VisionKG: Unleashing the Power of Visual Datasets via Knowledge Graph

Jicheng Yuan[1], Anh Le-Tuan[1], Manh Nguyen-Duc[1], Trung-Kien Tran[2], Manfred Hauswirth[1,3], and Danh Le-Phuoc[1,3]

[1] Open Distributed Systems, Technical University of Berlin
{jicheng.yuan,anh.letuan,duc.manh.nguyen, manfred.hauswirth,danh.lephuoc}@tu-berlin.de
[2] Bosch Center for Artificial Intelligence, Renningen, Germany
{TrungKien.Tran}@de.bosch.com
[3] Fraunhofer Institute for Open Communication Systems, Berlin, Germany

Abstract. The availability of vast amounts of visual data with heterogeneous features is a key factor for developing, testing, and benchmarking of new computer vision (CV) algorithms and architectures. Most visual datasets are created and curated for specific tasks or with limited image data distribution for very specific situations, and there is no unified approach to manage and access them across diverse sources, tasks, and taxonomies. This not only creates unnecessary overheads when building robust visual recognition systems, but also introduces biases into learning systems and limits the capabilities of data-centric AI. To address these problems, we propose the **Vision K**nowledge **G**raph (**VisionKG**), a novel resource that interlinks, organizes and manages visual datasets via knowledge graphs and Semantic Web technologies. It can serve as a unified framework facilitating simple access and querying of state-of-the-art visual datasets, regardless of their heterogeneous formats and taxonomies. One of the key differences between our approach and existing methods is that ours is knowledge-based rather than metadata-based. It enhances the enrichment of the semantics at both image and instance levels and offers various data retrieval and exploratory services via SPARQL. **VisionKG** currently contains 519 million RDF triples that describe approximately 40 million entities, and are accessible at `https://vision.semkg.org` and through APIs. With the integration of 30 datasets and four popular CV tasks, we demonstrate its usefulness across various scenarios when working with CV pipelines.

1 Introduction

Computer vision has made significant advances and visual datasets have become a crucial component in building robust visual recognition systems. The performance of the underlying deep neural networks (DNNs) in the systems is influenced not only by advanced architectures but also significantly by the quality of training data [59]. There are many available visual datasets, e.g., ImageNet [9],

OpenImage [28], and MS-COCO [33], which offer a range of visual characteristics in different contexts to improve the generalization capabilities of advanced machine learning models.

However, these datasets are often published in different data formats, and the quality of taxonomies and annotations varies significantly. Furthermore, labels used to define objects are available in diverse lexical definitions, such as WordNet [34], Freebase [4], or even just plain text. As a result, there may be inconsistencies in semantics across multiple datasets [30]. Isolated and non-unified datasets not only create unnecessary overhead when building robust visual recognition systems, but they also introduce biases into learning systems and limit the capabilities of data-centric AI [46].

Although researchers and practitioners have made efforts to unify visual datasets [29,19,36], a systematic approach to understanding the features and annotations underlying visual datasets is still lacking. For example, the DeepLake [19] can access data from multiple data sources in a unified manner, however, it does not bridge the gap in linking and managing these datasets. Fiftyone [36] can partially capture inconsistencies in multiple datasets by visualizing data sets and analyzing data pipeline failures. Although these works improve the performance of the learned model in a data-centric manner, training DNNs with high-quality data from multiple sources in a cost-effective way remains a formidable challenge for researchers and engineers [51].

Knowledge graph [20] offers a flexible and powerful way to organize and represent data that is comprehensible for both humans and machines. Thus, to systematically organize and manage data for computer vision, we built a knowledge graph of the visual data, named VisionKG. VisionKG is designed to provide unified and interoperable semantic representations of visual data that are used in computer vision pipelines. This knowledge graph captures the entities, attributes, relationships, and annotations of the image data, enabling advanced mechanisms to query training data and perform further analysis.

To address the data inconsistency problems mentioned above, VisionKG interlinks annotations across various datasets and diverse label spaces, promoting a shared semantic understanding and facilitating the retrieval of images that meet specific criteria and user requirements. For instance, for training and testing a specific system, developers may require images with specific types and attributes tailored to their particular scenarios across a range of visual tasks or sources. For example, pedestrian and vehicle detection in adverse weather conditions [42] or occlusion-aware pose estimation [23] both require such tailored image sets across multiple sources for training and testing. Our approach also enables users to better explore and understand relationships between entities using facet-based visualization and exploration powered by a graph data model. Graph queries powered by a graph storage can be employed to create declarative training pipelines from merged computer vision datasets, providing a convenient way to navigate and discover patterns among interlinked visual datasets such as KITTI [15], MS-COCO [33], and Cityscapes [7]. Additionally, VisionKG offers enhanced flexibility in terms of data representation and organization, enabling

faster and easier access to the necessary information, which supports developers in building training pipelines more conveniently and efficiently.

VisionKG is built based on the Linked Data principles [3], adhering to the FAIR [52] and open science guidelines [5], and encompasses various data sources. These sources have been defined and maintained by the research community, as they are widely used and have a significant impact on the development of computer vision algorithms and systems. Their popularity ensures that they will be regularly and frequently updated and extended. This makes VisionKG a valuable resource for researchers and developers who require access to the newest, high-quality image data. Our main contributions are summarized as follows:

- We provide a unified framework for representing, querying, and analysis of visual datasets. By aligning different taxonomies, we minimize the inconsistency between different datasets.
- We make these datasets accessible via standardized SPARQL queries. It is available in both web user interface and via APIs.
- We demonstrate the advantages of VisionKG via three use cases: composing visual datasets with unified access and taxonomy through SPARQL queries, automating training and testing pipelines, and expediting the development of robust visual recognition systems.
- Currently, VisionKG contains 519 million RDF triples that describe approximately 40 million entities from 30 datasets and four popular CV tasks.

The remainder of the paper is structured as follows. In Section 2, we present detailed steps that follow the Linked Data publishing practice [3] to enforce the FAIR principles [52] in VisionKG. Section 3 presents the infrastructure of our VisionKG framework. In Section 4, we demonstrate the MLOps use cases with VisionKG and how it promotes this process. Sections 5-6 discuss related works and conclusions, respectively.

2 Enforcing FAIR Principles for Visual Datasets

2.1 Making Visual Data Assets *Findable* and *Accessible*

To ensure the **findability** of visual data assets, VisionKG uses Uniform Resource Identifiers (URIs) to identify resources, including images and their associated metadata. These URIs provide a unique and persistent identifier for each resource, making it easy to find and access specific images or sets of images. Figure 1 ① illustrates an RDF data snippet linking images and their annotations in COCO [33], KITTI [15] and VisualGenome [25].

This pays the way to use standardized or popular vocabularies/ontologies, such as DCAT and Schema.org to enrich metadata associated with the content and context of image data in Section 2.2. These metadata can be used to facilitate searching, filtering, and discovery of images based on specific criteria, such as object category or image resolution as demonstrated later in Section 3.3. In particular, VisionKG links each piece of metadata to a URI for the corresponding

image to ensure that metadata clearly and explicitly describe the image they refer to, e.g 'containing' bounding boxes of 'person', 'pedestrian' or a 'man' in Figure 1 ①. This not only enables easy retrieval and exploration of images and their related ones based on their metadata but also ensures that more metadata can be incrementally enriched by simply adding more RDF triples linked to the corresponding image. Such desired features are powered by a triple storage in terms of storing, indexing and querying (cf. Section 3)

In this context, VisionKG can greatly facilitate the **accessibility** of data and metadata by using standardized communication protocols and supporting the decoupling of metadata from data. Its publication practice makes it easier for targeted users to access and reuse relevant data and metadata, even when the original data are no longer available. For instance, several images of Imagenet or MSCOCO were downloaded or extracted from web sources, the metadata will provide alternative sources even the original sources are no longer accessible.

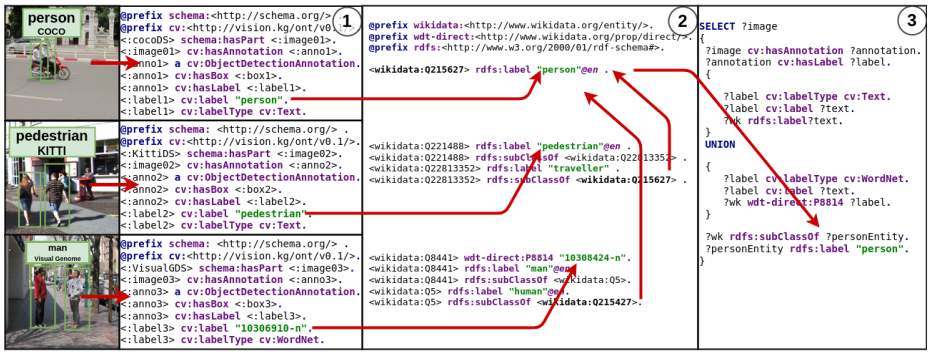

Fig. 1. FAIR for Visual Data Assets

To push the **accessibility** of VisionKG's data assets even further, users can access VisionKG through a well-documented web interface and a Python API. Both interfaces allow users to explore different aspects of VisionKG, such as the included tasks, images, and annotations with diverse semantics. Additionally, many SPARQL query examples.[4,5] enable users to explore the functionalities of VisionKG in detail and describe their requirements or specific criteria using RDF statements.

2.2 Ensure *Interoperability* across Datasets and Tasks

To make VisionKG **interoperable** across different datasets, computer vision tasks, and knowledge graph ecosystems, we designed its data schema as an

[4] `https://vision.semkg.org`
[5] `https://github.com/cqels/vision`

RDFS ontology as shown in Figure 2. This schema captures the semantics of the properties of visual data related to computer vision tasks. Our approach makes use of existing and well-developed vocabularies such as `schema.org` wherever possible. This ensures interoperability and backward compatibility with other systems that use these vocabularies and reduces the need for customized schema development.

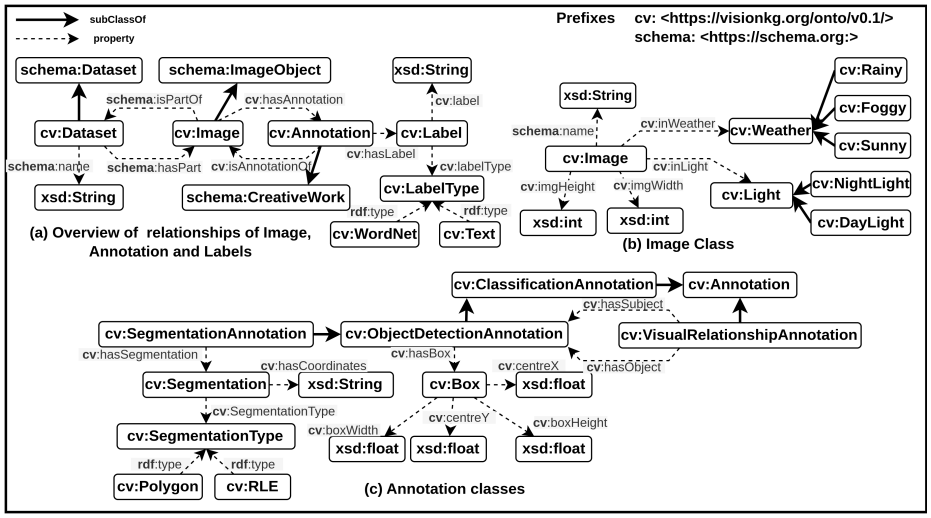

Fig. 2. VisionKG Data Schema

The key concepts in the CV datasets include images, annotations, and labels. To define these concepts, we reuse the `schema.org` ontology by extending its existing classes such as `<schema:ImageObject>`, and `<schema:CreativeWork>`. For example, we extend `<schema:ImageObject>` to create the `<cv:Image>` class, `<schema:Dataset>` to create the `<cv:Dataset>` class. By doing so, we are able to inherit existing properties, such as `<schema:hasPart>` or `<schema:isPartOf>`, to describe the relationships between datasets and images (Figure 2 ⓐ). Our created vocabulary offers the descriptors to capture the attributes of images that are relevant for training a computer vision (CV) model (Section 3.3), such as the image dimensions, illumination conditions, or weather patterns depicted in Figure 2 ⓑ. The concept `Annotation` refers to the labeling and outlining of specific regions within an image. Each type of annotation is used for a particular computer vision task. For instance, bounding boxes are utilized to train object detection models. However, annotations are also reusable for various computer vision tasks. For example, the bounding boxes of object detection annotations can be cropped to train a classification model that doesn't require bounding boxes. In order to enable interoperability of annotations across different computer vision tasks, we developed a taxonomy for them using RDFS ontology, as illustrated in Figure 2 ⓒ. In particular, defining the object detection annota-

tion class as a sub-class of the classification annotation enables the machine to understand that object detection annotations can be returned when users query annotations for a classification task. The cropping process can be performed during the pre-processing step of the training pipeline.

Annotations are associated with labels that define the object or relationship between two objects (visual relationship). However, labels are available in heterogeneous formats, and their semantics are not consistent across datasets. For instance, as shown in Figure 1 ①, the `pedestrian` in KITTI dataset or the `man` in Visual Genome dataset are annotated as `person` in MS-COCO dataset. Furthermore, in the Visual Genome dataset, WordNet [34] identification is used to describe the label. Such inconsistencies make it unnecessarily challenging to combine different datasets for training or testing purposes. To tackle this issue, we assign a specific label type that indicates how to integrate with other existing knowledge graphs to facilitate the **semantic interoperability** across datasets. Figure 1 ② and Figure 1 ③ exemplify how inconsistent labels from three datasets can be aligned using the RDFS taxonomies from WikiData.

2.3 Optimize *Reusability* through SPARQL Endpoint

To optimize the reusability of visual data assets, VisionKG provides a SPARQL endpoint [6] to enable users programmatically discover, combine and integrate visual data assets along with semantic-rich metadata with common vocabularies provided in Section2.2. In particular, users can use powerful SPARQL queries to automatically retrieve desired data across datasets for various computer vision tasks. We provided exemplar queries at `http://vision.semkg.org/`.

Moreover, we annotated VisionKG with data usage licenses for more than ten types[7] of licenses associated with datasets listed in Section 3.2. With this licensed data, users can filter datasets by their licenses to build their own custom datasets. For example, a user can pose a single SPARQL query to retrieve approximately 0.8 million training samples to train a classification model for `cars` with Creative Commons 4.0 license[8].

By linking images and annotations with the original sources and related data curation processes, we captured and shared detailed provenance information for images and their annotations, thus, VisionKG enables users to understand the history and context of data and metadata. By providing such detailed provenance information, VisionKG can enable users to better evaluate the quality and reliability of image and video data and metadata, promoting their reuse.

3 Unified Access for Integrated Visual Datasets

In this Section, we first provide a detailed overview of the architecture of VisionKG, and discuss how it supports access to various popular visual datasets and

[6] SPARQL Endpoint of VisionKG: `https://vision.semkg.org/sparql`
[7] List of dataset licenses in VisionKG: `http://vision.semkg.org/licences.html`
[8] CC BY 4.0: `https://creativecommons.org/licenses/by/4.0/`

computer vision tasks. We then demonstrate VisionKG's capabilities in providing unified access to integrated visual datasets via SPARQL queries, ultimately promoting and accelerating data streaming in CV pipelines. It shows the practical usefulness of our framework for MLOps [1] in Section 4 by exploiting knowledge graph features.

3.1 VisionKG Architecture to Facilitate Unified Access

Figure 3 presents an overview of our VisionKG framework and the process of creating and enriching our unified knowledge graph for visual datasets. We start by collecting popular computer vision datasets for CV from the PaperWithCode platform [9]. Next, we extract their annotations and features across datasets using a *Visual Extractor*. We use RDF Mapping Language (RML)[10] to map the extracted data into RDF. RDF data is generated using a *Semantic Annotator* implemented using RDFizer[21]. To enhance interoperability and enrich semantics in VisionKG, we link the data with multiple knowledge bases, such as WordNet [34] and Wikidata [38]. The *Semantic Enrichment Reasoner* expands the taxonomy by materializing the labels in each dataset using the ontology hierarchy. For instance, categories like `pedestrian` or `man isSubClassOf person` (Figure 1②). Based on the interlinked datasets and Semantic Enrichment Reasoner, users can access the data in VisionKG in a unified way (Figure1 ③). The SPARQL Engine maintains an endpoint for users to access VisionKG using the SPARQL query language.

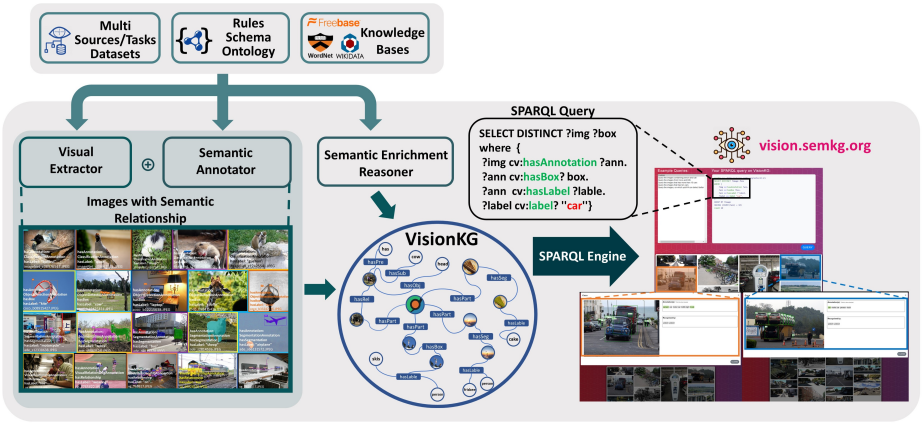

Fig. 3. Overview of VisionKG Platform

Moreover, VisionKG offers a front-end web interface that allows users to explore queried datasets, such as visualizing data distribution and their corresponding annotations (`https://vision.semkg.org/statistics.html`).

[9] `https://paperswithcode.com/datasets`

3.2 Linked Datasets and Tasks in VisionKG

The current version of our framework (by May 2023) integrates thirty most common-used and popular visual datasets, involved in the tasks for visual relationship detection, image classification, object detection, and instance segmentation. Table 1 gives an overview of the contained datasets, images, annotations, and triples in VisionKG. In total, it encompasses over 519 million triples distributed among these visual tasks.

Visual Tasks	**#Datasets**	**#Images**	**#Annotations**	**#Triples**
Visual Relationship	2	119K	1.2M	2.1M
Instance Segmentation	7	300K	3.9M	22.4M
Image Classification	9	1.7M	1.7M	16.6M
Object Detection	12	4.3M	50.8M	478.7M
Total	30	6.4M	57.6M	519.8M

Table 1. Statistics across various Visual Tasks in VisionKG

To enhance the effectiveness of our framework for image classification, we have integrated both large benchmark datasets, such as ImageNet [9], as well as smaller commonly used datasets, like CIFAR [26], the diversity of covered datasets enables users to quickly and conveniently validate model performance, thus avoid extra laborious work. Table 2 demonstrates that ImageNet comprises 1.2 million entities, dominating the distribution of the classification task in VisionKG. Thanks to the interlinked datasets and semantic-rich relationships across visual tasks, users can query different categories and the desired number of images to tailor training pipelines for specific scenarios.

	IMN	**SOP**	**CIFAR**	**MNIST**	**CART**	**Cars196**	**CUB200**
#Entities	2.7M	240K	240K	140K	77K	32.4K	23.6K
#Triples	13.3M	1.2M	1.2M	0.7M	0.4M	0.2M	0.1M

Table 2. Statistics of Triples and Entities in VisionKG for Image Classification. IMN:ImageNet[9], SOP: Stanford Online Products [39], CIFAR:CIFAR10/100 [26], CART: Caltech-101/-256 [16], CUB200: Caltech-UCSD Birds-200-2011 [48].

For object detection, Table 1 and Table 3 show that VisionKG comprises approximately 478 million triples for bounding boxes with dense annotations mainly contributed by large-scale datasets like OpenImages [28] and Objects365 [44]. The variety of visual features allows users to create diverse composite datasets based on their individual requirements for the size or the density of bounding boxes, which can be helpful to reduce biases solely introduced by a single dataset captured under specific conditions and scenarios, e.g., to enhance the model per-

	MSC	UAD	KIT	CAR	BDD	OID	O365	LVIS	MVD	VOC
#Entities	1.0M	678K	47K	32K	1.5M	14.3M	28.5M	1.6M	1.2M	138K
#Triples	9.7M	6.4M	0.4M	0.3M	15.1M	135.8M	277.4M	15.9M	11.9M	1.0M

Table 3. Statistics of Triples and Entities in VisionKG for Object Detection. MSC: MS-COCO[33], UAD: UA-DETRAC [50], KIT: KITTI [15], CAR: StanfordCars196 [24], BDD: BDD100K [55], OID: OpenImages [28], O365: Objects365 [44], LVIS [17], MVD [37], VOC [12]

formance on densely distributed small objects, which are typically challenging to localize and recognize [32,31].

For visual relationship detection, which aims to recognize relationships between objects in images, we have further integrated datasets such as VisualGenome [25] and SpatialScene [54], containing over 1.9 million triples for both bounding boxes and object-level relationships. Besides, VisionKG comprises 22.4 million triples for task instance segmentation, allowing users to retrieve and reuse masks of all instance-level objects for downstream scenarios, thus improving the pixel-level segmentation performance of models.

3.3 Visual Dataset Explorer powered by SPARQL

Organizing training data, which may be in heterogeneous formats and have distinct taxonomies, into one pipeline can be a time-consuming task. To reduce this effort, our framework provides a SPARQL web interface that enables users to access, explore, and efficiently combine data by leveraging the rich semantics of SPARQL. This empowers users to describe their requirements or specific criteria using graph query patterns

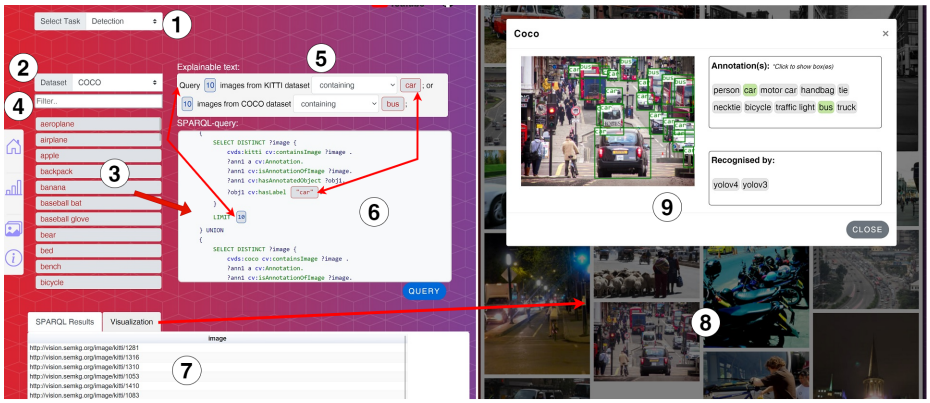

Fig. 4. VisionKG Web Interface

Figure 4 demonstrate our visual datasets explorer equipped with a live-interactive SPARQL web interface. Users can initiate their exploration by selecting a desired task, such as Detection, Classification, Segmentation, or Visual Relationship, from a drop-down menu in Figure4 ①. Upon task selection, the system will promptly generate a list of all compatible datasets that support the chosen task, as Figure4 ② illustrated.

Next, users may choose a dataset, such as COCO [33] or KITTI [15], from the list. This will prompt the system to display all available categories within that dataset in Figure 4 ③. To filter or select the desired categories, users can simply enter a keyword into the text box depicted in Figure 4 ④. This process is further facilitated by allowing users to drag and drop a category from Figure 4 ③ to the query box in Figure 4 ⑥. The system will then auto-generate a SPARQL query, accompanied by an explainable text in Figure 4 ⑤, designed to select images containing the specified category. It is noteworthy that multiple categories from different datasets can be selected. Users may modify the query by removing categories or adjusting the query conditions by selecting available options from boxes in Figure 4 ⑤ or Figure 4 ⑥. Additionally, users can also adjust the number of images to be retrieved.

Once the query is finalized, the user may click the "Query" button, and the results will be displayed in table format in Figure 4 ⑦. Additionally, users may select the "Visualization" tab to view the results graphically, as shown in Figure 4 ⑧. By clicking on an image, users may access additional information, such as annotations of that image and annotations generated from popular deep learning models shown in Figure 4 ⑨. Overall, the platform offers an intuitive and efficient method for dataset selection and querying for machine learning tasks.

4 VisionKG for MLOps

The term *MLOps* refers to the application of the DevOps workflow [11] specifically for machine learning (ML), where model performance is primarily influenced by the quality of the underlying data [1]. As demonstrated in Section 2 and Section 3, the detailed overview of our framework's architecture highlights its significant potential to boost the development of MLOps (e.g., data collection, preparation, and unified access to integrated data). In this section, we present three use cases that demonstrate how to carry out more complicated MLOps steps using our framework. These use cases demonstrate the ability to utilize VisionKG for composing visual datasets with unified access and taxonomy through SPARQL queries, automating training and testing pipelines, and expediting the development of robust visual recognition systems. VisionKG's features enable users to efficiently manage data from multiple sources, reduce overheads, and enhance the efficiency and reliability of machine learning models. More detailed features and tutorials about VisionKG can be found in our GitHub repository[10].

[10] `https://github.com/cqels/vision`

4.1 Composing Visual Datasets with a Unified Taxonomy and SPARQL

Data often comes in a variety of structures and schemas, and there is a need for consolidation of this information in a unified approach. Efficient data management with expressive query functionalities plays a pivotal role in MLOps. However, data from heterogeneous sources with inconsistent formats presents numerous challenges [19,41] that must be addressed to ensure the efficiency and reliability of machine learning models under development. Additionally, the quality and consistency of data and unified access to data are paramount in developing visual recognition systems.

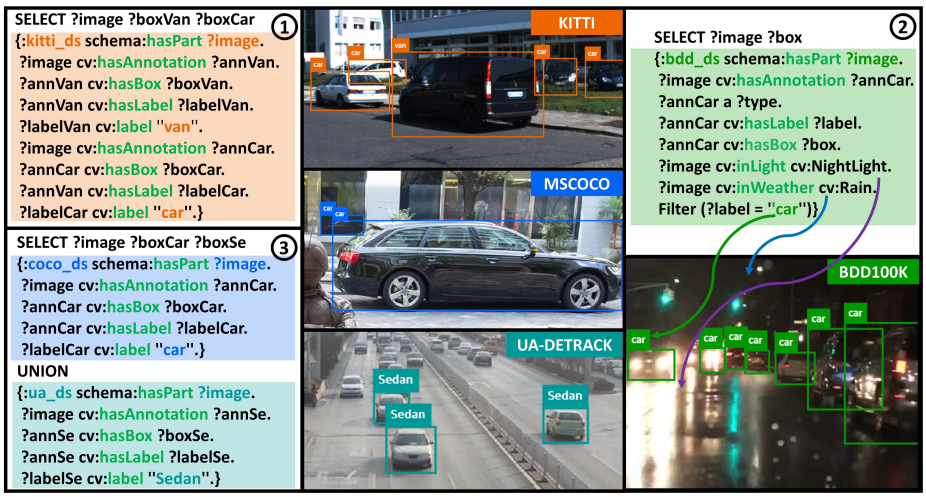

Fig. 5. Dataset-Exploration with SPARQL under various Conditions in VisionKG.

As discussed in Section 2 and 3, VisionKG is equipped with SPARQL engine allowing developers to programmatically build a composite dataset (from diverse sources with different annotated formats) to significantly reduce considerable effort in data preparation phase in MLOps. For instance, as demonstrated in Figure 5 ① and Figure 5 ③, users can query for part of images or categories from one dataset, e.g., images containing both `car` and `van` from KITTI [15]. Besides, as desired, they can also query for images from multiple sources with heterogeneous formats, e.g., images containing `car` from MS-COCO[33] and `sedan` from UA-DETRAC[50] datasets, even though they have far different annotated formats (i.e., annotations of MS-COCO and UA-DETRAC are organized in JSON and XML format, respectively). Furthermore, benefiting from the Semantic Enrichment Reasoner described in Section 3.1 and integrated knowledge bases (e.g., WordNet [34]), users can query for images containing `person` from MS-COCO, KITTI, and Visual Genome [25] (due to distinct taxonomies, `person` are annotated as `pedestrian` in KITTI and labeled as `man` Visual Genome) using a simple

query (Figure 1 ③) rather than a more complex query (Figure 1 ①) that covers all possible cases: e.g., images which containing `pedestrian` in KITTI and/or `man` in Visual Genome dataset, as users desired.

Thanks to the semantic interoperability (cf. Section 2.2) of interlinked annotations across diverse label spaces, users can create datasets from various sources with relevant definitions as desired. Along with the enrichment of semantic relationships, VisionKG provides users with composite visual datasets in a cost-efficient and data-centric manner and hence boosts the data flow in MLOps. Consider the Robust Vision Challenge[11] (RVC) in the context of object detection, where participants need to download terabyte-level datasets from the web (from different sources, taxonomies, and with inconsistent formats), and then train a unified detector to classify and localize objects across all categories in these datasets. To accomplish this, one approach is to unify the taxonomies from these datasets and mitigate the bias introduced by specific domains or similar categories (e.g., the `stop sign` in MS-COCO is a hyponym of `traffic sign`, which annotated in MVD[37]) from different taxonomies. Although the organizers provide manually aligned annotations as a good starting point, unifying labels from distinct taxonomies can still be a time-consuming process. With VisionKG, it is one step closer to achieving this. Users can carry out this process with the assistance of external knowledge bases like Wikidata [47], thereby leveraging external knowledge and facts. Additionally, the unified data model that leverages RDF and knowledge graphs, along with the SPARQL endpoint, allows users to conveniently query specific parts of the datasets as desired without the extra effort of parsing and processing the entire large datasets. This constitutes part of how VisionKG accelerates the MLOps workflow.

4.2 Automating Training and Testing Pipelines

One of the primary goals of MLOps is to automate the training and testing pipelines to accelerate the development and deployment of ML models [1]. Automated workflows enable rapid iteration and experimentation, avoiding the time-consuming process during the model development for both researchers and developers. However, despite the advancements of MLOps in increased productivity and reproducibility of experiments, there are also some limitations remain in current MLOps tools (e.g., Kubeflow [2] and MLflow [1]), such as limited support for complex data types and multi-modal data (e.g., images, videos, and audios). Besides, integrating these MLOps tools with existing diverse data infrastructures can be challenging and requires significant effort.

As described in Section 3.3, powered by SPARQL, VisionKG supports automated end-to-end pipelines for visual tasks. Users can start a training pipeline by writing queries to construct various composite visual datasets. As demonstrated in Figure 6 ①, users can query images and annotations with a few lines of SPARQL query to use RDF-based description to get desired data, such as images containing box-level annotations of `car` and `person` from interlinked datasets in VisionKG. In combination with current popular frameworks (e.g., PyTorch, TensorFlow) or toolboxes (e.g., MMDetection [6], Detectron2 [53]), users

[11] `http://www.robustvision.net/`

can further utilize the retrieved data to construct their learning pipelines in just a few lines of Python code without extra effort, as Figure 6 ② demonstrated. Users need to define solely the model they want to use and the hyperparameters they want to set.

Additionally, users can use VisionKG for their testing pipeline. The inference results can be integrated with data from VisonKG to provide quick insight about the potential model for specific scenarios. Figure 6 ③ demonstrates that one can gain quick overview of a trained YOLO model [22] to detect `car` on images containing `car` in crowded traffic scenes.

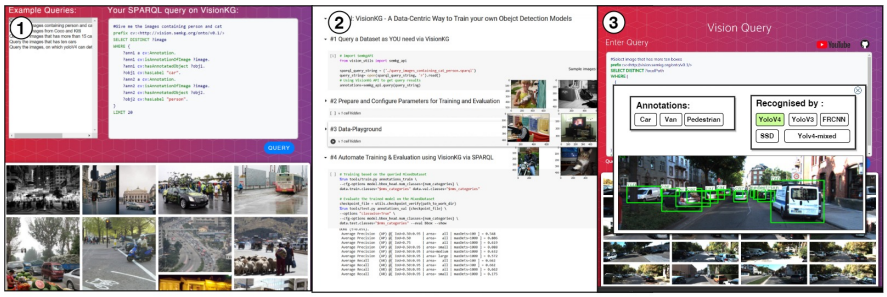

Fig. 6. Construct CV Pipelines Employing VisionKG.

These described features above significantly reduce the workload during data collection, preparation, pre-processing, verification, and model selection for MLOps. Further features of automated pipelines using VisionKG can be found in GitHub repository[12].

4.3 Robust Visual Learning over Diverse Data-Sources

The increasing demand for robust visual learning systems has led to the need for efficient MLOps practices to handle large-scale heterogeneous data, maintain data quality, and ensure seamless integration between data flow and model development. Moreover, a robust learning system should perform consistently well under varying conditions, such as invariance to viewpoint and scale, stable performance under instance occlusion, and robustness to illumination changes. However, many existing visual datasets are specifically designed and curated for particular tasks, often resulting in a limited distribution of image data applicable only in narrowly defined situations [40]. This not only imposes unnecessary burdens when developing robust visual recognition systems but also introduces biases within learning systems and constrains the robustness of visual recognition systems.

As discussed in Section 4.1 and 4.2, users can use VisionKG to compose datasets across interlinked data sources and semantic-rich knowledge bases and

[12] `https://github.com/cqels/vision`

automatically build training and testing pipelines starting from SPARQL queries. This paves the way to support the construction of robust learning systems exploiting features from VisionKG. For instance, when users want to develop a robust object detector, besides bounding boxes and annotated categories, other environmental situations should also be considered and incorporated as prior knowledge to boost the robustness of trained detectors, such as weather and illumination conditions. Using VisionKG, as demonstrated in Figure 5②, users can also employ fine-grained criteria for retrieving images with annotations, such as querying for "images captured at `night` showing `cars` in `rainy` weather conditions." This extends VisionKG's functionalities further for exploring and constructing datasets, allowing users to explore fruitful visual features as desired and build models that cater to various scenarios more robustly, e.g., images captured during adverse weather conditions or at different times of the day. This potential can assist users in evaluating the capability of domain transfer of models (e.g., if a detector trained on KITTI[15] is also robust to detect `cars` in `snowy` weather conditions) or handle rare categories and long-tail phenomenon [56] (e.g., query for a composite dataset containing specific categories which are rare in the source dataset to balance the data distribution).

These features reduce the bias arising from unrelated samples and also enable users to construct scenario-specific datasets covering rich semantics in a convenient fashion. In this way, it allows the users to build robust training pipelines in both data- and model-centric manners.

5 Related Work

Limitations in Existing Computer Vision Datasets
Modern computer vision models are data-intensive and rely heavily on available datasets to perform the learning progress and update learnable parameters. However, the majority of visual datasets are typically limited to specific domains with diverse taxonomies, and the imbalanced nature of class distribution, such as KITTI [15] and MS-COCO [33]. Model-centric approaches, like [49] [57], have trained models to deal with those issues, they require either a domain adapter or adopt an additional model to learn the distribution of unified datasets. However, both model-centric solutions demand extra computing power. Data-centric approaches such as MSeg [29] attempt to unify and interlink datasets manually which is labor-intensive. Besides, existing data toolchains or data hubs like Deep Lake [19], Hugging Face [13] and OpenDataLab [14] are well-established data infrastructures for organizing datasets from distinct web sources. However, these toolchains are based solely on meta-data and do not interlink images and annotations across datasets. In contrast, our framework employs knowledge graphs and diverse external knowledge bases to achieve this and adheres to the FAIR principles [52], enabling VisionKG to interlink images and annotations across visual datasets and tasks with semantic-rich relationships.

[13] Hugging Face. `https://huggingface.co/docs/datasets/index`
[14] Opendatalab: `https://github.com/opendatalab/opendatalab-python-sdk`

Knowledge Graph Technologies in Computer Vision
Knowledge graphs can enhance the utilization of background knowledge about the real world and capture the semantic relationships in images and videos through external knowledge and facts [8,58]. Approaches such as KG-CNet [13] integrate external knowledge sources like ConceptNet[45] to capture the semantic consistency between objects in images. KG-NN [35] facilitates the conversion of domain-agnostic knowledge, encapsulated within a knowledge graph, into a vector space representation via knowledge graph embedding algorithms. However, even these methods leverage external knowledge during learning or after the learning procedure, whereas our method utilizes not only the external knowledge bases but also interlinked datasets. In this way, the enhanced semantics can serve to render fruitful features for integrated datasets. The approach presented in [14] and [38] use Wikidata [47] to empower and interlink annotations for ImageNet [27]. Thanks to the knowledge and facts from the external knowledge base, the data quality has been improved, but both are labor-intensive and mainly target the specific dataset. Besides, the re-usability of these two approaches for other large visual datasets, such as OpenImages [28] and Objects365 [44], and knowledge bases, e.g., Freebase, have not been investigated. [18] employed knowledge graphs to interlink datasets. However, this approach mainly focuses on three datasets in the context of autonomous driving scenarios. In contrast, our framework, VisionKG, utilizes diverse knowledge bases such as WordNet [34], Wikidata [47], and Freebase [4] to enhance the semantics in both image- and instance-level. Additionally, KVQA [43] is a knowledge-based visual dataset employing Wikidata. It is restricted mainly to `person` entities. Different from it, our work interlinks various visual datasets and numerous entities across diverse taxonomies and domains.

6 Conclusions and Future Works

We provide a novel VisionKG that serves as a unified framework for accessing and querying state-of-the-art CV datasets, regardless of heterogeneous sources and inconsistent formats. With semantic-rich descriptions, high-quality, and consistent visual data, it not only helps to facilitate the automation of the CV pipelines but also is beneficial for building robust visual recognition systems.

As new large-scale datasets emerge, there is an increasing need to develop more efficient methods for querying and managing such a huge amount of data. As future work, we will utilize advanced indexing techniques, query optimization, and leveraging distributed computing technologies to improve scalability and integrate further datasets.

7 Acknowledgements

This work was funded by the German Research Foundation (DFG) under the COSMO project (ref. 453130567), the German Ministry for Education and Research via The Berlin Institute for the Foundations of Learning and Data (BIFOLD, ref. 01IS18025A and ref. 01IS18037A), and the European Union's Horizon WINDERA under the grant agreement No. 101079214 (AIoTwin), and RIA research and innovation programme under the grant agreementNo. 101092908 (SmartEdge).

References

1. Alla, S., Adari, S.K., Alla, S., Adari, S.K.: What is mlops? Beginning MLOps with MLFlow: Deploy Models in AWS SageMaker, Google Cloud, and Microsoft Azure pp. 79–124 (2021) 3, 4, 4.2
2. Bisong, E., Bisong, E.: Kubeflow and kubeflow pipelines. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners pp. 671–685 (2019) 4.2
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked data: The story so far. In: Semantic services, interoperability and web applications: emerging concepts, pp. 205–227. IGI global (2011) 1
4. Bollacker, K., Cook, R., Tufts, P.: Freebase: A shared database of structured general human knowledge. In: AAAI. vol. 7, pp. 1962–1963 (2007) 1, 5
5. Budroni, P., Claude-Burgelman, J., Schouppe, M.: Architectures of knowledge: the european open science cloud. ABI Technik **39**(2), 130–141 (2019) 1
6. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019) 4.2
7. Cordts, M., Omran, M., Ramos, S., Scharwächter, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset. In: CVPR Workshop on the Future of Datasets in Vision. vol. 2. sn (2015) 1
8. Cui, P., Liu, S., Zhu, W.: General knowledge embedded image representation learning. IEEE Transactions on Multimedia **20**(1), 198–207 (2017) 5
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 1, 3.2, 2
10. Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R.: Rml: A generic language for integrated rdf mappings of heterogeneous data. Ldow **1184** (2014) 3.1
11. Ebert, C., Gallardo, G., Hernantes, J., Serrano, N.: Devops. Ieee Software **33**(3), 94–100 (2016) 4
12. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**(2), 303–338 (Jun 2010) 3
13. Fang, Y., Kuan, K., Lin, J., Tan, C., Chandrasekhar, V.: Object detection meets knowledge graphs. International Joint Conferences on Artificial Intelligence (2017) 5
14. Filipiak, D., Fensel, A., Filipowska, A.: Mapping of imagenet and wikidata for knowledge graphs enabled computer vision. In: Business Information Systems. pp. 151–161 (2021) 5
15. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013) 1, 2.1, 3, 3.3, 4.1, 4.3, 5
16. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset (2007) 2
17. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5356–5364 (2019) 3
18. Halilaj, L., Luettin, J., Henson, C., Monka, S.: Knowledge graphs for automated driving. In: 2022 IEEE Fifth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE). pp. 98–105. IEEE (2022) 5

19. Hambardzumyan, S., Tuli, A., Ghukasyan, L., Rahman, F., Topchyan, H., Isayan, D., Harutyunyan, M., Hakobyan, T., Stranic, I., Buniatyan, D.: Deep lake: a lakehouse for deep learning (2023) 1, 4.1, 5
20. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G.d., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S., et al.: Knowledge graphs. ACM Computing Surveys (CSUR) **54**(4), 1–37 (2021) 1
21. Iglesias, E., Jozashoori, S., Chaves-Fraga, D., Collarana, D., Vidal, M.E.: Sdm-rdfizer: An rml interpreter for the efficient creation of rdf knowledge graphs. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. pp. 3039–3046 (2020) 3.1
22. Jiang, P., Ergu, D., Liu, F., Cai, Y., Ma, B.: A review of yolo algorithm developments. Procedia Computer Science **199**, 1066–1073 (2022) 4.2
23. Jin, K.M., Lee, G.H., Lee, S.W.: Otpose: Occlusion-aware transformer for pose estimation in sparsely-labeled videos. In: 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 3255–3260. IEEE (2022) 1
24. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13). Sydney, Australia (2013) 3
25. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017) 2.1, 3.2, 4.1
26. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009) 3.2, 2
27. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. NeurIPS (2012) 5
28. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision **128**(7), 1956–1981 (2020) 1, 3.2, 3, 5
29. Lambert, J., Liu, Z., Sener, O., Hays, J., Koltun, V.: Mseg: A composite dataset for multi-domain semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2879–2888 (2020) 1, 5
30. Le-Tuan, A., Tran, T.K., Nguyen, D.M., Yuan, J., Hauswirth, M., Le-Phuoc, D.: Visionkg: Towards a unified vision knowledge graph. In: ISWC (Posters/Demos/Industry) (2021) 1
31. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017) 3.2
32. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017) 3.2
33. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 1, 2.1, 3, 3.3, 4.1, 5
34. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995) 1, 2.2, 3.1, 4.1, 5

35. Monka, S., Halilaj, L., Schmid, S., Rettinger, A.: Learning visual models using a knowledge graph as a trainer. In: The Semantic Web–ISWC 2021: 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24–28, 2021, Proceedings 20. pp. 357–373. Springer (2021) 5
36. Moore, B.E., Corso, J.J.: Fiftyone. GitHub. Note: https://github.com/voxel51/fiftyone (2020) 1
37. Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the IEEE international conference on computer vision. pp. 4990–4999 (2017) 3, 4.1
38. Nielsen, F.Å.: Linking imagenet wordnet synsets with wikidata. In: Companion Proceedings of the The Web Conference 2018. pp. 1809–1814 (2018) 3.1, 5
39. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4004–4012 (2016) 2
40. Paullada, A., Raji, I.D., Bender, E.M., Denton, E., Hanna, A.: Data and its (dis) contents: A survey of dataset development and use in machine learning research. Patterns **2**(11), 100336 (2021) 4.3
41. Renggli, C., Rimanic, L., Gürel, N.M., Karlaš, B., Wu, W., Zhang, C.: A data quality-driven view of mlops. arXiv preprint arXiv:2102.07750 (2021) 4.1
42. Rothmeier, T., Huber, W.: Performance evaluation of object detection algorithms under adverse weather conditions. In: Intelligent Transport Systems, From Research and Development to the Market Uptake: 4th EAI International Conference, INTSYS 2020, Virtual Event, December 3, 2020, Proceedings. pp. 211–222. Springer (2021) 1
43. Shah, S., Mishra, A., Yadati, N., Talukdar, P.P.: Kvqa: Knowledge-aware visual question answering. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8876–8884 (2019) 5
44. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8430–8439 (2019) 3.2, 3, 5
45. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: An open multilingual graph of general knowledge. AAAI (2017) 5
46. Tran, T.K., Le-Tuan, A., Nguyen-Duc, M., Yuan, J., Le-Phuoc, D.: Fantastic data and how to query them. arXiv preprint arXiv:2201.05026 (2022) 1
47. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014) 4.1, 5
48. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011) 2
49. Wang, X., Cai, Z., Gao, D., Vasconcelos, N.: Towards universal object detection by domain attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 5
50. Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M.C., Qi, H., Lim, J., Yang, M.H., Lyu, S.: Ua-detrac: A new benchmark and protocol for multi-object detection and tracking. Computer Vision and Image Understanding **193**, 102907 (2020) 3, 4.1
51. Whang, S.E., Roh, Y., Song, H., Lee, J.G.: Data collection and quality challenges in deep learning: A data-centric ai perspective. The VLDB Journal pp. 1–23 (2023) 1
52. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.:

The fair guiding principles for scientific data management and stewardship. Scientific data **3** (2016) 1, 5
53. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. `https://github.com/facebookresearch/detectron2` (2019) 4.2
54. Yang, K., Russakovsky, O., Deng, J.: Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2051–2060 (2019) 3.2
55. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2636–2645 (2020) 3
56. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 4.3
57. Zhou, X., Koltun, V., Krähenbühl, P.: Simple multi-dataset detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7571–7580 (2022) 5
58. Zhu, C., Chen, F., Ahmed, U., Shen, Z., Savvides, M.: Semantic relation reasoning for shot-stable few-shot object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8782–8791 (2021) 5
59. Zhu, X., Vondrick, C., Fowlkes, C.C., Ramanan, D.: Do we need more training data? International Journal of Computer Vision **119**(1), 76–92 (2016) 1