common annotation issues in current image datasets, evaluate the quality of these datasets, and provide data-driven insights to support the improvement of image dataset construction methodologies in the future.

Table 3.1: Statistics about annotation problems in the ImageNet musical instruments subset.

| Number | Musical Instrument | Stringed Instrument | Keyboard Instrument | Wind Instrument | Guitar | Dulcimer | Koto | Acoustic Guitar | Electric Guitar |
|---|---|---|---|---|---|---|---|---|---|
| **Good Images** | 175 | 134 | 242 | 19 | 204 | 69 | 18 | 263 | 123 |
| **Multi-Object Images** | 326 | 345 | 209 | 244 | 236 | 209 | 166 | 188 | 202 |
| **Single-Object Ambiguous Images** | 3 | 12 | 17 | 19 | 46 | 24 | 1 | 33 | 30 |
| **Mislabled Images** | 2 | 6 | 11 | 8 | 18 | 14 | 3 | 21 | 20 |
| **All Images** | 506 | 497 | 479 | 290 | 504 | 316 | 188 | 505 | 375 |

### 3.4.1 Experimental Design

In this experiment, we selected a subset of the ImageNet dataset related to musical instruments, comprising a total of 3,660 images across 9 classes: Musical Instrument, Stringed Instrument, Keyboard Instrument, Wind Instrument, Guitar, Dulcimer, Koto, Acoustic Guitar, and Electric Guitar. Following the classification described in Section 3.3, the images were divided into four categories: Good Images, Multi-Object Images, Single-Object Ambiguous Images, and Mislabelled Images. The classification results are reported in Table 3.1, with each column corresponding to the respective categories under consideration.

### 3.4.2 Results and Analysis

The results reveal that the number of Good Images meeting our image annotation classification criteria, although highly variable and dependent on the category, consistently accounts for less than half of the total images. This indicates that even relatively simple categories exhibit a limited proportion of high-quality annotations. In contrast, Multi-Object Images is the largest category across all classes, highlighting the widespread challenges of annotating scenes with multiple objects. For instance, in the Wind Instrument class, the number of Multi-Object Images reached 244, comprising 84.1% of the total images in that category. This underscores the difficulties of accurately annotating objects in complex scenes.

Although Single-Object Ambiguous Images are fewer in number, they exhibit significant ambiguity in certain classes. For example, in the Guitar class, annotation ambiguities arise due to insufficient visual granularity stemming from annotators' incomplete contextual knowledge, a phenomenon typical of SOIA. In contrast, ambiguities in the Dulcimer class are often due to the SOIL, where annotators from different regions perceive the term "Dulcimer" to represent distinct instruments.

Further analysis shows that Mislabelled Images account for 2.8% of the total images, closely aligning with the 2.7% reported in prior work [145], namely 270 mislabelled images out of approximately 10,000. This similarity demonstrates that our evaluation aligns with well-established benchmarks. Furthermore, our analysis provides a more detailed understanding of annotation mistakes. If all annotation mistakes were considered, this percentage would significantly increase due to our stricter definitions of annotation mistakes. For instance, a guitar-shaped wooden object lacking key features such as taut strings might be labeled as "Guitar" in [145], but according to our standards, such images fall under annotation mistakes, namely the SOII. This highlights the sensitivity and effectiveness of our methodology in identifying annotation mistakes.

## 3.5 Summary

The Semantic Gap Problem presents multiple challenges for constructing visual datasets. First, it increases inconsistency among annotators, leading to confusion in label distribution. Second, this issue is particularly pronounced in cross-cultural and multilingual tasks, as different cultural backgrounds may assign varying semantics to the same object. Finally, the SGP directly impacts a model's ability to learn the mapping between visual features and semantic labels, thereby reducing its generalization capability in real-world scenarios.

The experimental results reveal common issues in current image datasets, in-

cluding incomplete annotations for multi-object scenes, inconsistent annotation standards, and a significant proportion of annotation errors. These challenges not only compromise the quality of the dataset but also negatively impact model training and performance evaluation. To address these issues, it is necessary to enhance methodologies for constructing image datasets, refine annotation tools and processes, and implement efficient quality control mechanisms to reduce annotation mistakes. Through an in-depth analysis of annotation mistakes and the SGP in current image datasets, this Chapter provides specific observations and theoretical support for addressing annotation ambiguities and improving semantic consistency. It also lays the groundwork for the methodological designs discussed in subsequent chapters.

# Chapter 4

# vTelos: Image Dataset Construction Methodology

## 4.1 Introduction

In this Chapter, we propose `vTelos`, an integrated Natural Language Processing (NLP), Knowledge Representation (KR) and CV methodology whose main goal is to generate high-quality datasets. There are two main underlying intuitions. The first is that the *purpose* of an annotation effort, i.e., what the ML model trained by the dataset should be used for, should be made explicit via precise guidelines, that we organize around *four main Design Choices*, as follows:

- C1: for each *image*, which object(s) should be considered?
- C2: for each *object*, which set of properties should be considered?
- C3: for each *set of properties*, which label should be selected to properly describe them?
- C4: for each *label*, if *polysemic*,[1] how do we unambiguously represent its meaning?

[1] For instance, in WordNet, the average number of meanings per noun and per polysemous noun are 1.25 and 4, respectively, where the more common a noun, the higher the polysemy [39].

The second underlying intuition is that the role of the annotator should be split in two, i.e., that of the *Classificationist* and that of the *Classifier*, each associated with different tasks (this terminology being adopted from KR and Knowledge Organization (KO) [116]). Thus, the classificationist, will first define the space of labels to be used (Design Choices C3, C4) and, for each of them, the (visual) properties they describe. Then, in a second phase, the classifier will classify images using the labels defined by the classificationist (Design Choices C1, C2), where Choice C2 is based on the output of Choice C3. Notice how the proposed decision process is reversed with respect to how image annotation is usually performed, as also described above. The key idea is that we should first select the space of labels to be used, with a clear understanding of their intended meaning, in terms of the visual properties they are taken to describe, and then exploit this clarity when selecting, for each input image, the label which most properly describes its content. The proposed approach actually follows the standard practice in KR, for instance when classifying products in an online catalogue [51].

In this process, *the intended natural language semantics of natural language terms* (as selected during C3, C4) *and the relevant visual information of images* (i.e., objects and their visual properties, as selected during C1) *are natively aligned* (during C2), this being the reason why the SGP does not appear. The annotation bias is dealt with not by building an all-encompassing dataset, which would be impossible [10, 45] but, rather, by making explicit the annotation choices, as in [31]. This advantage is further amplified by the fact that, instead of performing choices C3 and C4 from scratch, we take, as a predefined source of natural language labels, the *WordNet Genus-Differentia lexico-semantic hierarchy* [103]. In WordNet, each label is associated with a *gloss*, which, in fact, is an *unambiguous linguistic definition of the properties of the objects it describes*. This opens up the possibility of full and general alignment between natural language lexical semantics and object recognition, thus enabling a general-purpose

solution to the SGP many-to-many mapping problem.

## 4.2 Four Annotation Design Choices

As from the introduction, the `vTelos` annotation process is articulated into four main choices C1-C4. Let us analyze them in detail.

1. *Choice C1: Object localization.* Choose one of the (one or more) objects in an image, thus eliminating any possible source of *object ambiguity*.

2. *Choice C2: Visual classification.* Choose the relevant *visual properties* of the object under consideration and annotate the image with one of the labels selected in Choice C3. This choice is conceptually similar to the usual annotation choice. The key difference is that what is selected is not the label but, rather, the linguistic properties that describe the visual properties of the object. Thus for instance, if the object occurring in the image is a guitar, the annotator would not choose the label *Guitar* but, rather the property *having six strings*. The label *Guitar* would then be automatically associated to the label because of the definition of guitar generated by Choice C3. This eliminates any source of *visual ambiguity*.

3. *Choice C3: Label generation.* Choose the space of labels used to annotate images. Here the meaning of each label is *defined* in terms of linguistically defined properties encoding a selected set of visual properties. One such example is the definition of a *guitar* as being a *"string instrument with six strings"*. This eliminates any source of *linguistic ambiguity* intrinsic to *informally defined labels* and it allows to say that two labels are *synonyms*.

4. *Choice C4: Label disambiguation.* Choose a unique concept identifier for each of the (one or more) meanings of each label. This eliminates any source of *linguistic ambiguity* intrinsic to *polysemous labels* (see footnote [1] in this Chapter).

Table 4.1: The vTelos four choices. MOI and MI stand for Multi-Object and Mislabelled Images, respectively. SOII, SOIL, SOIA stand for Single-Object Images where the source of subjectivity is the Image, the Label and the Annotator background knowledge, respectively.

| ***Role*** | ***Annotation Choice*** | ***Purpose*** | ***Source of Subjectivity*** | ***Solution*** |
|---|---|---|---|---|
| Classificationist | C3: Label generation | Choose label and its properties | MI, SOII | Linguistic Genus-Differentia |
| Classificationist | C4: Label disambiguation | Choose identifier | SOIL | Alinguistic Identifiers |
| Annotator | C1: Object localization | Choose object | MOI | Bounding Polygons |
| Annotator | C2: Visual classification | Choose label via its properties | MI, SOII, SOIL, SOIA | Visual Genus-Differentia |

These four choices allow us to deal with the annotation mistakes described in Chapter 3. Consider Table 4.1, where the first column reports the two vTelos roles (with *Annotator* can also be read as *Classifier*), the second, the design choice, the third, the purpose behind the choice, the fourth, the source of subjectivity which is addressed and the last, the solution that is enforced.

In Table 4.1, the first and the last column are crucial to the vTelos approach for three reasons. The first is the splitting of the annotation tasks into two roles, i.e., *classificationist* and *annotator*, where the first is in charge of preparing a well-specified annotation task for the second. The second reason is that, opposite to the annotation standard practice, in vTelos, Choices C3,C4 are performed before Choices C1,C2. This is in order to force the annotator to select a label within of a (possibly very large) pre-selected *controlled vocabulary*. Following vTelos this vocabulary should be built by a domain expert. The third reason is that classificationist and classifier coordinate through the use of the *Visual* and the *Linguistic Genus-Differentia* (see the last column in Table 4.1). The idea behind the notion of Linguistic Genus-Differentia is that the meaning of a label (e.g., *Guitar*) is defined in terms of a linguistic description composed of two parts, the *(linguistic) Genus*, which describes what the object has in common with other objects (in the case of a guitar the fact that it is a *string instrument*), and the *(linguistic) Differentia* which describes which properties differentiate the object from the other objects with the same genus (in the case of a guitar, the fact that it has *six strings*). In turn, given a label whose

meaning is defined in terms of Linguistic Genus-Differentia, any object denoted by that label, e.g., a guitar, must have a *(visual) Genus* which is described by the linguistic genus of the label and similarly for the *(visual) differentia*. We often talk of linguistic and visual differentia in terms of *linguistic* and *visual properties*.[2]

For what concerns the classifier, the goal of Choice C1 is to select the image *main object* (see Section 3). Consider for instance, in Figure 3.2 (II), the objects in the MOI Image #271, i.e., a *Koto* but also a *Flute*, *Music Stand* etc., all labelled as *Koto*. Choice C1 solves the problem of the MOI source of subjectivity. The possible strategies are: (i) removing the image just because of multi-object, (ii) splitting the image into sub-images, one per object, or (iii) isolating objects, e.g., via bounding polygons, in this case allowing for multi-label images. Notice how Choices C1, C2 are crucially based on the distinction between *object localization* and *visual (image) classification*, where the former is an *inherently visual activity* where an object is localized but not recognized, while the latter is an *inherently semantics-driven activity* which determines the relevant properties of an object. This distinction, while being well known, is often not enforced in many state-of-the-art approaches, which collapse object detection and recognition (see [58, 123]).

Choice 2 enables the classifier to select the label whose genus and differentia align with the visual genus and differentia of the object in the image. There are only two possibilities. Either such a label does not exist, in which case the object does not belong to that class, or it does exist, in which case the opposite holds. The case of multiple possible labels cannot occur as two different objects with the same genus cannot have the same differentia. This is the key guideline enforced by the classificationist to avoid the occurrence of the SGP many-to-many problem, modulo mistakes on the side of the annotator, i.e., the selection

[2]The distinction between the visual properties of an object and the linguistic properties which are used, in natural language, to define the label used to name the object, is based on the *Teleosemantics* theory of meaning [104, 105], as adapted to CV in [50].

of a label that does not describe the contents of the image. Thus, for instance, a positive example is Image #134 in Figure 3.2-(I), where the ImageNet gloss of the label *Keyboard Instrument* forces the selection of the visual property of being a *"a musical instrument that is played by means of a keyboard"*. Dually, Choice 2 forces the rejection of all images which do not satisfy the meaning of the label. Thus, for instance, `vTelos` will reject the MI Image #483, as the ImageNet label *Acoustic Guitar* is defined as *"a guitar with no input jack"*; but the picture does not represent a guitar and, more in general, not even a musical instrument.

## 4.3 The Role of Annotators

As introduced in Section 4.1, the proposed `vTelos` methodology divides the annotation process into two distinct roles: the Classificationist and the Classifier. This division not only clarifies task boundaries but also significantly enhances annotation accuracy and consistency through collaborative processes at different stages.

### 4.3.1 The Classificationist

The core responsibility of the Classificationist is to establish the foundational structure for the annotation process. They define and organize the labeling system, including the semantic definitions of each label (addressing C3 and C4) and their associations with visual properties. By performing this role, the annotation tasks are semantically structured, mitigating inconsistencies caused by ambiguous label definitions.

A typical example involves using lexical resources such as WordNet to define labels. For instance, the Classificationist may define "Guitar" as a "string instrument with six strings" and further specify it as "Acoustic Guitar". This eliminates semantic ambiguities in the labels, providing clear guidance for the

Classifier's subsequent tasks. The Classificationist also handles label disambiguation (C4), ensuring consistency by assigning unique identifiers to each possible meaning of a label. For example, the term "Dulcimer" may represent different instruments in various cultures (e.g., American vs. Chinese dulcimers). The Classificationist can resolve this ambiguity by assigning distinct identifiers to each cultural variant.

Additionally, the Classificationist explicitly links visual properties to labels. For instance, for the category "Piano", the Classificationist may define it as "a musical instrument with keys" and specify visual characteristics such as "88-key layout" or "upright design". This clarity reduces subjectivity during annotation and ensures consistency in label assignment.

### 4.3.2 The Classifier

The Classifier is responsible for executing the annotation tasks based on the label system and guidelines defined by the Classificationist (addressing C1 and C2). Their primary responsibilities include identifying target objects in images and selecting the labels that align most accurately with the semantic definitions.

In the object localization stage (C1), the Classifier identifies the primary object within an image. For instance, in a stage scene image, the Classifier analyzes the visual elements and selects "Stage" as the main object, rather than "Microphone" or "Speaker". This process leverages the definitions and prioritization rules provided by the Classificationist, reducing errors caused by subjective preferences.

In the visual classification stage (C2), the Classifier further analyzes the target object's features. For example, when annotating an acoustic guitar image, the Classifier does not directly choose the label "Guitar" but first identifies its visual attributes, such as "wooden body" and "six strings". These attributes are then matched with the semantic properties defined by the Classificationist (e.g., WordNet glosses), automatically determining the label "acoustic guitar". This

process effectively eliminates ambiguities arising from unclear features.

The Classifier plays a crucial role in annotating multi-object images. For example, in the ImageNet dataset, an image may contain multiple objects such as an "electric guitar", "microphone", and "speaker". Following the Classificationist's defined rules, the Classifier can delineate these objects using bounding polygons and annotate them separately. Alternatively, priority rules can be applied to select the most important object as the focus.

### 4.3.3 Benefits of the Annotator Roles

The collaboration between the Classificationist and the Classifier establishes a systematic and standardized annotation process. The Classificationist focuses on defining labels and their semantic hierarchies through linguistic genus-differentia, providing clear instructions to the Classifier by aligning these definitions with visual genus-differentia. For example, the Classificationist may define the linguistic genus-differentia for "String Instruments" and refine visual properties like "six strings" and "bridge design", guiding the Classifier to annotate an image as an "Acoustic Guitar". This approach ensures logical clarity and consistency, avoiding errors caused by ambiguous label definitions or semantic discrepancies.

The Classifier, in turn, leverages the label system provided by the Classificationist to focus on visual analysis and label assignment. This division of labor significantly reduces annotation mistakes by standardizing the annotation process through a controlled labeling system, ensuring alignment between semantics and visual representations. For instance, after identifying an object's attributes (e.g., "wooden appearance"), the Classifier references the Classificationist's semantic definitions to generate precise label mappings. This collaborative framework not only enhances annotation quality and consistency but also provides effective solutions to semantic gap problems and annotation mistakes, laying a robust theoretical and practical foundation for the `vTelos` methodol-

ogy's construction of high-quality datasets.

## 4.4 Summary

In this Chapter, we introduce the proposed `vTelos`, a novel image dataset construction methodology, which integrates theories from NLP, KR, and CV to address the pervasive annotation challenges in current image datasets. By introducing four carefully designed annotation choices and clearly defined roles for annotators, this approach significantly enhances dataset quality, consistency, and semantic clarity.

This methodology leverages the Classificationist role to construct a robust and systematic label framework, ensuring that labels are unambiguous and aligned with visual and linguistic properties. This precision reduces inconsistencies in annotations caused by ambiguous label definitions. Simultaneously, the Classifier role focuses on applying these predefined labels to images through a rigorous process of object localization and feature-based classification. This division of labor minimizes annotation mistakes and ensures a seamless integration of semantic and visual information, effectively mitigating issues arising from the semantic gap problems.

Furthermore, `vTelos` introduces a novel annotation pipeline that incorporates linguistic genus-differentia and visual genus-differentia principles. This framework ensures that annotations not only align with linguistic definitions but also accurately reflect the visual characteristics of objects. Such alignment is particularly critical for addressing the challenges of multi-object and single-object images, as well as resolving polysemy and synonymy in annotation labels.

By implementing these innovations, `vTelos` offers a scalable and adaptable methodology for constructing high-quality datasets. It provides a comprehensive solution to the annotation errors and semantic ambiguities that have long

hindered the development of reliable and generalizable machine learning models. This systematic approach paves the way for improved dataset reliability and supports a wide range of applications in computer vision and beyond.

# Chapter 5

# Classificationist: Category Definition

## 5.1 Introduction

In this chapter, we present a detailed exploration of the responsibilities of the proposed Classificationist, a role designed to guide the systematic definition of image datasets based on visual attributes. Specifically, we introduce key principles, such as canons, that are used to distinguish visual objects, providing methodological guidelines for dataset creation. Additionally, this chapter elaborates on the object categories defined by the Classificationist and the associated visual properties that guide the Classifier in annotating image data. This study is based on the work presented in [44, 47, 48].

Research on visual semantics is a cornerstone of computer vision, aiming to construct semantically consistent and high-quality datasets that improve the alignment between visual data and semantic representations [44]. In the training and evaluation of machine learning models, the quality of datasets directly impacts their generalization ability and performance. However, existing image datasets often suffer from the Semantic Gap Problem, a disconnect between the information extracted from visual data and its semantic interpretation. This gap not only leads to annotation errors but also limits the applicability and robustness of models in diverse real-world scenarios.

The root of the semantic gap lies in the disparity between the fields of Knowl-

edge Representation (KR) and Knowledge Organization (KO). KR, a pivotal area of artificial intelligence, focuses on "how to represent knowledge in symbolic form within intelligent systems" [11]. It encompasses a wide range of technologies and methodologies, such as the semantic web technology stack [23] and ontology construction [56], to facilitate the structuring of knowledge through models like conceptual hierarchies and knowledge graphs (KGs). Despite its technical maturity, KR approaches often face criticism for their neglect of modeling quality. For example, datasets like ImageNet are plagued by issues such as label ambiguity and annotation inconsistencies [26, 146], which highlight fundamental flaws in KR-based dataset design, resulting in data biases and reduced model performance [124, 79].

On the other hand, KO focuses on the systematic organization of information resources, often using methodologies like Facet Analysis, initially proposed by Ranganathan [120, 119]. KO methods provide clear principles for constructing high-quality models through classification hierarchies and canons. However, these methods have primarily been applied in information science and have not been fully exploited in technically driven dataset modeling. This gap in methodological integration leaves existing datasets without clear guidance for label definitions and semantic consistency. Consequently, systematic flaws in these datasets, such as annotation errors and ambiguous labels, are inherited during model training, thereby limiting the applicability of data-driven models in cross-cultural and multilingual tasks.

To address these challenges, we introduce the Classificationist role within the proposed `vTelos` framework, which integrates KR and KO methodologies to systematically resolve the semantic gap and enhance dataset quality and consistency. The Classificationist plays a pivotal role in defining classification systems, ensuring alignment between semantic and visual attributes. Drawing on KO's facet analysis methods, the Classificationist establishes explicit guidelines, including:

- Canons: Systematic rules for classification, such as selecting visual features as classification criteria.

- Glosses: Semantic definitions to eliminate label ambiguity. For instance, "Guitar" can be defined as "a string instrument with six strings".

- Visual Properties: Associations between visual features and semantic definitions to ensure consistency between visual and linguistic descriptions.

- Hierarchy: A structured framework built on visual properties, integrating visual genus and visual differentia to systematically organize and classify object categories. This hierarchy ensures semantic clarity and consistency in image datasets and serves as a guiding structure for annotation processes.

The role of Classificationist is vital to `vTelos`. By addressing ambiguities in label definitions and establishing hierarchical classifications and semantic associations, the Classificationist significantly enhances the modeling quality of datasets. Their efforts ensure systematic and consistent annotation processes, laying a strong foundation for the construction of high-quality datasets.

In the following sections, we delve into the responsibilities of the Classificationist within the `vTelos` framework. We discuss the guiding principles for distinguishing visual objects, including canons, glosses, and visual properties, and explore how these methods are employed to define object categories. Additionally, we examine how these principles provide clear guidance to the Classifier for annotating image data. Through these discussions, we demonstrate how `vTelos` integrates the strengths of KR and KO to construct systematically robust datasets and offer a comprehensive solution to the semantic gap problem.

## 5.2 Canons

When designing high-quality image datasets, ensuring the consistency between semantic and visual attributes is crucial. To address this, we adopted Ranganathan's Facet Analysis methodology [116], which defines visual objects hierarchically through a set of normative principles known as canons. These canons regulate dynamic and accurate knowledge classification, forming the core of hierarchical design. They systematically organize visual features and semantic representations, thereby alleviating the SGP. This section elaborates on the detailed design of canons and their practical applications, demonstrating their essential role in visual object classification.

To construct a structured taxonomy for visual objects, Ranganathan's faceted approach divides the classification process into four phases:

- **Pre-Idea Plane**: Focuses on generating perceptual concepts by extracting perceptual features of visual objects.
- **Idea Plane**: Centers on organizing concepts into hierarchical structures based on visual properties, serving as the core stage for visual classification.
- **Verbal Plane**: Maps hierarchical structures to lexical semantics, ensuring semantic consistency through naming conventions.
- **Notational Plane**: Formalizes the hierarchy in a language-independent manner using unique identifiers.

Among these, the Idea Plane plays the most critical role in defining visual classification hierarchies through the application of core canons. The following sections introduce the design principles for each plane in detail.

### 5.2.1 Pre-Idea Plane

The Pre-Idea Plane focuses on generating perceptual concepts, forming the foundational stage of the taxonomy. It emphasizes the accumulation of perceptual features through iterative observations, enabling the gradual formation of object concepts.

**The Canon of Perceptual Accumulation.** The Canon of Perceptual Accumulation underscores that concept generation must rely on accumulated observations rather than isolated experiences. For instance, the concept of "Guitar" develops progressively through observations across multiple images, capturing its shape, string arrangement, and structural details.

**The Canon of Perceptual Consistency.** The Canon of Perceptual Consistency ensures that perceptual features remain consistent across varying conditions and environments. For example, whether a guitar is photographed against a complex or minimalistic background, its key perceptual attributes, such as the strings and outline, must remain identifiable.

**The Canon of Dynamic Adaptation.** Lastly, the Canon of Dynamic Adaptation allows concepts to evolve as new perceptual details emerge. For instance, initially unfamiliar stringed instruments might be broadly categorized under "String Instruments" but could later be refined into specific categories such as "Dulcimer" or "Koto" as more features are observed.

These norms lay the groundwork for building a taxonomy that is gradual, consistent, and adaptable to evolving perceptual data.

### 5.2.2 Idea Plane

The Idea Plane serves as the central phase for constructing the classification hierarchy, focusing on the selection and systematic organization of visual features. Its primary goal is to define and structure visual objects based on their distinguishing characteristics, ensuring a consistent and logical classification

system.

**The Canon of Differentiation.** The Canon of Differentiation requires that every chosen feature distinctly separates entities into at least two categories, ensuring clear divisions in the classification process. For example, in musical instruments, the sound production mechanism serves as a reliable feature: string instruments produce sound through vibrations of strings, while wind instruments rely on the airflow through an embouchure.

**The Canon of Relevance.** The Canon of Relevance ensures that the chosen features align closely with the classification's primary purpose. For instance, in musical instrument classification, sound production methods are more relevant than superficial characteristics like color, which do not contribute to the functional differentiation of the instruments.

**The Canon of Ascertainability.** The Canon of Ascertainability emphasizes the importance of selecting features that are clearly defined and measurable. For instance, the distinction between acoustic and electric guitars can be made based on the presence of an input jack—a feature that is both observable and verifiable.

**The Canon of Permanence.** The Canon of Permanence dictates that the selected features should remain stable over time, as long as the classification's purpose remains the same. For example, the classification of instruments by their sound generation mechanisms—such as strings, winds, or percussions—is a timeless standard that is unlikely to change.

**The Canon of Concomitance.** Beyond feature selection, the Idea Plane also introduces norms for organizing features in a logical and sequential manner. The Canon of Concomitance prevents the selection of overlapping features that appear frequently across multiple objects, as this can lead to redundancy and confusion. For instance, classifying musical instruments by both their construction material and primary use could result in ambiguous overlaps, such as acoustic and electric guitars both being categorized under "wooden instruments", ignor-

ing their functional differences.

**The Canon of Relevant Succession.** The Canon of Relevant Succession ensures that the order of feature application follows a logical progression aligned with the classification goal. For example, in musical instruments, classification could begin with the sound production mechanism (e.g., strings, winds), then proceed to shape (e.g., saxophones, flutes), and finally to specific attributes such as the number of strings.

**The Canon of Consistent Succession.** The Canon of Consistent Succession reinforces uniformity across the classification system by requiring that the same sequence of features be consistently applied to similar categories. For instance, if guitars are classified by type (e.g., acoustic, electric) and then by string count, the same sequence should be applied to other string instruments, such as violins and zithers, to maintain structural coherence.

**The Canon of Exhaustiveness.** Furthermore, the Canon of Exhaustiveness ensures that all known objects within a classification are accounted for, leaving no gaps. For example, the classification hierarchy for string instruments must include all varieties, such as guitars, zithers, and kotos, to prevent omissions.

**The Canon of Exclusiveness.** The Canon of Exclusiveness mandates that each object belongs to only one unique category within the hierarchy, avoiding overlaps. For instance, while a piano has strings, its classification under keyboard instruments reflects its primary sound production mechanism, ensuring a clear and exclusive categorization.

**The Canon of Modulation.** Finally, the Canon of Modulation ensures the inclusion of intermediate categories in the classification chain. For example, when classifying electric guitars, it is essential to include "Guitar" as an intermediate category between "String Instruments" and "Electric Guitar" to maintain a logical progression.

Through these canons, the Idea Plane establishes a robust framework for organizing visual objects, addressing the SGP by creating a structured alignment

between visual features and semantic attributes. By applying these principles systematically, the classification system ensures clarity, consistency, and adaptability for complex datasets.

### 5.2.3 Verbal Plane

The Verbal Plane focuses on mapping the hierarchical classification to lexical semantics, ensuring that concepts are clearly defined and consistently named.

**The Canon of Clarity.** The Canon of Naming Clarity requires that each concept's name be accompanied by an explicit semantic definition, eliminating ambiguity. For example, the term "Guitar" should be defined as "a string instrument with six strings" to ensure precise understanding.

**The Canon of Linguistic Consistency.** The Canon of Linguistic Consistency ensures uniformity in naming conventions across the taxonomy. For instance, all string instruments should follow a consistent naming structure, such as "Acoustic Guitar" and "Electric Guitar", rather than mixing disparate naming formats.

**The Canon of Multilingual Support.** The Canon of Multilingual Support emphasizes cross-language consistency, ensuring that concept definitions remain identical across languages. For instance, the term "Guitar" and its properties should be mapped uniformly in English, French, and Chinese using multilingual lexical resources like WordNet [103] or UKC [43].

These norms ensure that the taxonomy's linguistic layer is clear, consistent, and adaptable across languages.

### 5.2.4 Notational Plane

The Notational Plane provides a language-independent formalization of the classification hierarchy, using unique identifiers to maintain consistency and traceability.

**The Canon of Unique Identification.** The Canon of Unique Identification en-

sures that every concept is assigned a distinct identifier. For example, the concept of "Guitar" might be assigned the number 12345, guaranteeing its uniqueness in the system.

**The Canon of Semantic Fidelity.** The Canon of Semantic Fidelity requires that identifiers accurately reflect the semantic definition of the associated concept. For instance, identifier 12345 for "Guitar" must correspond precisely to its semantic attributes, such as "six strings" and "acoustic construction".

**The Canon of Hierarchical Consistency.** The Canon of Hierarchical Consistency ensures that identifiers represent the structural relationships within the hierarchy. For instance, the identifier for "String Instruments" might serve as a parent identifier for the "Guitar", reflecting their hierarchical connection.

These norms provide the foundation for a formalized and language-independent taxonomy, ensuring consistency and scalability in technical implementations.

Overall, the Pre-Idea Plane, Idea Plane, Verbal Plane, and Notational Plane collectively form a comprehensive framework for visual object classification. From perceptual concept generation to hierarchical organization, semantic representation, and formalized presentation, these stages ensure a structured and progressive approach to taxonomy design. The canons guiding each phase address the Semantic Gap Problem by systematically aligning visual and semantic attributes. This framework not only enhances the scientific rigor and semantic clarity of classifications but also supports scalable and high-quality image dataset construction.

## 5.3 Glosses

In constructing high-quality image datasets, Glosses serve as a cornerstone for ensuring semantic consistency within the classification framework. Glosses provide each visual object with clear and unambiguous semantic definitions, articulated through language to describe its essential attributes. These semantic

definitions act as a critical bridge connecting visual features to linguistic labels in object classification, significantly enhancing the quality of dataset organization. By integrating glosses into the Verbal Plane, the Semantic Gap Problem (SGP) is effectively addressed. Glosses ensure a precise alignment between visual features and semantic labels, minimizing errors arising from label interpretation discrepancies during annotation. They not only provide annotators with clear guidance but also strengthen semantic consistency across multilingual and multicultural tasks. As such, glosses form the semantic foundation of classification systems.

The core of glosses lies in offering precise semantic definitions for each visual object, articulated through language to highlight their key attributes. This process is a pivotal step in bridging language and visual features within classification frameworks. Below, we elaborate on the core principles and applications of gloss design, including the Canon of Semantic Clarity, the Canon of Consistency, and the Canon of Multilingual Support.

**The Canon of Semantic Clarity**. The Canon of Semantic Clarity is the foundational principle of glosses, ensuring that each classification label is defined with precision. This principle is particularly critical when classifying complex categories such as musical instruments. For example, the gloss for "String Instruments" should describe their defining attribute of sound production through vibrating strings, rather than relying on superficial descriptors like appearance. By accurately describing essential characteristics, glosses effectively delineate boundaries between categories, eliminating semantic ambiguity.

**The Canon of Consistency**. The Canon of Consistency mandates that all glosses within a classification system adhere to uniform naming conventions and linguistic structures. For instance, when describing a group of string instruments, all glosses must follow a consistent format and semantic framework to prevent inconsistencies in classification results. This uniformity not only reduces subjective biases during classification but also provides a stable semantic

foundation for the hierarchical expansion of classifications.

**The Canon of Multilingual**. In terms of multilingual applications, the Canon of Multilingual Support ensures that glosses can be consistently mapped across multiple languages. By leveraging multilingual lexical resources such as WordNet or UKC, glosses for classification labels can achieve uniform semantic expression across languages. For example, the gloss for "Guitar" can consistently convey its core semantic features in English, Chinese, and French, ensuring uniformity regardless of linguistic or cultural differences.

As a semantic backbone for classification systems, glosses support the construction of high-quality image datasets through clear definitions, linguistic consistency, multilingual applicability, and hierarchical alignment. The contributions of glosses to object classification include: (1) Eliminating Semantic Ambiguity: Clear glosses prevent classification errors caused by divergent interpretations among annotators, thereby enhancing the consistency of classification results. (2) Enhancing Multilingual Compatibility: Multilingual glosses ensure the classification framework's applicability across global contexts while maintaining semantic uniformity. (3) Improving Annotation Efficiency: Explicit glosses provide annotators with intuitive guidance, accelerating the annotation process and reducing uncertainty arising from subjective interpretation.

Glosses also maintain close interconnections with other stages of the visual object classification framework. In the Pre-Idea Plane, perceptual accumulation generates initial object concepts, which are organized into hierarchical classifications through the Idea Plane. Within the Verbal Plane, glosses clarify these concepts' semantic definitions, enabling their applicability in multilingual and multicultural scenarios. Finally, in the Notational Plane, glosses are formalized through language-independent identifiers, ensuring semantic stability and consistency.

Through this interconnected, multi-stage framework, glosses ensure semantic clarity and precision, providing both theoretical and practical foundations

for constructing high-quality image datasets. Accurate glosses inject semantic rigor into classification systems, significantly enhancing their usability and scientific reliability.

## 5.4 Visual Properties

In constructing a classification system for visual objects, we propose Visual Properties as the key elements for defining object categories and distinguishing between different objects. The proposed Visual Properties include Visual Genus (G) and Visual Differentia (D), which are used to organize and classify data systematically. In the classification of visual objects, Visual Genus and Visual Differentia play a critical role in ensuring systematic and precise categorization. These concepts are derived from the classical notions of Genus and Differentia established in the study of Objects as Classification Concepts. In this traditional framework, Genus represents the general shared attributes of a category, while Differentia captures the distinctive features that differentiate objects within the same Genus.

In object recognition tasks, the division of Visual Properties enables the hierarchical structure of objects to be defined and presented in a clear and consistent manner. Translating this idea into the domain of visual classification, we adapt these principles to define object hierarchies based on observable visual attributes. Visual Genus establishes the foundational shared properties of a category, ensuring a baseline for grouping objects with common characteristics. Visual Differentia, on the other hand, introduces specific distinguishing features that allow for detailed categorization within the same genus. Together, these properties provide a structured and coherent framework for visual object classification, bridging the gap between linguistic semantics and visual observation.

### 5.4.1 Visual Genus

Visual genus refers to the shared, general visual properties that are common across a set of objects within a category. These properties establish a baseline or a common ground for grouping objects into broader categories. For instance, in the hierarchy of musical instruments, the visual genus for the "String Instrument" category is the presence of taut strings, a feature that is shared by all objects within this group, including guitars, kotos, and dulcimers. The visual genus serves as the starting point for classification, ensuring that objects within the same category exhibit a foundational similarity. It is important to note that visual genus is not specific to a particular instance but rather encapsulates the general characteristics that define a category at its highest level. In this way, the concept of visual genus aligns with the first role of genus in the original framework, ensuring that objects with different visual genus properties are inherently different categories.

### 5.4.2 Visual Differentia

Building upon the baseline established by the visual genus, visual differentia introduces novel and distinguishing visual properties that allow for further differentiation within a category. These properties are unique to specific subcategories or objects and are essential for creating finer distinctions within a classification hierarchy. For example, within the "String Instrument" category, visual differentia such as the number of strings (e.g., six for guitars and thirteen for kotos) and the presence or absence of an input jack (e.g., differentiating acoustic guitars from electric guitars) are used to distinguish between subcategories. Visual differentia ensures that objects within the same visual genus can be systematically categorized based on additional, more specific attributes. This concept aligns with the second role of differentia in the classical framework, wherein objects with the same visual genus but different visual differentia are considered

distinct.

## 5.5 Hierarchy

In this section, we discuss the construction of hierarchical structures based on Visual Properties, which consist of Visual Genus and Visual Differentia. Together, these elements define the categories of visual objects and distinguish between them. The design of this hierarchical structure draws inspiration from the genus-differentia gramework established in Objects as Classification Concepts [44], adapting its principles to suit the requirements of image classification tasks by leveraging observable visual features.

The classification approach using Visual Genus and Visual Differentia ensures the scientific rigor and consistency of the hierarchy. Visual Genus represents shared attributes across a class of objects, while Visual Differentia captures the specific characteristics unique to each object within that class. For instance, the Visual Genus of "String Instruments" can be defined as "instruments with taut strings", and the Visual Differentia of "Guitar" can be specified as "having six strings". Within this framework, each category can be recursively generated using a more general Visual Genus and progressively detailed Visual Differentia. For example, the definition of "Koto" can be traced to "an instrument with taut strings" and further specified as "having 13 strings". This recursive generation facilitates a multi-level classification system suitable for diverse visual objects.

The mainstream approach for generating such hierarchies is the *Faceted Classification* methodology [116], a well-established paradigm for creating classification schemes tailored to various types of information resources. A key innovation of this study is the proposal to use the *WordNet Lexico-Semantic Hierarchy* [103] and its evolved versions as a genus-differentia classification hierarchy, as developed in Computational Linguistics. In our implementation,

this methodology recursively applies genus and differentia to closely align the semantic hierarchy of visual objects with their observable characteristics. By leveraging lexical semantic hierarchies such as WordNet, we provide explicit semantic definitions and visual mappings for each category.

For example, in constructing an image dataset, this hierarchy based on Visual Properties offers a clear classification framework. Starting from the root node "Musical Instruments", its Visual Genus is defined as "devices with sound-producing mechanisms". This hierarchy is then further divided into "Keyboard Instruments", "String Instruments", and "Wind Instruments". Under "String Instruments", the hierarchy further branches into categories such as "Guitar", "Koto" and "Dulcimer". Subcategories of "Guitar" include "Acoustic Guitar" and "Electric Guitar", differentiated by their Visual Differentia of "without the input jack" and "with the input jack", respectively.

The hierarchical structure based on Visual Properties offers three key advantages. First, it enables the reuse of lexical semantic hierarchies to generate high-quality datasets, significantly reducing annotation costs. Second, the subsequent application of classificationist choices C1 and C2 ensures that datasets are inherently aligned with both linguistic and visual semantics, effectively addressing the semantic gap problem (SGP) of many-to-many mappings. Finally, this approach allows annotators to assign labels to images based solely on Visual Differentia, enhancing the efficiency and consistency of the annotation process.

The hierarchical framework based on Visual Properties provides a systematic and scientifically grounded method for constructing image datasets. By applying the principles of Visual Genus and Visual Differentia to classification design, this approach delivers explicit semantic definitions and visual boundaries for each label in the dataset. Whether in theoretical research or practical applications, this method demonstrates significant advantages, particularly in resolving semantic gaps and improving classification accuracy. Ultimately, it establishes a robust foundation for constructing high-quality image datasets and

*Word*: piano

| | |
|---|---|
| Sense ID: 03934354; piano#1 | Sense ID: 04998511; piano#2 |
| Synset = piano, forte-piano, ... | Synset = piano, pianissimo, ... |
| Gloss = "a keyboard instrument that is played by depressing keys" | Gloss = "(music) low loudness" |

offers new directions and methodologies for future data-driven computer vision tasks.

## 5.6 Results

Based on the content introduced earlier, we will provide a detailed explanation of how the classificationist employs canons to consequence choices C3 and C4, thereby generating a *genus-differentia classification hierarchy* which suitably organizes the meaning of labels. We draw upon the *WordNet lexico-semantic hierarchy* [103] and its evolutions, as developed in computational linguistics, to design the genus-differentia classification hierarchy. For instance, we provide the following definition of *piano*:
From this definition, we know that the label *piano* has two unique meanings, "piano#1" and "piano#2", called *senses*, each associated with a unique sense *identifier*, e.g., `03934354`. Different synonymous words for the same sense are clustered into *synsets*. Further, each synset is associated with a *gloss* which is a natural language definition of the concept expressed by the synset in terms of genus-differentia. For instance, the synset {*piano, forte-piano, ...*}, associated to the sense "piano#1", is defined by a gloss which encodes its genus *"keyboard instrument"* and its differentia *"depressing keys"*.

We use the subset of musical instruments that we mentioned in Section 3.4 as an example to illustrate how the classificationist defines object categories recursively based on visual properties, including visual genus and visual differentia.

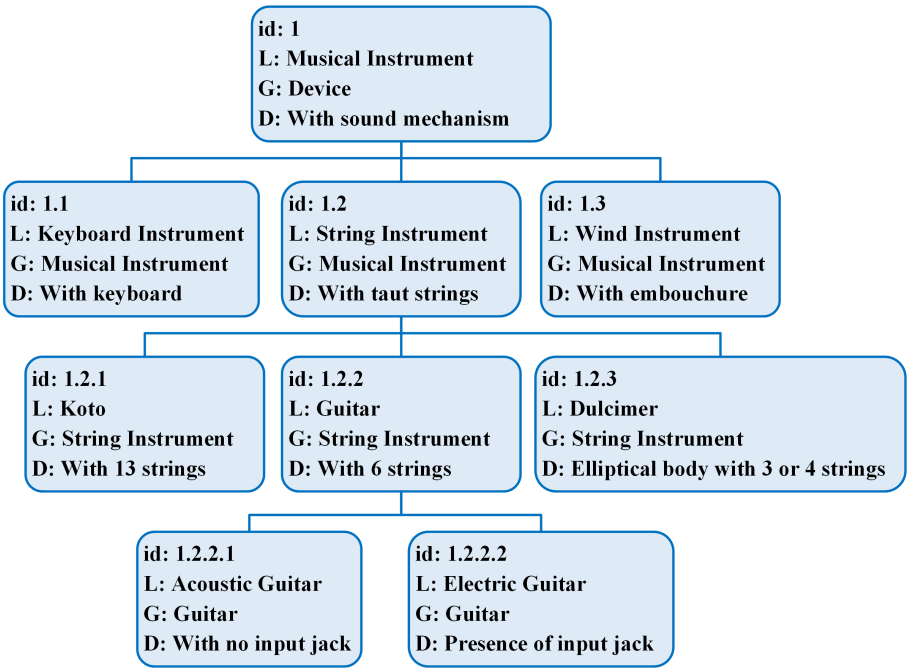


Figure 5.1: The hierarchy defined by classifications in vTelos, exemplified by the musical instrument subset. Each node is associated with a unique linguistic identifier (id), the label (L) of the category in ImageNet, a Genus (G), and a Differentia (D).

This process creates a detailed and consistent hierarchy from the top-level class to specific subclasses, as shown in Figure 5.1. The hierarchy that the classificationist designed also includes a comprehensive representation of the salient attributes of objects. For instance, the visual genus of "Dulcimer" is defined as "String Instrument", while its visual differentia is specified as "Elliptical body with 3 or 4 strings". Similarly, the visual genus of "Koto" is also "String Instrument", but its visual differentia is refined to "With 13 strings". These precise definitions enable the classification system to clearly distinguish between different subclasses within the same class, thereby providing annotators with clear guidance and significantly reducing annotation errors caused by semantic ambiguity.

The hierarchy constructed through the classificationist's definition process offers several notable advantages:

- Systematic Structure: By defining visual genus and visual differentia at

each level, a top-down classification framework is established, ensuring logical consistency between semantic and visual properties.

- Precision: Fine-grained category definitions reduce semantic ambiguity and provide high-quality training data for machine learning models.
- Scalability: The recursive definition approach allows new object classes to be seamlessly integrated into the existing hierarchy.

In our study, the classificationists designed visual property-based hierarchical classifications not only for musical instruments but also for subsets such as "birds" and "vehicles", encompassing a total of 44 categories. Acting as classificationists, we defined comprehensive glosses, visual genus, and visual differentia for each category. All the definitions are compiled in Appendix A. These definitions demonstrated exceptional effectiveness in addressing the semantic gap problem and improving classification accuracy, providing a reliable direction for future image dataset design and optimization.

The visual properties defined by classificationists based on the canons, along with the constructed hierarchical classification, resolved inconsistencies between semantic and visual features, thereby alleviating the SGP. Moreover, they established a solid foundation for designing and annotating image datasets. By combining theory with practice, this approach offers a novel perspective for building high-quality datasets.

## 5.7 Summary

In this chapter, we provided a detailed explanation of the responsibilities of the Classificationist, one of the key roles in the proposed methodology for constructing image datasets. The Classificationist operates within the framework of faceted classification, which aims to enhance the semantic consistency and classification accuracy of datasets through the systematic design and organization

of classification systems. At the core of this framework are three fundamental components: Canons, Glosses, and Visual Properties, which together form the theoretical foundation for building high-quality image datasets.

As the central figure in the classification process, the Classificationist is responsible for defining the classification structure and establishing semantic boundaries, ensuring both scientific rigor and consistency in the classification system. Canons offer the foundational principles for building classification systems, encompassing aspects such as feature selection, feature sequencing, category grouping, and hierarchical chaining. Glosses provide precise semantic definitions for classification labels, reducing annotation errors, enhancing multilingual compatibility, and ensuring global applicability. Visual Properties, defined through Visual Genus and Visual Differentia, are used to organize and differentiate object categories. Visual Genus captures the shared properties of a category, while Visual Differentia identifies specific characteristics that refine and clarify classifications. Additionally, this chapter presented a practical application of the Classificationist's work, showcasing the definition of image dataset categories based on Canons, Glosses, and Visual Properties.

In summary, the proposed Canons, Glosses, and Visual Properties collectively establish a comprehensive and systematic classification framework, guiding the Classificationist in defining dataset categories. By seamlessly integrating these components, this approach effectively addresses the semantic gap problem, ensuring consistency between semantic and visual attributes, and providing a robust foundation for training machine learning models. This framework not only offers theoretical insights but also demonstrates significant practical advantages, paving the way for innovative perspectives and methods in the design and optimization of future datasets.

# Chapter 6

# Classifier: Annotation Process

## 6.1 Introduction

In the previous chapter, we provided a detailed overview of the role of the Classificationist and constructed a classification hierarchy based on their definitions. Building upon this foundation, this chapter focuses on the role of the Classifier and their core responsibilities in the image annotation process. The Classifier's primary task is to perform object localization and classification, namely C1 & C2 we introduced in Section 4.2, according to the annotation system and guidelines provided by the Classificationist. For large-scale image datasets, this process is further enhanced by the integration of crowdsourcing techniques, leveraging collective intelligence to improve annotation efficiency and quality. This study is based on our publications [49, 28].

The Classifier serves as the central executor in the image annotation process, as an annotator. Their main responsibility is to identify target objects in images and assign the labels that best match the semantic definitions. During the object localization stage (C1), the Classifier identifies the primary object in complex visual scenes. For example, in an image depicting a stage scene, the Classifier relies on the prioritization rules and definitions established by the Classificationist to select "Guitar" as the primary object rather than "Microphone" or "Speaker". This process effectively minimizes annotation errors caused by sub-

jective biases.

In the visual classification stage (C2), the Classifier further analyzes the visual attributes of the target object. For example, when annotating an image of an acoustic guitar, the Classifier does not directly stop at the label "Guitar." Instead, it will continue to identify the object's visual properties, such as "six strings". Then the Classifier identify it as a "Acoustic Guitar", which is matched with the semantic properties defined by the Classificationist. This process significantly reduces ambiguities resulting from unclear object features.

The annotation process involving the Classifier is an iterative optimization procedure comprising three levels of loops: the top-level loop, the vertical loop, and the horizontal loop. In the top-level loop, the Classifier performs object localization, while in the vertical and horizontal loops, further refinement of label selection is achieved through visual classification. Object localization at the top level ensures precise identification of the main object in the image, whereas the vertical and horizontal loops iteratively refine the classification by analyzing the object's visual properties. This iterative approach allows the annotation process to continually improve and produce high-quality labels under the semantic framework defined by the Classificationist.

To address the demands of annotating large-scale datasets, the Classifier's tasks are extended and optimized through crowdsourcing techniques. The construction of non-generative image datasets in the field of computer vision typically involves crowdsourcing techniques [90]. During the crowdsourcing annotation process, annotators are presented with images and asked to select the most appropriate label from a set of words that are predefined categories, as introduced in some datasets, e.g., ImageNet [26] and CIFAR [86]. The labelling of images in such datasets is obtained by requiring annotators to verify if the image matches a predefined category or synonym set. For instance, if the synonym set is "bear," all images deemed by annotators to contain "bear" are categorized under this label. A widely discussed issue in this construction methodology is

the many-to-many matching problem between categories and images [145, 48]. This issue, where an image containing multiple objects is described by a single label, introduces confusing erroneous information into object recognition models, negatively impacting recognition accuracy. The advent of datasets designed for object detection tasks, such as COCO [93] and PASCAL VOC [32], has somewhat alleviated the problem of information clutter caused by multiple objects. The method of marking specific objects with polygonal boxes allows images to be labelled one-to-one. However, the approach to single-object category labelling remains unchanged, with the annotation process still matching images to predefined words.

The complexity and polysemy of natural language indeed introduce ambiguity when describing the visual information in images and their semantic interpretation, which is still caused by the semantic gap problem. This gap exacerbates the challenge of annotation inconsistency in image datasets. A fundamental issue with existing annotation methodologies is the lack of a clearly defined process for categorizing and standardizing image annotations. As a result, there is a pressing need for an image annotation methodology that minimizes the impact of subjective interpretations by annotators, aims to ensure consistency across annotators with diverse backgrounds, and maintains high quality in image datasets.

Our annotation method incorporates the classification hierarchy defined by the Classificationist, providing clear guidance rules for Classifiers, namely annotators, within the crowdsourcing framework. By requiring annotators to verify the match between images and predefined visual properties, this method ensures a more efficient and consistent annotation process.

In this chapter, we will explore in detail how the Classifier performs annotation tasks based on the hierarchy defined by the Classificationist and how crowdsourcing techniques are utilized to optimize the process. We will also examine the advantages of this approach in addressing the semantic gap problem

and improving annotation efficiency, supported by specific experimental cases and application results. By combining theory and practice, this chapter aims to provide effective guidance for the construction of high-quality image datasets.

## 6.2 The Iterative Annotation Process

In this section, we detail the image annotation methodology. This is an iterative refinement process, which includes three loops, namely the top-level loop, the vertical loop, and the horizontal loop. The C1 step, object localization, is applied in the top-level loop, and the C2 step, visual classification, produced by the vertical loop and horizontal loop. Choice C2, as carried out by the classifier, takes in input the hierarchy generated by Choices C3, C4 and associates each single image to a node in the hierarchy. The resulting algorithm is reported in Algorithm 1. This algorithm takes in input a hierarchy $H$, e.g., the hierarchy in Fig.5.1, and an Image $I$ and produces in output the image label $L$ (Lines 3, 12). The annotators are also provided with the conceptual identifiers $id$, *visual genus*, and *visual differentia* $vD$ of each node $N$. Note that the hierarchy $H$ allows easy viewing and navigation of subcategories. We detail the three loops below.

**The Top-level Loop.** The top-level loop is designed to continuously offer *good images* for the iterative annotation method. In this process, the *object localization* step is introduced to avoid *object ambiguity*. An object localization model [121] is introduced to automatically locate objects by machine in an image, and crop the image based on the coordinates of objects, aiming to obtain images with a single object. As a result, we can obtain multiple single-object images from one multi-object image. Note that, although the square bounding box sometimes contains parts of other objects in the cropped image, the main object of these images will be more defined than the original image, thus, we treat the cropped image as a single-object image. Then, all images are input to the

**Algorithm 1** $L = VisClassify(I)$

$H$ = Tree of ($N$);

$N =< id, vD >$; {For a category node, $id$ is the conceptual identifier, $vD$ is the visual differentia properties.}

**Input**: Image $I$;

**Output**: Image label $L$;

$R = getRoot(H)$;
**if** $askDif(I, R.vD) = False$ **then**
  **return** $L =$ "$Discharged$";
**end if**
$L = R.vD$;
**while** $hasChild(R) = True$ **do**
  **while** $C \in getChild(R)$ **do**
    **if** $askDif(I, C.vD) = True$ **then**
      $R = C$; $break$;
    **end if**
  **end while**
  $L = R.vD$;
**end while**
**return** $L$;

following vertical loop and horizontal loop one by one for annotation.

**The Vertical Loop.** The goal of the vertical loop is to iteratively refine the label of the input image through layer-by-layer computation to label it precisely. There is also a hierarchy that is input at the same time as the image, which will be gradually enriched by adding nodes and samples in the current iterative process. When inputting a new object, the similarity is applied to compare with the categories already stored in the hierarchy, and the category with the highest similarity is regarded as the "candidate". Next, humans join this loop to determine whether the object has the common *visual genus* as the "candidate". If the answer is "False", proceed to compare this object to other categories at the same layer with "candidate" in the hierarchy, and enter the horizontal loop until "True" is obtained. None of "True" obtained means the object does not share a

*visual genus* with any categories, then it will be labeled as a new category and stored in the hierarchy. During this process, the machine is responsible for recommending initial labeled options and performing the method, while humans take charge of determining the *visual genus* to decide whether further labeling refinement is needed.

**The Horizontal Loop.** The goal of the horizontal loop is to label the most refined category for the input object by comparing them within a domain category. It is triggered by the vertical loop and compares the input object with the subcategories of the candidate (if the candidate category has no subcategories, the input object is labeled as the candidate category). The process starts from the subcategories with the highest similarity of the object and asks humans whether they have *visual differentia*. If the answer is "False", it means that they belong to the same category, and the input object is labeled as the current subcategory; if the answer is "True", the same comparison is proceed with the next subcategory. If all subcategories in the candidate has *visual differentia* from the input image, it means the object does not belong to any subcategory, and it is labeled as a new category and added to the hierarchy as a new subcategory of the candidate. In the process, we complete the annotation and enrich the hierarchy at the same time. This incremented hierarchy will also continue to be utilized in the next top-level loop.

**Observation and Analysis** There are two important observations. Firstly, the fact that, though we have a detailed set of canonical principles for ensuring the *visual subsumption hierarchy* to be ontologically thorough, the task becomes particularly critical due to the tradeoff between the appropriate vertical and horizontal choice in uniquely classifying an object. The choice must be guided by the specific object recognition task that must be performed by the model trained using the dataset generated. In other words, there cannot be a *fits-it-all* dataset. Instead, we envisage a future where this methodology will allow the construction of datasets with clear and precisely specified *semantic* properties, which

will then be introduced in the ML models by using them for training. Secondly, as a consequence of the first observation, human supervision can often be necessary for determining the exact (succession of) differentia in sync with the egocentric hierarchy in the mind of the user. The key observation underlying both observations, also factoring in other phases, is that the faceted classification process, while (of course) not eliminating human subjectivity, does provide the guidelines for *enforcing a one-to-one mapping* between visual and linguistic properties. Overall, although the proposed iterative refinement process of image annotation still suffers from human subjectivity, it does provide the guidelines for *enforcing a one-to-one mapping* between visual and linguistic properties.

## 6.3 Crowdsourcing Image Annotation

This section details its implementation within a crowdsourcing framework. Our method utilizes hierarchical classification [131] and detailed analysis of visual properties [48], engaging annotators in answering a series of questions related to the visual properties, i.e., *visual genera* and *visual differentia* of object categories. These questions are linked to a predefined object hierarchy encompassing all inheriting superordinate categories of the objects to be annotated.

### 6.3.1 Label Preparation.

Before initiating the annotation phase, the first two steps of the Image Annotation Methodology are completed, i.e., *Label Generation* and *Label Disambiguation*. Drawing upon established knowledge bases, we construct an object hierarchy with all object categories in a dataset. To ensure clarity and eliminate linguistic ambiguities, we define each category with a set of specific visual properties based on the guidance of canons [77, 117, 116, 118]. This preparation lays the groundwork for the annotation process.

### 6.3.2 Image Preparation.

The primary goal of this phase is to supply quality images for annotation continuously. We screen "good images[1]" for the dataset. It is worth mentioning that, in our subsequent experiments, to reduce the cost of image collection and compare the accuracy and effectiveness of annotations with existing image datasets, we utilized data from pre-existing image datasets and re-annotated these images. In this project, we selected ImageNet [26] as the foundation, extracted subsets and images from ImageNet, and re-annotated these images. The *Object Localization* step applies machine learning models [121] to automatically identify and crop images, focusing on offering single-object images. In the absence of significant noise, we consider these cropped images as single objects, even if minor parts of other objects are present. This step significantly reduces object ambiguities and prepares for finer categorization in subsequent phases.

### 6.3.3 Crowdsourcing Annotation Collection.

After preparing labels and images, the classifiers proceed with *Visual Classification*. Given the diverse backgrounds of crowdsourcing workers[2], we optimize human-machine interaction for high-quality annotations. Annotators are provided with a hierarchy $H$ of all categories, along with the conceptual identifiers $id$, *visual genus*, and *visual differentia* $vD$ of each node $N$. This hierarchy allows easy viewing and navigation of subcategories. The interface presents the image $I$ alongside questions about its visual properties. These questions, designed to iteratively refine labels, are automated based on the object hierarchy and annotator responses, as detailed in Algorithm 1.

In this process, the vertical loop aims to refine the labels of input images iteratively, moving from root to leaf nodes in the object hierarchy to accurately

[1] A "good image" clearly features a single main object, captured optimally with minimal noise or distortion, ensuring clear visibility of its defining visual properties.

[2] We refer to the crowdsourcing workers as "annotators" in the following paper.

categorize them. By asking annotators if the image shares a *visual genus* with specific categories in the hierarchy, candidate categories are determined, and then triggering the corresponding horizontal loop. The goal of the horizontal loop is to ascertain the most refined category within its domain for the input image, involving comparisons with subcategories of the candidate. Annotators identify *visual differentia* to decide the specific subcategory, ensuring accurate and reliable labelling through these interrelated loops.

### 6.3.4 Quality Control.

During the annotation phase, each image was annotated by three individuals to enhance reliability. In the quality assessment phase, labels that received consistent annotations from at least two out of the three annotators were adopted as the final label for the dataset. In cases where there was no consensus among the initial three annotators, a fourth annotator was introduced to reassess and provide an additional annotation for the image. This step was crucial in ensuring the accuracy and consistency of the labels in our dataset, thereby enhancing the overall quality of the annotated data.

This innovative crowdsourcing methodology combines machine learning efficiency with human cognition, creating a dynamic, improving process. By using an object hierarchy and focusing on visual properties, this method ensures precise and consistent image annotations, which is crucial for advancing machine learning and AI in object recognition.

## 6.4 Crowd sourcing experiments

To validate our image annotation methodology in a crowdsourcing framework, we conducted experiments focusing on machine learning results and annotator consistency to assess dataset quality.

Table 6.1: The number of images for categories we obtained by the proposed image annotation methodology from three groups of annotators, where "I, II, ..., XII" denote twelve categories. The "Discharged" indicates instances where annotators did not annotate the image into any of the twelve specific categories but rather assigned it to an upper-level category, e.g., its parent category. "Alpha" refers to Krippendorff's alpha measure, a statistical tool for assessing the reliability of these annotations.

| Categories | I | II | III | IV | V | VI | VII | VIII | IX | X | XI | XII | Discharged | Alpha |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Annotators group 1 | 98 | 101 | 86 | 102 | 97 | 103 | 92 | 95 | 96 | 96 | 90 | 95 | 49 | |
| Annotators group 2 | 91 | 105 | 93 | 98 | 101 | 98 | 91 | 94 | 87 | 104 | 88 | 92 | 58 | 0.934 |
| Annotators group 3 | 95 | 100 | 91 | 101 | 99 | 102 | 95 | 92 | 90 | 98 | 93 | 98 | 46 | |

Table 6.2: Results of implementing three different image annotation methodologies, including inter-class agreement evaluation with Krippendorff's alpha measure, time cost on single tasks, and payment to one annotator for one task. To evaluate the proper number of images per task, we present a comparison of results for tasks with 50 images and 100 images.

| Annotation methodology | 50 images per task | | | 100 images per task | | |
|---|---|---|---|---|---|---|
| | Alpha | Time (min) | Payments(£/p) | Alpha | Time (min) | Payments(£/p) |
| Method A | 0.912 | 4.12 | 1 | 0.896 | 8.27 | 2 |
| Method B | 0.937 | 5.08 | 1.5 | 0.914 | 9.22 | 3 |
| Method C (Ours) | 0.974 | 5.47 | 1.5 | 0.958 | 9.47 | 3 |

### 6.4.1 Experimental Setup

**Image and Labels Collection.** A total of 1200 images were collected, representing 12 different object categories. These categories spanned three domains: birds (5), vehicles (3), and musical instruments (4). Guided by faceted classification theory [116, 117], we defined a pair of *visual genus* and *visual differentia* for each category and organized them into a hierarchy to guide the annotation process.

**Crowdsourcing development.** We developed a new webpage[3] linked to a crowdsourcing platform to collect user annotations by asking a series of questions. The questions were dynamically updated based on the hierarchy and feedback from the annotators. Building on the proposed annotation method-

[3]The webpage will be published, currently anonymous due to blind review.

ology, we emphasized flexible task deployment, using Prolific[4] as our primary crowdsourcing platform. All annotator responses were recorded on our system. Image labels are selected based on the appropriate category level in the dataset's hierarchy to meet the dataset's specific requirements, typically using the most fine-grained level.

### 6.4.2 Inter-annotator agreement.

We introduce Krippendorff's Alpha[71] to measure inter-annotator agreement. We assign 1200 images across twelve object categories to three annotator groups, with each image annotated by three different annotators. Table 6.1 shows the number of images per category in the obtained dataset. A consistency score of 0.934 indicates a high level of agreement, validating the reliability of our annotation method.

### 6.4.3 Experiments on crowdsourcing setting

Before deploying large-scale crowdsourcing, it's crucial to establish the optimal setup for the newly proposed annotation methodology to ensure high-quality results. We designed experiments to collect annotations under various crowdsourcing parameters, aiming to identify the most effective interaction settings by assessing annotation quality and analyzing annotator feedback.

**Experiment 1: Annotation Methodology Evaluation.** This experiment is designed to evaluate image annotation quality by comparing three annotation methodologies, including:

- Method A: Name labels only. Annotators label images using only the category names (present as words or phrases), as the existing annotation methodology used, e.g., ImageNet.

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

- Method B: Name labels in an object hierarchy. Annotators label images using names with a predefined hierarchy.
- Method C: Visual property labels in an object hierarchy. Annotators label images using visual properties (*visual genus* and *visual differentia*) within a predefined hierarchy, namely our proposed method.

The results are shown in Table 6.2. Method A resulted in quicker annotations but lower inter-class agreement. This illustrates that Method A relies heavily on semantic processing, which can be subjective and ambiguous, leading to semantic gaps among annotators. Method B improved agreement, while Method C achieved the highest accuracy and consistency, though it required more time. This phenomenon can be explained by the theory of cognitive science [30] that deeper information processing, such as analyzing *visual genus* and *visual differentia*, leads to better memory and understanding. Methods B and C reduced cognitive load [141] by providing structured frameworks for annotation, with Method C further reducing ambiguity by integrating visual properties, leading to higher-quality annotations. This experiment demonstrates the importance of structured visual property-based annotation methods for improving dataset quality. Although more time-consuming, our proposed methodology offers a robust framework for accurate image annotation, which is crucial for tasks requiring high annotation accuracy and consistency.

**Experiment 2: Costs Efficiency Analysis.** We compared the cost efficiency of three image annotation methodologies introduced in Experiment 1. We assessed annotation consistency (Krippendorff's Alpha [107]), time taken, and payments made to annotators for 50 and 100 image tasks. As shown in Table 6.2, key findings include:

- Time vs Quality Trade-off: Method C consistently had the highest annotation consistency, despite taking slightly more time and payment. The higher Alpha values indicate a worthwhile trade-off for better quality.

- Learning Curve Efficiency: Annotation speed increased from 50 to 100 images as annotators became more familiar with the hierarchy and visual properties. This indicates that efficiency improves over time, suggesting that our method can be scaled to annotate large image datasets.

- Cost Implications: The higher consistency of Method C justifies the increased cost per image, especially for applications where accuracy is critical.

As a result, while our annotation method slightly requires more time and costs, its superior consistency and accuracy justify the additional resources, especially where high data quality is essential.

**Experiment 3: Optimal Number of Images per Task.** This experiment explored the balance between task length and annotator motivation in crowdsourcing. We tested tasks with 10, 30, 50, and 100 images and collected annotators‘ feedback. Feedback indicated that tasks with 10 images were completed quickly but offered low compensation, reducing their appeal. Conversely, tasks with 100 images led to boredom and decreased enthusiasm. Most annotators (8 out of 10) preferred tasks with 50 images.

This feedback aligns with our understanding. Tasks with too many images cause cognitive overload [38] and reduced annotation quality due to waning attention, while too few images fail to engage crowdsourcing workers. Balancing task size, time investment, and compensation is crucial for maintaining annotators' motivation. A moderate task size (around 50 images) keeps annotators engaged without overwhelming them, ensuring sustained attention and higher-quality annotations.

## 6.5 Summary

In this chapter, we explored the critical role of the Classifier in the image annotation process, focusing on their responsibilities in object localization and classification (C1 and C2) as guided by the classification hierarchy defined by the Classificationist. The Classifier's task of identifying target objects and assigning precise labels based on semantic definitions ensures both accuracy and consistency in annotations. By leveraging iterative refinement processes—comprising top-level, vertical, and horizontal loops—the Classifier systematically reduces annotation errors and enhances the quality of labeled data.

We also examined the integration of crowdsourcing techniques to scale annotation efforts for large datasets. By incorporating the Classificationist-defined hierarchy, crowdsourced annotators are provided with clear and consistent guidance, resulting in more efficient and reliable annotations. This approach not only addresses the semantic gap problem by aligning visual features with semantic properties but also ensures a streamlined annotation process suitable for large-scale applications.

Through the combination of theoretical underpinnings and practical methodologies, this chapter highlighted how the collaborative efforts of Classifiers and Classificationists, supported by crowdsourcing, contribute to the development of high-quality image datasets. The framework presented here offers a robust solution to the challenges of semantic ambiguity and annotation inconsistency, laying the foundation for improved machine learning models and their applications in computer vision tasks.

# Chapter 7

# Evaluation and Application on vTelos

## 7.1 Introduction

In this chapter, we conduct a comprehensive evaluation of the dataset constructed using the proposed image dataset construction methodology *vTelos*. The dataset, named vTelos-img, is assessed from multiple perspectives, including quality validation, annotation efficiency analysis, and assessment of practical application, to verify the scientific rigor and practicality of the *vTelos* framework. The vTelos-img dataset comprises 44 categories across three subdomains: “Birds”, “Vehicles”, and “Musical Instruments”, totaling 55,086 images. Through systematic experimental design and result analysis, we aim to provide insights for high-quality dataset construction and optimization, showcasing the potential of the *vTelos* methodology in the domain of image classification. This study is based on our publications [48, 128, 18].

First, we conducted a quality evaluation, focusing on inter-annotator agreement to assess the reduction of subjective biases. By introducing Krippendorff’s alpha as a consistency measure, we verified the reliability and coherence of the proposed annotation approach. Additionally, to analyze annotation efficiency and cost, we compared the time spent by non-expert annotators on various annotation tasks, exploring the economic advantages of the *vTelos* method in large-scale dataset annotation. Second, we performed machine learning ex-

periments to evaluate the impact of the vTelos-img dataset on model performance improvement. Using multiple state-of-the-art machine learning models, we compared their performance when trained on the original ImageNet subset and the refined dataset optimized through the *vTelos*. This analysis highlighted how improved data quality enhances model generalization capabilities and accuracy. To further evaluate the core components of the *vTelos* methodology, we designed ablation studies focusing on object localization and visual classification, assessing their respective contributions to final dataset quality.

Through these experiments, this chapter not only validates the high-quality construction of the vTelos-img dataset but also explores its potential in practical applications. The experimental results provide valuable insights for addressing the semantic gap problem and improving annotation processes while laying the foundation for future image dataset design and optimization based on the *vTelos* methodology.

## 7.2 Quality evaluation

### 7.2.1 Inter-annotator agreement

We applied 450 of the 3660 images populating the Musical Instrument subset in ImageNet, introduced in Section 3.4. This subset includes nine categories, 50 images per category. To minimize data bias, these 450 labelled image dataset, let us call GT1 (where GT stands for *Ground Truth*) was randomly generated. Then, a second dataset GT2 was generated by an expert annotator $EA_1$ by relabelling the 450 images of GT1 using Choice C2 only, as described in Section 4.2. We then selected 16 volunteer *non-expert annotators*, with no previous annotation experience. The annotators were university students and collaborators, with an average age of 29 years. To maximize diversity, we selected people from different disciplines and from 3 continents and 7 countries. The annotators were organised in 2 groups of 8 people, each annotator relabelling GT1 and

Table 7.1: Inter-annotator agreement evaluation with Choice C2 only. The images in GT1 have ImageNet labels. GT2, GT3, GT4 images have labels selected by the expert annotator, and by the 8+8 non-expert annotators in Groups 1 and 2, respectively. $\text{NA}_{j,i}$ stands for annotator $i$ of Group $j$. Id. is the label identifier, as from Figure 3.2. The last row reports the value of Krippendorff's alpha.

| Id. | GT1 | GT2 | GT3 (Annotation via Differentia labels) | | | | | | | | GT4 (Annotation via Category labels) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Differentia | $\text{NA}_{1.1}$ | $\text{NA}_{1.2}$ | $\text{NA}_{1.3}$ | $\text{NA}_{1.4}$ | $\text{NA}_{1.5}$ | $\text{NA}_{1.6}$ | $\text{NA}_{1.7}$ | $\text{NA}_{1.8}$ | Categories | $\text{NA}_{2.1}$ | $\text{NA}_{2.2}$ | $\text{NA}_{2.3}$ | $\text{NA}_{2.4}$ | $\text{NA}_{2.5}$ | $\text{NA}_{2.6}$ | $\text{NA}_{2.7}$ | $\text{NA}_{2.8}$ |
| 1 | 50 | 41 | with Sound Mechanism | 33 | 12 | 27 | 25 | 28 | 29 | 12 | 18 | Musical Instrument | 16 | 42 | 100 | 17 | 27 | 20 | 19 | 26 |
| 1_1 | 50 | 123 | with Taut Strings | 46 | 97 | 71 | 133 | 112 | 83 | 62 | 79 | Stringed Instrument | 162 | 78 | 0 | 115 | 144 | 106 | 161 | 74 |
| 1_1_1 | 50 | 34 | with 6 Strings | 37 | 13 | 40 | 34 | 58 | 34 | 19 | 31 | Guitar | 77 | 18 | 41 | 40 | 13 | 9 | 0 | 8 |
| 1_1_1_1 | 50 | 40 | with No Input Jack | 66 | 77 | 53 | 54 | 26 | 67 | 68 | 50 | Acoustic Guitar | 13 | 81 | 66 | 43 | 70 | 70 | 82 | 71 |
| 1_1_1_2 | 50 | 43 | with Input Jack | 54 | 61 | 67 | 43 | 28 | 45 | 82 | 78 | Electric Guitar | 74 | 65 | 86 | 44 | 70 | 71 | 68 | 74 |
| 1_1_2 | 50 | 31 | Elliptical body with 3 or 4 strings | 61 | 37 | 36 | 32 | 29 | 30 | 37 | 30 | Dulcimer | 0 | 6 | 31 | 31 | 0 | 16 | 6 | 46 |
| 1_1_3 | 50 | 27 | with 13 Strings | 42 | 32 | 42 | 16 | 18 | 44 | 53 | 45 | Koto | 0 | 47 | 47 | 39 | 0 | 41 | 0 | 43 |
| 1_2 | 50 | 47 | with Keyboard | 49 | 46 | 41 | 49 | 43 | 47 | 50 | 46 | Keyboard Instrument | 41 | 49 | 0 | 44 | 40 | 47 | 52 | 44 |
| 1_3 | 50 | 47 | with Embouchure | 62 | 60 | 60 | 63 | 50 | 55 | 60 | 59 | Wind Instrument | 61 | 57 | 55 | 62 | 54 | 61 | 60 | 59 |
| | 0 | 17 | Unrecognised | 0 | 15 | 13 | 1 | 58 | 12 | 7 | 14 | Unrecognised | 6 | 7 | 24 | 15 | 32 | 9 | 2 | 5 |
| all | 450 | 450 | Krippendorff's alpha | 0.5973 | | | | | | | | Krippendorff's alpha | 0.5047 | | | | | | | |

generating a dataset as follows:

- GT3, as created by each member of Group 1 The annotators were asked to perform the annotation in two steps. First, they had to annotate images based on Choice C2. Here they were allowed to create a new category "*Discharged*" for all the images they could not classify. Then, in the second step, they were asked to label the categories generated in the first step S2 with their "most suitable" label, not necessarily an ImageNet label (which was not provided).

- GT4, as created by each member of Group 2. This group was instructed to annotate images using the ImageNet labels plus the label "Discharged".

The motivation for focusing $\text{EA}_1$ and Group 1 on Choice C2 was that of highlighting the use of the differentia in place of the label. The results are reported in Table 7.1. Notice how this table reports the number of images associated with each label for all four datasets and for each annotator working on GT3 and GT4. As it can be seen, the number of images (and the images) associated with the same label are different. We measure the inter-annotator agreement of GT3 and GT4 using *Krippendorff's alpha* [84, 2] (last row). Krippendorff's alpha is $0.5973$ for GT3 and $0.5047$ for GT4, that is, GT3 is around $18\%$ higher than

Table 7.2: Inter-annotator agreement with Choices C1, C2 (full `vTelos-img`), generating two GT2* and eight GT3* datasets.

| Index | Categories | GT2* | | GT3* (Only Single-Object images) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\mathbf{EA}_1$ | $\mathbf{EA}_2$ | $\mathbf{NA}_{1.1}$ | $\mathbf{NA}_{1.2}$ | $\mathbf{NA}_{1.3}$ | $\mathbf{NA}_{1.4}$ | $\mathbf{NA}_{1.5}$ | $\mathbf{NA}_{1.6}$ | $\mathbf{NA}_{1.7}$ | $\mathbf{NA}_{1.8}$ |
| 1 | with Sound Mechanism | 17 | 17 | 16 | 2 | 5 | 11 | 11 | 11 | 4 | 4 |
| 1_1 | with Taut Strings | 42 | 42 | 21 | 41 | 32 | 43 | 41 | 39 | 28 | 32 |
| 1_1_1 | with 6 Strings | 21 | 20 | 21 | 8 | 23 | 19 | 26 | 24 | 11 | 15 |
| 1_1_1_1 | with No Input Jack | 21 | 22 | 31 | 34 | 29 | 28 | 18 | 32 | 33 | 29 |
| 1_1_1_2 | with Input Jack | 22 | 22 | 24 | 32 | 31 | 21 | 20 | 21 | 39 | 39 |
| 1_1_2 | Elliptical body with 3 or 4 strings | 13 | 13 | 24 | 13 | 13 | 14 | 12 | 13 | 15 | 14 |
| 1_1_3 | with 13 Strings | 12 | 12 | 12 | 7 | 10 | 8 | 9 | 10 | 13 | 9 |
| 1_2 | with Keyboard | 33 | 33 | 33 | 33 | 29 | 34 | 29 | 31 | 34 | 33 |
| 1_3 | with Embouchure | 10 | 10 | 20 | 20 | 21 | 23 | 14 | 15 | 19 | 21 |
| | Unrecognised | 11 | 11 | 0 | 12 | 9 | 1 | 22 | 6 | 6 | 6 |
| all | Krippendorff's alpha | 0.9877 | | 0.7595 | | | | | | | |

GT4. This is an important improvement but still not good enough.[1] The cause of this, not fully satisfactory, result was the Multi-object images which are more than 58% of the total (2125 out of 3660 images). We have therefore performed a second experiment where Choice C1, C2 were enforced and where MOI images were proposed multiple times with a bounding box around the object to be annotated. This experiment was performed by Group 1 (generating eight datasets GT2*) and two expert annotators $\mathrm{EA}_1$, $\mathrm{EA}_2$ (generating two datasets GT3*). Table 7.2 reports the results obtained. Here, Krippendorff's alpha goes up to $0.7595$ among non-expert annotators and $0.9877$ between the two experts, namely up to near-perfect agreement.

An interesting final observation is that in Group 2, but not in Group 1, some annotators were unable to annotate some images (see, e.g., the "0"s in Table 7.1). Two such examples are "Dulcimer" and "Koto".[2] This observation is also confirmed in Table 7.3, which reports some example labels provided during the second part of the Group 1 experiment. The observation here is that, despite properly labeling images via differentia, some annotators did not know

[1] As from [71][Table 2, p. 417], a value of alpha in the range $(0.6, 0.8]$ indicates substantial agreement while an index in the range $(0.8, 1]$ indicates near-perfect agreement.

[2] Koto is a Japanese instrument.

Table 7.3: Category Labels suggested by Group 1 annotators.

| Anno. | Acoustic Guitar | Dulcimer | Koto |
|---|---|---|---|
| $\mathbf{NA}_{1.1}$ | Guitar | Appalachian Dulcimer | Biwa |
| $\mathbf{NA}_{1.2}$ | Guitar | IDK (I Don't Know) | IDK |
| $\mathbf{NA}_{1.3}$ | Wooden guitar | IDK | Zither |
| $\mathbf{NA}_{1.4}$ | Wooden guitar | 3 or 4 String Musical Instrument | 13 String Koto |
| $\mathbf{NA}_{1.5}$ | Hawaiian Guitar | 3-4 Stringed Elliptical Instrument | 13-String Instrument |
| $\mathbf{NA}_{1.6}$ | 6 Stringed Instrument no Jack | Fretted Stringed Instrument | 13 Stringed Instrument |
| $\mathbf{NA}_{1.7}$ | Classic Guitar | Elliptical Stringed Instrument | Rectangular Stringed Instrument |
| $\mathbf{NA}_{1.8}$ | Non Powered Guitars | Short-Stringed Music Instruments | Japanese Stringed Instrument |

the names of some categories, or used wrong or more generic labels, or in one case used a more specific label.

### 7.2.2 Annotation cost

As from Section 6, `vTelos` requires that each image is recursively compared with the node one level down in the lexico-semantic hierarchy till it meets a differentia which it does not satisfy. So the question arises of what is the cost of the multiple checks. The preliminary observation is that, while containing an excess of 100K concepts, WordNet is a broad and flat hierarchy whose paths have a depth in the range of around 7 to 15 nodes (with different depths for different versions) and where the nodes higher in the hierarchy are very abstract and have no visual content. A 4-node deep hierarchy, like the one in Figure **??**, is therefore a good proxy for evaluating the annotation cost. We have therefore compared the time consumed by two non-expert annotators in the generation of two GT3 and GT4-like datasets restricted to contain one hundred images. The results show that the average annotation time per image is 5.565 seconds for GT3 and 3.473 seconds for GT4, that is, a time increase for `vTelos` of around 60% which one would guess it would become a little more than double with 8-node level deep hierarchies. The time increase correlates to the images classified in intermediate nodes. The fact that it is a low increase seems to

correlate to the fact that each single choice is simpler given that: (i) the same image is seen more than once, (ii) the choice is restricted within a precise set of labels, (iii) which have similar meanings, because selected following the pattern induced by the tree structure, thus ultimately improving the automation of the process.

### 7.2.3 Machine Learning Accuracy

We conducted experiments using eight state-of-the-art machine learning (ML) methods on two datasets, referred to as "ImageNet" and "vTelos-img". It is important to emphasize that the images in these two datasets are identical; the distinction lies in their annotations. Specifically, we collected images from ImageNet, discarded the ImageNet labels, and then re-annotated them based on the *vTelos* methodology. This process involved leveraging the Classificationist to define precise glosses and visual properties and organizing them into a well-structured hierarchy. Subsequently, the Classifier annotated the images following the hierarchy, resulting in the newly constructed vTelos-img dataset. The vTelos-img dataset consists of 55,086 images spanning 44 categories across three domains: "Birds", "Vehicles", and "Musical Instruments". Unlike the original ImageNet dataset, which retains its original, vTelos-img features refined annotations enriched with visual property-based labels.

For both datasets, we allocated 80% of the images as the training set and the remaining 20% as the test set. The all above settings aim to ensure a fair comparison of the ML methods' performance across the two datasets. By training the same models under identical conditions—except for the dataset labels—we aim to isolate the impact of improved annotation quality on the models' performance.

All the experiments have been implemented with PyTorch with identical settings. During the training, we performed the same data augmentation (horizontal flipping and random scaling with multiple times) on each dataset and

Table 7.4: Classification results for ImageNet and `vTelos-img`.

| Methods | Accuracy | | Improvement |
|---|---|---|---|
| | ImageNet | vTelos-img | |
| AlexNet [85] | 0.543 | 0.596 | 9.76% |
| ZFNet [161] | 0.612 | 0.657 | 7.35% |
| VGG [132] | 0.655 | 0.743 | 13.44% |
| GoogleNet [142] | 0.734 | 0.835 | 13.76% |
| ResNet [59] | 0.593 | 0.732 | 23.44% |
| DenseNet [70] | 0.724 | 0.793 | 9.53% |
| RAN [151] | 0.713 | 0.784 | 9.96% |
| SENets [68] | 0.728 | 0.811 | 11.40% |

randomly sampled $224 \times 224$ crops from augmented images. All experiments have been optimized using Adaptive Moment Estimation with learning rate initialized to $0.0002$, momentum initialized to $0.9$, and weight decay to $10^{-8}$. All models have been pre-trained on ImageNet with ResNet101 [59]. The results are reported in Table 7.4. It can be noticed that all models trained on vTelos-img get significant improvement in accuracy, especially ResNet which achieves up to 23.44% improvement.

Finally, if we look at the heat maps in Figure 7.1, we can see how the models pay attention to the visual regions described by the linguistic differentia, see for instance the keyboard in the case of the label *Keyboard Instrument*.

### 7.2.4 Ablation study

We conducted two ablation studies, the first focusing on Choice C1, the second of Choice C2 for all the eight systems tested in the previous experiment. The results are reported in Table 7.5 where the first and the last row replicate the results in Table 7.4. The key observation is that, in all cases, the improvement in performance due to Choice C2 with respect to that due to Choice C1 is

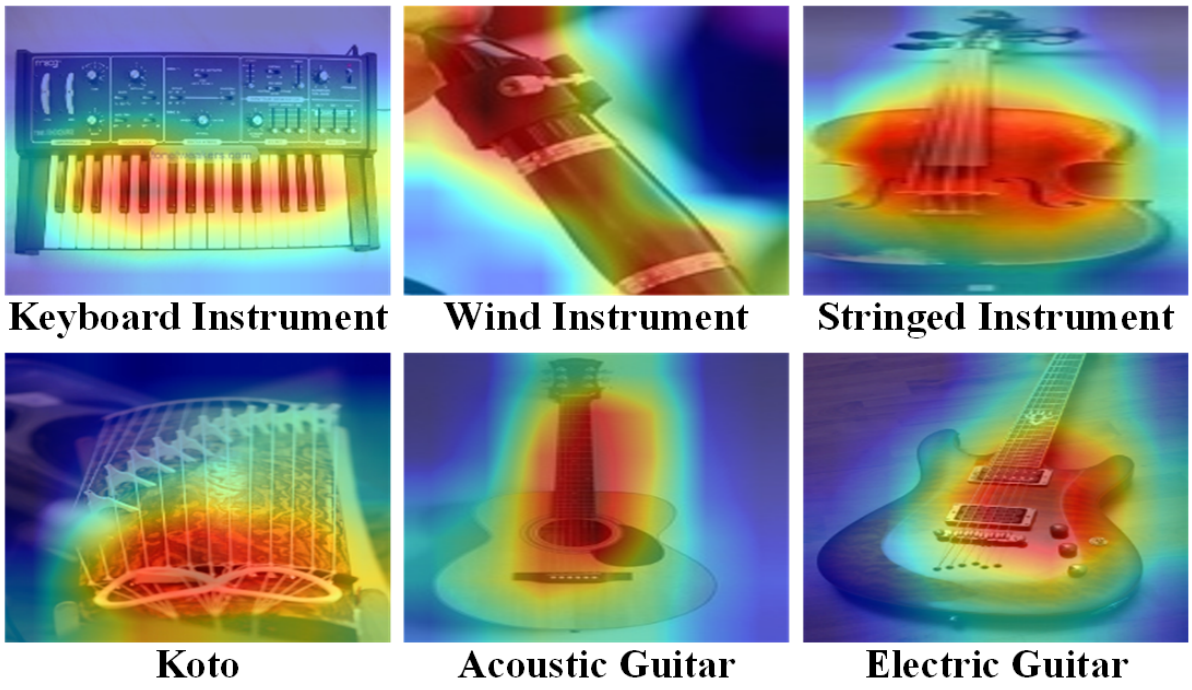


Figure 7.1: Attention heat maps.

substantial (e.g., 5 times with AlexNet, 2.9 times with ZFNet, 13.5 times with RAN, 8 times with ResNet) thus providing further confirmation of the intuition suggested by the heat maps in Figure 7.1. Taking into account the relevance of Choice C1 which transforms Multi-Object Images into Single-Object images, we take this as an important indicator of the fact that *labeling images using differentia labels is the way to go.*

## 7.3 Summary

In this chapter, we evaluated the effectiveness and applicability of the *vTelos* methodology through comprehensive experiments on the newly constructed vTelos-img dataset. We performed evaluations across multiple dimensions, including annotation quality, efficiency, practical application in machine learning tasks, and ablation studies. The inter-annotator agreement analysis demonstrated the reliability and consistency of the *vTelos* annotation process, achiev-

Table 7.5: Ablation study. "Acc." is the accuracy of object recognition, "Impro.(%)" is the improvement in performance of the dataset over ImageNet.

| Methods | | ImageNet | vTelos - Differentia | Impro.(%) | vTelos - Detection | Impro.(%) | vTelos-img | Impro.(%) |
|---|---|---|---|---|---|---|---|---|
| Accuracy | **AlexNet** | 0.543 | 0.55 | 1.47% | 0.58 | 7.37% | 0.596 | 9.76% |
| | **ZFNet** | 0.612 | 0.62 | 1.80% | 0.64 | 5.23% | 0.657 | 7.35% |
| | **VGG** | 0.655 | 0.67 | 2.55% | 0.73 | 11.82% | 0.743 | 13.44% |
| | **GoogleNet** | 0.734 | 0.75 | 1.55% | 0.83 | 12.94% | 0.835 | 13.76% |
| | **ResNet** | 0.593 | 0.61 | 2.86% | 0.73 | 22.93% | 0.732 | 23.44% |
| | **DenseNet** | 0.724 | 0.73 | 1.24% | 0.78 | 8.01% | 0.793 | 9.53% |
| | **RAN** | 0.713 | 0.72 | 0.74% | 0.77 | 8.04% | 0.784 | 9.96% |
| | **SENets** | 0.728 | 0.75 | 2.54% | 0.81 | 10.87% | 0.811 | 11.40% |

ing near-perfect agreement. Cost analysis revealed the efficiency of the methodology, highlighting its suitability for large-scale dataset construction. Experiments with state-of-the-art ML methods confirmed that models trained on vTelos-img significantly outperformed those trained on the original ImageNet subset, validating the enhanced quality and semantic consistency of the dataset. Ablation studies further underscored the importance of key components of the *vTelos* process, such as object localization and visual classification, in achieving superior dataset quality.

Overall, this chapter highlights the robustness and practicality of the *vTelos* methodology for constructing high-quality datasets. The results demonstrate its potential to bridge the semantic gap and enhance the performance of machine learning models, providing a strong foundation for future advancements in dataset design and optimization.

# Chapter 8

# Conclusion

## 8.1 Main Conclusion

This thesis addresses the systemic flaws and the semantic gap problems inherent in current image dataset construction processes by proposing a novel image dataset construction methodology called `vTelos`. By integrating techniques from natural language processing, knowledge representation, and computer vision, we tackle the challenges posed by the many-to-many mapping between visual information and semantic labels from both theoretical and practical perspectives. The primary contributions of this thesis include: First, existing datasets suffer from issues such as mislabeled cases, semantic ambiguities, and cultural biases, closely tied to the semantic gap problems. To tackle this, we conducted a comprehensive analysis of these issues and proposed a systematic annotation methodology `vTelos`. Second, the ambiguity in current annotation processes highlights the need for a structured approach. Our methodology introduces a hierarchical framework based on visual properties, supported by standardized annotation guidelines to reduce inconsistencies and improve dataset quality. Third, effective evaluation schemes are necessary to validate dataset construction methods. We implemented a multidimensional evaluation framework to demonstrate the practical effectiveness of `vTelos` in improving dataset reliability and applicability. These contributions establish a robust foun-

dation for resolving the SGP and advancing dataset construction methodologies in computer vision.

In detail, Chapter 2 provides a comprehensive review of existing methods for constructing image datasets, with a focus on analyzing the key challenges faced during the development of current benchmark datasets. This chapter also introduces several categorization methods and examines their impact on dataset quality, laying the foundation for the proposed methodology.

Chapter 3 delves into the annotation mistakes prevalent in current image datasets, introducing the concept and definition of the Semantic Gap Problem (SGP) and discussing its manifestations in image datasets. We specifically discuss and classify the situations of annotation mistakes, providing support for solving such annotation mistakes.

In Chapter 4, we present a systematic approach to addressing the semantic gap problems, detailing the design of the proposed `vTelos` methodology. We introduce the division of roles and collaborative mechanisms between the Classificationist and Classifier and discuss the four key annotation design choices that guide the construction process. It demonstrates how explicit classification rules and semantic definitions can effectively address the critical challenges of dataset annotation.

Chapter 5 focuses on the core role of the Classificationist, providing an in-depth analysis of classification canons, semantic glosses, and the design of hierarchy structures based on visual properties. Through specific examples, this chapter illustrates how the Classificationist applies canons to construct scientifically robust classification systems, ensuring semantic consistency and integrity for image datasets.

Chapter 6 explores the pivotal role of the Classifier in the image annotation process and introduces the application of crowdsourcing techniques for large-scale dataset construction. Through a series of experiments, this chapter demonstrates how Classifiers use the hierarchies defined by the Classificationist

to perform high-quality annotation tasks and how iterative refinement improves annotation efficiency and consistency.

In Chapter 7, we present a comprehensive evaluation of the vTelos methodology through experiments. The analysis presents four aspects: quality validation, annotation efficiency, the performance improvement of machine learning models on the dataset, and ablation studies of the core components of the methodology. These evaluations validate the practicality and scalability of the vTelos approach.

Overall, these chapters form a comprehensive and systematic research framework, showcasing the significant advantages of the vTelos methodology in addressing the semantic gap problems and constructing high-quality image datasets from both theoretical and practical perspectives.

## 8.2 Prospects for Future Work

In this concluding section of our final chapter, we wish to deliberate on potential avenues for future research. Future research will focus on the further development and application of the vTelos methodology, including the construction of larger-scale image datasets, the application of the methodology to more complex computer vision tasks, the expansion to other types of multimedia datasets, and the integration of advanced intelligent technologies. First, we plan to use the vTelos methodology to construct larger-scale image datasets that encompass a broader range of visual objects and semantic categories. This will validate the scalability and generalizability of the methodology in large-scale dataset construction. Such efforts will enhance the diversity and representativeness of the datasets, providing higher-quality foundational data to support the training of machine learning models.

Second, we aim to apply datasets constructed using the vTelos methodology to more complex computer vision tasks, such as zero-shot image clas-

sification and fine-grained image classification. By deploying these datasets in practical applications, we can validate their potential in improving model generalization and adaptability. Additionally, this will further demonstrate the advantages of the `vTelos` methodology in addressing the Semantic Gap Problem (SGP) and improving task performance, showcasing its broad applicability in the field of computer vision.

Another promising direction is to extend the `vTelos` methodology to the construction of other types of multimedia datasets, including video datasets and multimodal datasets. For video datasets in particular, the methodology must address unique challenges posed by temporal dimensions and dynamic object variations, such as capturing and annotating temporal sequence features and representing dynamic semantics in hierarchical structures. These research directions will further expand the applicability of the methodology and enhance its ability to support complex multimodal tasks.

Future research directions also include the integration of advanced intelligent technologies, such as large language models (LLMs) and vision-language models (VLMs), to optimize the annotation process and enhance the semantic representation capabilities of datasets. These technologies are expected to significantly improve annotation efficiency, reduce human subjective bias, and enrich the semantic hierarchical structure of datasets, enabling them to better reflect the complex semantic relationships found in real-world scenarios.

# Bibliography


[1] Yehya Abouelnaga, Ola S Ali, Hager Rady, and Mohamed Moustafa. Cifar-10: Knn-based ensemble of classifiers. In *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1192–1195. IEEE, 2016.

[2] Ron Artstein. Inter-annotator agreement. In *Handbook of linguistic annotation*, pages 297–313. Springer, 2017.

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[4] A. Bendale and T. E. Boult. Towards open world recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 1893–1902, 2015.

[5] Alexander C Berg, Tamara L Berg, Hal Daume, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, et al. Understanding and predicting importance in images. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3562–3569. IEEE, 2012.

[6] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

[7] Shahzad Sarwar Bhatti, Yiding Chang, Xiaofeng Gao, and Guihai Chen. Affinitive diversity-aware task allocation in spatial crowdsourcing. *2020 IEEE International Conference on Web Services (ICWS)*, pages 27–36, 2020.

[8] Carlo Bianchini and Stefano Bargioni. Automated classification using linked open data. a case study on faceted classification and wikidata. *Cataloging & Classification Quarterly*, 59:835 – 852, 2021.

[9] Christian Bors, Theresia Gschwandtner, Simone Kriglstein, Silvia Miksch, and Margit Pohl. Visual interactive creation, customization, and analysis of data quality metrics. *Journal of Data and Information Quality (JDIQ)*, 10(1):1–26, 2018.

[10] Paolo Bouquet and Fausto Giunchiglia. Reasoning about theory adequacy. a new solution to the qualification problem. *Fundamenta Informaticae*, 23(2, 3, 4):247–262, 1995.

[11] Ronald J Brachman. *Knowledge Representation and Reasoning*. Morgan Kaufman/Elsevier, 2004.

[12] Cynthia Brame. Active learning. *Vanderbilt University Center for Teaching*, 2016.

[13] Alec Burmania, Srinivas Parthasarathy, and Carlos Busso. Increasing the reliability of crowdsourcing evaluations using online quality assessment. *IEEE Transactions on Affective Computing*, 7(4):374–388, 2015.

[14] Luca Cagliero, Paolo Garza, Giuseppe Attanasio, and Elena Baralis. Training ensembles of faceted classification models for quantitative stock trading. *Computing*, 102:1213 – 1225, 2020.

[15] Lijuan Cai and Thomas Hofmann. Hierarchical document categorization with support vector machines. In *Proceedings of the thirteenth ACM*

*international conference on Information and knowledge management*, pages 78–87, 2004.

[16] Wenshi Chen, Bowen Zhang, and Mingyu Lu. Uncertainty quantification for multilabel text classification. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6):e1384, 2020.

[17] Zhihao Chen, Bin Hu, Chuang Niu, Tao Chen, Yuxin Li, Hongming Shan, and Ge Wang. Iqagpt: Image quality assessment with vision-language and chatgpt models. *arXiv preprint arXiv:2312.15663*, 2023.

[18] Yang Chi, Fausto Giunchiglia, Daqian Shi, Xiaolei Diao, Chuntao Li, and Hao Xu. Zinet: Linking chinese characters spanning three thousand years. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3061–3070, 2022.

[19] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46, 1960.

[20] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[21] Jesse Cummings, Elías Snorrason, and Jonas Mueller. Detecting dataset drift and non-iid sampling via k-nearest neighbors. In *ICML Workshop on Data-centric Machine Learning Research*, 2023.

[22] F. Daniel, P. Kucherbaev, C. Cappiello, B. Benatallah, and M. Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys*, 51(1):1–40, 2018.

[23] Roberto De Virgilio, Fausto Giunchiglia, and Letizia Tanca. *Semantic web information management: a model-based perspective*. Springer Science & Business Media, 2010.

[24] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *Workshops of Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, pages 52–59, 2019.

[25] Ofer Dekel, Joseph Keshet, and Yoram Singer. Large margin hierarchical classification. In *Proceedings of the twenty-first international conference on Machine learning*, page 27, 2004.

[26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. computer vision and pattern recognition*, pages 248–255, 2009.

[27] Xiaolei Diao and Fausto Giunchiglia. Building a visual semantics aware object hierarchy. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence IJCAIECAI 2022*, 2022.

[28] Xiaolei Diao and Fausto Giunchiglia. Crowdsourcing image annotation by visual properties. *arXiv preprint arXiv:2501.15663*, 2025.

[29] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 256–263, 2000.

[30] GEORGE DUNBAR. Towards a cognitive analysis of polysemy, ambiguity, and vagueness. *Cognitive Linguistics*, 12(1):1–14, 2001.

[31] L. Erculiani, A. Bontempelli, A. Passerini, and F. Giunchiglia. Egocentric hierarchical visual semantics. In *HHAI Conf.*, 2023.

[32] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[33] Jianping Fan, Yuli Gao, and Hangzai Luo. Hierarchical classification for automatic image annotation. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 111–118, 2007.

[34] Jianping Fan, Yuli Gao, Hangzai Luo, and Ramesh Jain. Mining multilevel image semantics via hierarchical classification. *IEEE Transactions on Multimedia*, 10(2):167–187, 2008.

[35] Jianping Fan, Ji Zhang, Kuizhi Mei, Jinye Peng, and Ling Gao. Cost-sensitive learning of hierarchical tree classifiers for large-scale image classification and novel category detection. *Pattern Recognition*, 48(5):1673–1687, 2015.

[36] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382, 1971.

[37] Niki Maria Foteinopoulou, Christos Tzelepis, and Ioannis Patras. Estimating continuous affect with label uncertainty. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2021.

[38] Julia R Fox, Byungho Park, and Annie Lang. When available resources become negative resources: The effects of cognitive overload on memory sensitivity and criterion bias. *Communication Research*, 34(3):277–296, 2007.

[39] Abed Alhakim Freihat. An organizational approach to the polysemy problem in wordnet. 2014.

[40] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

[41] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016.

[42] F. Giunchiglia, M. Bagchi, and X. Diao. Aligning visual and lexical semantics. In *Conf. on Information (iConference)*, 2023.

[43] F. Giunchiglia, K. Batsuren, and G. Bella. Understanding and exploiting language diversity. In *International Joint Conference on Artificial Intelligence*, pages 4009–4017, 2017.

[44] F. Giunchiglia, L. Erculiani, and A. Passerini. Towards visual semantics. *Springer Nature Computer Science (SNCS)*, 2(6), 2021.

[45] F. Giunchiglia, S. Kleanthous, J. Otterbacher, and T. Draws. Transparency paths-documenting the diversity of user perceptions. In *Adjunct Proc. of the 29th ACM Conference UMAP*, pages 415–420, 2021.

[46] Fausto Giunchiglia, Mayukh Bagchi, and Xiaolei Diao. Visual ground truth construction as faceted classification. *ArXiv*, abs/2202.08512, 2022.

[47] Fausto Giunchiglia, Mayukh Bagchi, and Xiaolei Diao. Aligning visual and lexical semantics. In *International Conference on Information*, pages 294–302, 2023.

[48] Fausto Giunchiglia, Mayukh Bagchi, and Xiaolei Diao. A semantics-driven methodology for high-quality image annotation. In *European Conference on Artificial Intelligence (ECAI)*, 2023.

[49] Fausto Giunchiglia, Xiaolei Diao, and Mayukh Bagchi. Incremental image labeling via iterative refinement. In *IWCIM@ICASSP*, 2023.

[50] Fausto Giunchiglia and Mattia Fumagalli. Concepts as (recognition) abilities. In *FOIS*, pages 153–166, 2016.

[51] Fausto Giunchiglia, Ilya Zaihrayeu, and Uladzimir Kharkevich. Formalizing the get-specific document classification algorithm. In *TPDL*, pages 26–37. Springer, 2007.

[52] Claudio Gnoli. Faceted classifications as linked data: A logical analysis. *KNOWLEDGE ORGANIZATION*, 2021.

[53] Hui Wen Goh and Jonas Mueller. Activelab: Active learning with relabeling by multiple annotators. In *ICLR Workshop on Trustworthy ML*, 2023.

[54] Hui Wen Goh, Ulyana Tkachenko, Jonas Mueller, and Cleanlab Cleanlab Cleanlab. Utilizing supervised models to infer consensus labels and their quality from data with multiple annotators. *CoRR*, 2022.

[55] Robert L Goldstone, Alan Kersten, and Paulo F Carvalho. Concepts and categorization. *Comprehensive handbook of psychology*, 4:599–621, 2003.

[56] Nicola Guarino, Daniel Oberle, and Steffen Staab. What is an ontology? In *Handbook on ontologies*, pages 1–17. Springer, 2009.

[57] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE intelligent systems*, 24(2):8–12, 2009.

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015.

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

[60] Marti Hearst, Ame Elliott, Jennifer English, Rashmi Sinha, Kirsten Swearingen, and Ka-Ping Yee. Finding the flow in web site search. *Communications of the ACM*, 45(9):42–49, 2002.

[61] Marti A. Hearst. Clustering versus faceted categories for information exploration. *Communications of the ACM*, 49:59 – 61, 2006.

[62] Alessa Hering, Lasse Hansen, Tony CW Mok, Albert CS Chung, Hanna Siebert, Stephanie Häger, Annkristin Lange, Sven Kuckertz, Stefan Heldmann, Wei Shao, et al. Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, 42(3):697–712, 2022.

[63] Chien-Ju Ho, Tao-Hsuan Chang, Jong-Chuan Lee, Jane Yung-jen Hsu, and Kuan-Ta Chen. Kisskissban: a competitive human computation game for image annotation. In *Proceedings of the acm sigkdd workshop on human computation*, pages 11–14, 2009.

[64] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*, pages 419–429, 2015.

[65] Chien-Ju Ho and Jennifer Wortman Vaughan. Online task assignment in crowdsourcing markets. In *AAAI Conference on Artificial Intelligence*, 2012.

[66] Saihui Hou, Yushan Feng, and Zilei Wang. Vegfru: A domain-specific dataset for fine-grained visual categorization. In *Proceedings of the IEEE international conference on computer vision*, pages 541–549, 2017.

[67] Jeff Howe et al. The rise of crowdsourcing. *Wired magazine*, 14(6):176–183, 2006.

[68] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, pages 7132–7141, 2018.

[69] Mengqiu Hu, Yang Yang, Fumin Shen, Luming Zhang, Heng Tao Shen, and Xuelong Li. Robust web image annotation via exploring multi-facet and structural knowledge. *IEEE Transactions on Image Processing*, 26:4871–4884, 2017.

[70] G. Huang, Z.ang Liu, L. VanDer Maaten, and KQ. Weinberger. Densely connected convolutional networks. In *Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 2017.

[71] John Hughes. Krippendorffsalpha: An r package for measuring agreement using krippendorff's alpha coefficient. *The R Journal*, 1(1), 2021. Also: arXiv preprint arXiv:2103.12170.

[72] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67, 2010.

[73] Johannes Jakubik, Michael Vössing, Niklas Kühl, Jannis Walk, and Gerhard Satzger. Data-centric artificial intelligence. *Business & Information Systems Engineering*, pages 1–9, 2024.

[74] Kaige Jiang, Yingjie Wang, Haipeng Wang, Zhaowei Liu, Qilong Han, Ao Zhou, Chaocan Xiang, and Zhipeng Cai. A reinforcement learning-based incentive mechanism for task allocation under spatiotemporal crowdsensing. *IEEE Transactions on Computational Social Systems*, 11:2179–2189, 2024.

[75] Nan-Jiang Jiang and Marie-Catherine de Marneffe. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374, 2022.

[76] Ashish Kapoor, Eric Horvitz, and Sumit Basu. Selective supervision: Guiding supervised learning with decision-theoretic active learning. In *IJCAI'07 Proceedings of the 20th international joint conference on Artifical intelligence*, pages 877–882, 2007.

[77] Prithvi N Kaula. Canons in analytico-synthetic classification. *KO KNOWLEDGE ORGANIZATION*, 7(3):118–125, 1980.

[78] Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456, 2008.

[79] Bernard Koch, Emily Denton, Alex Hanna, and Jacob G Foster. Reduced, reused and recycled: The life of a dataset in machine learning research. In *NIPS*, 2021.

[80] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2973–2980, 2012.

[81] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017.

[82] Ivan Krasin, Tom Duerig, Neil Alldrin, Andreas Veit, Sami Abu-El-Haija, Serge Belongie, David Cai, Zheyun Feng, Vittorio Ferrari, Victor Gomes, Abhinav Gupta, Dhyanesh Narayanan, Chen Sun, Gal Chechik, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2016.

[83] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[84] Klaus Krippendorff. Computing krippendorff's alpha-reliability. *Computing*, 1:1–25, 2011.

[85] A. Krizhevsky, I. Sutskever, and GE Hinton. Imagenet classification with deep convolutional neural networks. *Neurips*, 2012.

[86] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[87] Johnson Kuan and Jonas Mueller. Model-agnostic label quality scoring to detect real-world label errors. In *ICML DataPerf Workshop*, 2022.

[88] Ashnil Kumar, Shane Dyer, Jinman Kim, Changyang Li, Philip HW Leong, Michael Fulham, and Dagan Feng. Adapting content-based image retrieval techniques for the semantic annotation of medical images. *Computerized Medical Imaging and Graphics*, 49:37–45, 2016.

[89] Vedang Lad and Jonas Mueller. Estimating label quality and errors in semantic segmentation data via any model. In *ICML Workshop on Data-centric Machine Learning Research*, 2023.

[90] Edith Law and Luis von Ahn. Human computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2011.

[91] Matthew Lease. On quality control and machine learning in crowdsourcing. In *Workshops at the twenty-fifth AAAI conference on artificial intelligence*, 2011.

[92] Johann Li, Guangming Zhu, Cong Hua, Mingtao Feng, Basheer Bennamoun, Ping Li, Xiaoyuan Lu, Juan Song, Peiyi Shen, Xu Xu, et al. A systematic collection of medical image datasets for deep learning. *ACM Computing Surveys*, 56(5):1–51, 2023.

[93] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[94] Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 68–76, 2010.

[95] Baoyuan Liu, Fereshteh Sadeghi, Marshall Tappen, Ohad Shamir, and Ce Liu. Probabilistic label trees for efficient large scale image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 843–850, 2013.

[96] Shixia Liu, Changjian Chen, Yafeng Lu, Fangxin Ouyang, and Bin Wang. An interactive method to improve crowdsourced annotations. *IEEE Transactions on Visualization and Computer Graphics*, 25:235–245, 2019.

[97] Sidong Liu, Siqi Liu, Sonia Pujol, Ron Kikinis, Dagan Feng, and Weidong Cai. Propagation graph fusion for multi-modal medical content-based retrieval. In *2014 13th International Conference on Control Automation Robotics & Vision (ICARCV)*, pages 849–854. IEEE, 2014.

[98] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019.

[99] Chang Lu, Yanyun Qu, Cuiting Shi, Jianping Fan, Yang Wu, and Hanzi Wang. Hierarchical learning for large-scale image classification via cnn and maximum confidence path. In *Pacific Rim Conference on Multimedia*, pages 236–245. Springer, 2015.

[100] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[101] Winter Mason and Duncan J Watts. Financial incentives and the" performance of crowds". In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 77–85, 2009.

[102] Carolyn B Mervis, Eleanor Rosch, et al. Categorization of natural objects. *Annual review of psychology*, 32(1):89–115, 1981.

[103] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[104] Ruth Garrett Millikan. *On clear and confused ideas: An essay about substance concepts*. Cambridge University Press, 2000.

[105] Ruth Garrett Millikan. Neuroscience and teleosemantics. *Synthese*, pages 1–9, 2020.

[106] Henning Müller and Devrim Unay. Retrieval from and understanding of large-scale multi-modal medical datasets: a review. *IEEE transactions on multimedia*, 19(9):2093–2104, 2017.

[107] Joseph Nassar, Viveca Pavon-Harr, Marc Bosch, and Ian McCulloh. Assessing data quality of annotations with krippendorff alpha for applications in computer vision. *arXiv preprint arXiv:1912.10107*, 2019.

[108] Geng Niu, Pan Yang, Yi Zheng, Ximing Cai, and Huapeng Qin. Automatic quality control of crowdsourced rainfall data with multiple noises: A machine learning approach. *Water Resources Research*, 57(11):e2020WR029121, 2021.

[109] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

[110] Stefanie Nowak and Stefan Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *IEEE-MIPR*, pages 557–566, 2010.

[111] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

[112] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9617–9626, 2019.

[113] Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Xiao-Ping Zhang, Dong Wu, and Tao Mei. Boosting video representation learning with multi-faceted

integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14030–14039, June 2021.

[114] Habibur Rahman, Senjuti Basu Roy, Saravanan Thirumuruganathan, Sihem Amer-Yahia, and Gautam Das. Task assignment optimization in collaborative crowdsourcing. *2015 IEEE International Conference on Data Mining*, pages 949–954, 2015.

[115] Harish Ramaswamy, Ambuj Tewari, and Shivani Agarwal. Convex calibrated surrogates for hierarchical classification. In *International Conference on Machine Learning*, pages 1852–1860. PMLR, 2015.

[116] S R Ranganathan. *Prolegomena to Library Classification*. Asia Publishing House (Bombay and New York), 1967.

[117] S R Ranganathan. *Philosophy of library classification*. Sarada Ranganathan Endowment for Library Science (Bangalore, India), 1989.

[118] Shiyali Ramamrita Ranganathan. Self-perpetuating scheme of classification. *Journal of Documentation*, 1949.

[119] Shiyali Ramamrita Ranganathan. Prolegomena to library classification. *Asia PublishingHouse*, 1967.

[120] Shiyali Ramamrita Ranganathan and Arashanapalai Neelameghan. Classified catalogue code: with additional rules for dictionary catalogue code. *(No Title)*, 1964.

[121] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[122] Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439, 1976.

[123] O. Russakovsky, J. Deng, H. Su, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[124] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[125] Thomas Serre. Models of visual categorization. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(3):197–213, 2016.

[126] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

[127] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622, 2008.

[128] Daqian Shi, Ting Wang, Hao Xing, and Hao Xu. A learning path recommendation model based on a multidimensional knowledge graph framework for e-learning. *Knowledge-Based Systems*, 195:105618, 2020.

[129] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016.

[130] Carlos N. Silla and Alex A. Freitas. A survey of hierarchical classification across different application domains. *Data mining and knowledge discovery*, 22:31–72, 2011.

[131] Carlos N Silla and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data mining and knowledge discovery*, 22:31–72, 2011.

[132] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[133] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[134] Prerna Singh. Systematic review of data-centric approaches in artificial intelligence and machine learning. *Data Science and Management*, 2023.

[135] A. WM. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (12):1349–1380, 2000.

[136] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263, 2008.

[137] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consis-

tency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[138] Pia Sommerauer, Antske Fokkens, and Piek Vossen. Would you describe a leopard as yellow? evaluating crowd-annotations with justified and informative disagreement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4798–4809, 2020.

[139] Alexander Sorokin and David Forsyth. Utility data annotation with amazon mechanical turk. In *2008 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 1–8. IEEE, 2008.

[140] Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing annotations for visual object detection. In *Workshops at the twenty-sixth AAAI conference on artificial intelligence*, 2012.

[141] John Sweller. Cognitive load theory. In *Psychology of learning and motivation*, volume 55, pages 37–76. Elsevier, 2011.

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

[143] Ulyana Tkachenko, Aditya Thyagarajan, and Jonas Mueller. Objectlab: Automated diagnosis of mislabeled images in object detection data. In *ICML Workshop on Data-centric Machine Learning Research*, 2023.

[144] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, pages 1521–1528. IEEE, 2011.

[145] D. Tsipras, S. Santurkar, L. Engstrom, A. Ilyas, and A. Madry. From imagenet to image classification: Contextualizing progress on benchmarks.

In *International Conference on Machine Learning (ICML 2020)*, pages 9625–9635. PMLR, 2020.

[146] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning*, pages 9625–9635. PMLR, 2020.

[147] Douglas Tudhope, Traugott Koch, and Rachel Heery. Terminology services and technology: Jisc state of the art review. 2006.

[148] Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021.

[149] Epaphrodite Uwimana and Miguel E. Ruiz. Automatic classification of medical images for content based image retrieval systems (cbir). *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 52:788 – 792, 2008.

[150] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[151] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pages 3156–3164, 2017.

[152] Wei-Chen Wang and Jonas Mueller. Detecting label errors in token classification data. *NeurIPS 2022 Workshop on Interactive Learning for Natural Language Processing (InterNLP)*, 2022.

[153] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and R Summers. Hospital-scale chest x-ray database and bench-

marks on weakly-supervised classification and localization of common thorax diseases. In *IEEE CVPR*, volume 7, page 46. sn, 2017.

[154] Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. The multidimensional wisdom of crowds. *Advances in neural information processing systems*, 23, 2010.

[155] Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal*, 32(4):791–813, 2023.

[156] K. Yang, K. Qinami, L. Fei-Fei, J. Deng, and O. Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *FACT Conf.*, pages 547–558, 2020.

[157] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8934–8954, 2022.

[158] Dunhui Yu, Yi Wang, and Zhuang Zhou. Software crowdsourcing task allocation algorithm based on dynamic utility. *IEEE Access*, 7:33094–33106, 2019.

[159] S. Yun, Sj. Oh, B. Heo, D. Han, J. Choe, and S. Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *Conference on Computer Vision and Pattern Recognition (CVPR 2021)*, pages 2340–2350, 2021.

[160] Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord. Explainability of deep vision-based autonomous driving systems: Review and challenges. *International Journal of Computer Vision*, 130(10):2425–2452, 2022.

[161] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014.

[162] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*, 2023.

[163] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020.

[164] Xiang Zhang, Guoliang Xue, Ruozhou Yu, Dejun Yang, and Jian Tang. Truthful incentive mechanisms for crowdsourcing. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 2830–2838. IEEE, 2015.

[165] Yan Zhao, Kai Zheng, Yue Cui, Han Su, Feida Zhu, and Xiaofang Zhou. Predictive task assignment in spatial crowdsourcing: A data-driven approach. *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 13–24, 2020.

[166] Chen Zhou, Mohit Prabhushankar, and Ghassan AlRegib. On the ramifications of human label uncertainty. *arXiv preprint arXiv:2211.05871*, 2022.

[167] Hang Zhou, Jonas Mueller, Mayank Kumar, Jane-Ling Wang, and Jing Lei. Detecting errors in numerical data via any regression model. In *ICML Workshop on Data-centric Machine Learning Research*, 2023.

[168] Jingbo Zhu, Huizhen Wang, Eduard Hovy, and Matthew Ma. Confidence-based stopping criteria for active learning for data annotation. *ACM*

*Transactions on Speech and Language Processing (TSLP)*, 6(3):1–24, 2010.

# Appendix A

# vTelos-img

We constructed the vTelos-img dataset based on the proposed image dataset construction methodology, `vTelos`. The vTelos-img dataset comprises 55,086 images, covering 44 categories across three domains: “Birds”, “Vehicles”, and “Musical Instruments”. Acting as the Classificationists, we designed comprehensive glosses and visual properties, including visual genus and visual differentia, for all 44 categories in vTelos-img and organized them into a hierarchical structure. All definitions can be found in Table A.1.

Table A.1: Categories, glosses, and visual properties we defined in vTelos-img, where visual properties consist of visual genus and visual differnetia.

| Words | glosses | visual genus | visual differentia |
|---|---|---|---|
| hen | adult female bird — adult female chicken with non-arched tail feathers | chicken | non-arched tail feathers |
| brambling | Eurasian finch — with orange breast and white belly | finch | with orange breast and white belly |
| goldfinch | small European finch having a crimson face and yellow-and-black wings | finch | crimson face and yellow-and-black wings |
| house_finch | small finch originally of the western United States and Mexico — having white wing bar | finch | white wing bar |
| junco | small North American finch seen chiefly in winter — slate coloured having white belly | finch | slate coloured having white belly |
| indigo_bunting | small deep blue North American bunting — with a short, conical beak | bunting | with a short, conical beak |
| robin | large American thrush having a rust-red breast and abdomen | thrush | rust-red breast and abdomen |
| bulbul | nightingale spoken of in Persian poetry — with an erect crest | nightingale | erect crest |
| magpie | long-tailed black-and-white corvine that utters a raucous chattering call | corvine_bird | long-tailed black-and-white |
| chickadee | any of various small grey-and-black titmouse songbirds of North America — with white cheek | titmouse | grey-and-black, white cheek |
| water_ouzel | small diving stocky dark grey oscine bird having long, unwebbed feet | oscine | stocky dark grey, and long unwebbed feet |
| black_swan | large Australian swan having black plumage and a red bill | swan | black plumage and a red bill |
| white_stork | the common stork of Europe; white with black wing feathers and a red bill | stork | white with black wing feathers |
| black_stork | Old World stork that is glossy black above and white below | stork | glossy black above and white below |
| spoonbill | wading birds having a long flat bill with a tip like a spoon | wading_bird | long flat bill with a tip like a spoon |
| flamingo | large pink to scarlet web-footed wading bird with down-bent bill; inhabits brackish lakes | wading_bird | pink to scarlet |
| limpkin | wading bird of Florida, Cuba and Jamaica having a distinctive wailing call, with a drab plumage—dark brown with an olive luster above | wading_bird | drab plumage—dark brown with an olive luster above |
| bustard | large heavy-bodied chiefly terrestrial wading bird capable of powerful swift flight, with a cryptic plumage | wading_bird | cryptic plumage |
| little_blue_heron | small bluish-grey heron of the western hemisphere, with a rich purple-maroon head and neck | heron | rich purple-maroon head and neck |
| American_egret | a common egret of the genus Egretta found in America, with a long S -curved neck | egret | long S -curved neck |
| bittern | relatively small compact tawny-brown heron with nocturnal habits and a booming cry, having a distinctive streaked neck | heron | a distinctive streaked neck |
| American_coot | a coot found in North America — with a white frontal shield | coot | a white frontal shield |
| ruddy_turnstone | common Arctic turnstone that winters in South America and Australia, with a harlequin-like plumage pattern of black and white | turnstone | a harlequin-like plumage pattern of black and white |
| red-backed_sandpiper | small common sandpiper that breeds in northern or Arctic regions and winters in southern United States or Mediterranean regions, with a droopy bill | sandpiper | droopy bill |
| redshank | a common Old World sandpiper with long red legs | sandpiper | long red legs |
| dowitcher | a snipe, with a long bill and dark body scalloping | snipe | dark body scalloping |
| oystercatcher | black-and-white shorebird with stout legs and massive long orange or red bill | shorebird | long orange or red bill |
| pelican | large long-winged warm-water pelicanifrom seabird having a large bill with a distensible pouch for fish | pelecaniform_seabird | a large bill with a distensible pouch for fish |
| king_penguin | large penguin on islands bordering the Antarctic Circle, with vivid orange, tear-shaped patches on each side of the head | penguin | vivid orange, and tear-shaped patches on each side of the head |
| European_gallinule | purple gallinule of southern Europe, with a red frontal shield | purple_gallinule | a red frontal shield |
| bicycle-built-for-two | a bicycle with two sets of pedals and two seats | bicycle | two sets of pedals and two seats |
| mountain_bike | a bicycle with a sturdy frame and fat tires; originally designed for riding in mountainous country | bicycle | a sturdy frame and fat tires |
| freight_car | a railway car that carries freight; composed of covered wagons | railway car | covered wagons |
| grand_piano | a piano with the strings on a horizontal harp-shaped frame; usually supported by three legs | piano | horizontal harp-shaped frame |
| chime | a percussion instrument consisting of a set of tuned bells that are struck with a hammer; used as an orchestral instrument | percussion_instrument | a set of tuned bells |
| drum | a musical percussion instrument; usually consists of a hollow cylinder with a membrane stretched across each end | percussion_instrument | hollow cylinder with a membrane stretched |
| gong | a percussion instrument consisting of a metal plate that is struck with a softheaded drumstick | percussion_instrument | a metal plate |
| maraca | a percussion instrument consisting of a hollow gourd containing pebbles or beans; often played in pairs | percussion_instrument | hollow gourd |
| marimba | a percussion instrument with wooden bars tuned to produce a chromatic scale and with resonators; played with small mallets | percussion_instrument | wooden bars, resonators |
| steel_drum | a concave percussion instrument made from the metal top of an oil drum; has an array of flattened areas that produce different tones when struck (of Caribbean origin) | percussion_instrument | metal top, and an array of flattened areas |
| koto | a string instrument has a rectangular wooden sounding board and usually 13 strings. | string_instrument | 13 strings |
| dulcimer | a string instrument with an elliptical body and three or four strings | string_instrument | an elliptical body, and three or four strings |
| acoustic _guitar | a guitar with no an input jack | guitar | with no an input jack |
| electric_guitar | a guitar with an input jack | guitar | with an input jack |