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, Berlin, Germany
{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 diverse and fruitful features is a key factor for developing, verifying, and benchmarking advanced computer vision (CV) algorithms and architectures. Most visual datasets are created and curated for specific tasks or with limited data distribution for very specific fields of interest, 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 VisionKG is not only based on metadata but also utilizes a unified data schema and external knowledge bases to integrate, interlink, and align visual datasets. It enhances the enrichment of the semantic descriptions and interpretation at both image and instance levels and offers data retrieval and exploratory services via SPARQL and natural language empowered by Large Language Models (LLMs). VisionKG currently contains 617 million RDF triples that describe approximately 61 million entities, which can be accessed at https://vision.semkg.org and through APIs. With the integration of 37 datasets and four popular computer vision tasks, we demonstrate its usefulness across various scenarios when working with computer vision pipelines.

Keywords: Computer Vision · Knowledge Graph · Linked Data · Ontology · RDF

Resource Type: Datasets/Knowledge Graph
Repository: https://github.com/cqels/vision
Homepage: https://vision.semkg.org
License: MIT

A. Meroño Peñuela et al. (Eds.): ESWC 2024, LNCS 14665, pp. 75–93, 2024.
https://doi.org/10.1007/978-3-031-60635-9_5

1 Introduction

Computer vision (CV) has made significant advances, with visual datasets emerging as a crucial component in developing robust visual recognition systems (VRSs). The performance of the underlying deep neural networks (DNNs) is influenced not only by advanced architectures but also significantly by the quality of learning data [59]. There exist many open visual datasets, e.g., ImageNet [9], OpenImage [26], and MS-COCO [31], which offer a wide range of visual characteristics in different contexts to support the generalization capabilities of DNNs.

However, a challenge arises as these datasets often come in varied data formats and the quality of their taxonomies and annotations differs significantly. Furthermore, labels used to define objects vary in diverse lexical definitions, from structured lexical definitions, such as WordNet [32], Freebase [4], to unstructured plain text. This often leads to semantic inconsistencies across different datasets [28]. Additionally, the lack of semantic integration not only causes unnecessary overhead in developing robust VRSs but also introduces biases in learning systems, thereby constraining the potential of data-centric AI [45].

Although researchers and practitioners have made efforts to unify visual datasets [18,27,34], a systematic approach to understanding the features and unifying semantics underlying visual datasets is yet to be achieved. For instance, DeepLake [18] can access data from multiple sources in a unified manner, however, it does not bridge the gap in semantics alignments across datasets. Conversely, Fiftyone [34] can partially identify inconsistencies in annotations and analyze data pipeline failures by interlinking metadata manually, while this labor-intensive solution limits the effectiveness of building computer vision pipelines. While these works improve the performance of learning systems 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 [50].

To address the aforementioned data inconsistency problems, leveraging the capabilities of Knowledge Graphs, which offer a flexible and powerful way to organize and represent data that is comprehensible for both humans and machines, we built a knowledge graph for visual data, named VisionKG, to systematically organize and manage data for computer vision. This graph is specifically designed to provide unified and interoperable semantic representations of visual data, seamlessly integrated into computer vision pipelines. Furthermore, VisionKG interlinks taxonomies across diverse datasets and label spaces, promoting a shared semantic understanding and enabling efficient retrieval of images that meet specific criteria and user requirements.

Additionally, our approach enables users to better explore and comprehend relationships between entities using facet-based visualization and exploration powered by a graph data model. Graph queries, facilitated by graph storage, can be employed to create declarative training pipelines from merged computer vision datasets, providing a convenient way to navigate and investigate patterns among interlinked visual datasets such as KITTI [15], MS-COCO [31], and Cityscapes [7]. Moreover, VisionKG enhances flexibility in terms of data

representation and organization, ensuring rapid and effortless access to essential visual data information, thereby supporting developers and users in constructing computer vision pipelines conveniently and efficiently.

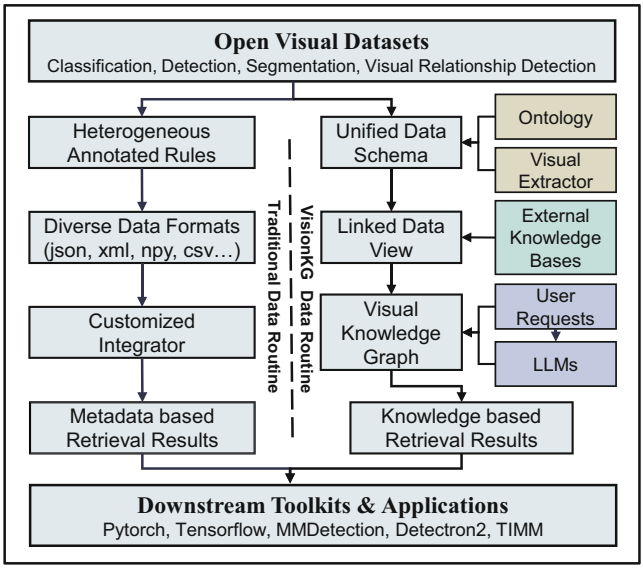

Fig. 1. Differences between VisionKG- and traditional-Data Pipelines

As illustrated in Fig. 1, VisionKG is constructed following the Linked Data principles [3], adhering to the FAIR [51] and open science guidelines [5], and encompasses a large range of 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 VRSs. Their popularity ensures that they will be updated frequently and extended regularly. This makes VisionKG a valuable resource for researchers and developers who require access to the latest, high-quality image data with enriched semantics. Our main contributions are summarized as follows:

- We provide a unified framework for representing, querying, and analyzing visual datasets. By aligning different taxonomies, we minimize the inconsistencies among different datasets.
- Datasets interlinked in VisionKG are made accessible through standardized SPARQL queries, available via both a web user interface and APIs.
- We demonstrate the advantages of VisionKG through 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 comprises 617 million RDF triples that describe approximately 61 million entities from 37 datasets, covering four popular computer vision tasks.

The remainder of the paper is structured as follows: Sect. 2 discusses related work. Section 4 presents the detailed steps to enforce the FAIR principles within VisionKG that follow the Linked Data publishing practice. Section 3 describes the infrastructure of the VisionKG framework. In Sect. 5, we describe the example use cases for VisionKG, e.g., accessing image data with enriched semantics, thereby enhancing the data pipeline for computer vision through standardized SPARQL queries. We present our conclusions in Sect. 6.

2 Related Work

Limitations in Existing Computer Vision Datasets: Modern CV models are data-intensive and their effectiveness is significantly influenced by the quality and diversity of the datasets employed. However, the majority of visual datasets are typically limited to specific domains using heterogeneous taxonomies and exhibit imbalanced class distributions, for example, KITTI [15] and MS-COCO [31] datasets. To address these challenges, model-centric approaches, such as [48,57], either employ a domain adapter or incorporate auxiliary models to discern the distribution across diverse datasets. However, these solutions necessitate additional computational resources and even lead to negative transfer. Data-centric approaches, e.g., MSeg [27], manually unify and interlink datasets, albeit at the cost of increased labor intensity. Moreover, existing visual content extractors [39] or data hubs such as Deep Lake [18], Hugging Face[1], OpenDataLab[2] and MetaVD [53] are well-established data infrastructures for organizing datasets from distinct web sources. However, these toolchains primarily rely on metadata, lacking the ability to interlink images and annotations across datasets. In contrast, our framework leverages knowledge graphs and diverse external knowledge bases, and adheres to the FAIR principles [51], enabling VisionKG to interlink and extend visual datasets and tasks with semantically rich relationships.

Knowledge Graph Technologies in Computer Vision: Knowledge graphs can augment real-world visual recognition with background knowledge, capturing 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 [43] to encapsulate the semantic consistency among objects within images. Similarly, KG-NN [33] enables the conversion of domain-agnostic knowledge, encapsulated within a knowledge graph, into a vector space representation and enhances the model's robustness against domain shift. However, these methods leverage external knowledge during- or post-learning procedure, whereas our solution utilizes not only external knowledge bases but also the knowledge inside interlinked datasets to alleviate the data inconsistency and reduce learning bias. Hence, the enhanced semantics can render fruitful features for integrated datasets benefiting from the interaction between diverse visual features and external knowledge. Similarly, the approach proposed in [14] and

[1] https://huggingface.co/docs/datasets/index.
[2] https://github.com/opendatalab/opendatalab-python-sdk.

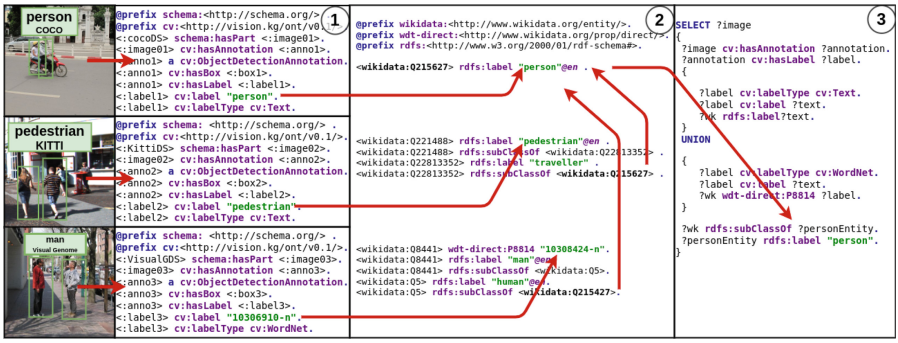

Fig. 2. FAIR for Visual Data Assets

[36] make use of Wikidata [46] to enable and interlink annotations for the ImageNet [25] dataset. Benefiting from that external knowledge and facts, the data quality has been improved, although these approaches are labor-intensive and primarily target a specific dataset. The reusability of these approaches with other extensive visual datasets, such as OpenImages [26] and Objects365 [42], and knowledge bases like Freebase, have not been investigated so far. Approaches proposed by [20,21] use WordNet [32] noun hierarchies and bidirectional ontology engineering to integrate pre-existing knowledge resources and the selection of visual objects within images. In contrast, our VisionKG framework utilizes diverse knowledge bases including WordNet [32], Wikidata [46], and Freebase [4] to enrich the semantics at both image-level and instance-level. Additionally, KVQA [41], another knowledge-based visual dataset employing Wikidata, is restricted mainly to `person` entities, while our work interlinks various visual datasets and numerous entities across diverse taxonomies and domains.

3 Unified Access for Integrated Visual Datasets

This section provides an overview of the VisionKG framework and the details of the integrated visual datasets, and we demonstrate VisionKG's capabilities in providing unified access to those supported visual datasets with enriched semantics, ultimately integrating various data sources with different formats and schemas into CV pipelines seamlessly.

3.1 VisionKG Architecture to Facilitate Unified Access

Figure 3 illustrates the pipeline of creating and enriching the unified knowledge graph of VisionKG for visual datasets. We start by collecting popular computer vision datasets from the PaperWithCode platform[3]. Next, we extract their annotations and features across datasets using a *Visual Extractor*. We use the RDF

[3] https://paperswithcode.com/datasets.

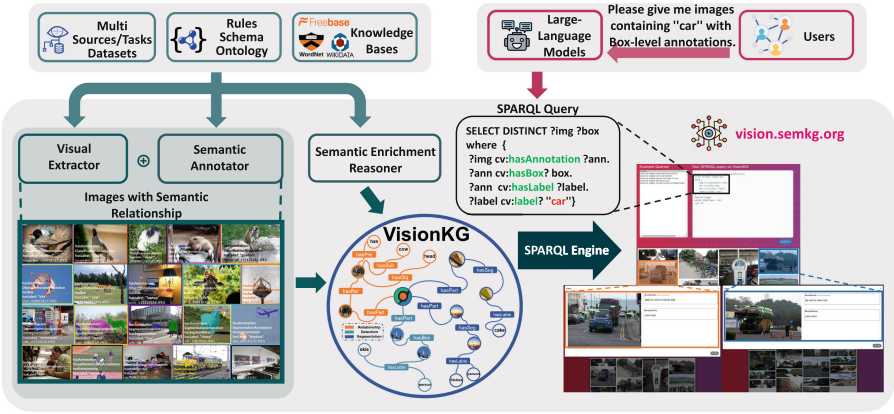

Fig. 3. Overview of VisionKG Platform

Mapping Language (RML) [10] to map the extracted data into RDF. RDF data is generated using a *Semantic Annotator* implemented using RDFizer [19]. To enhance interoperability and enrich semantics in VisionKG, we link the data with multiple knowledge bases, such as WordNet [32] and Wikidata [36]. 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` (Fig. 2 ②). Based on the interlinked datasets and with the Semantic Enrichment Reasoner, users can access the data in VisionKG in a unified way (Fig. 2 ③). The SPARQL Engine maintains an endpoint for users to access VisionKG efficiently. To simplify access to the data, we also built a text-to-SPARQL parser into VisionKG's GUI to translate user requests from natural language into SPARQL using OpenAI APIs.

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

3.2 Linked Datasets and Tasks in VisionKG

As of November 2023, VisionKG integrates the 37 most commonly used visual datasets, across the tasks for visual relationship detection, image classification, object detection, and instance segmentation. Table 1 provides an overview of the datasets, images, annotations, and triples integrated into VisionKG. In total, it encompasses over 617 million triples distributed among these visual tasks. 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 [24]. The diversity of covered datasets enables users to quickly and conveniently validate the effectiveness of learning systems, thus avoiding extra laborious work. As shown in Fig. 4, ImageNet comprises approximately 2.7 million entities across 1.3 million images, dominating

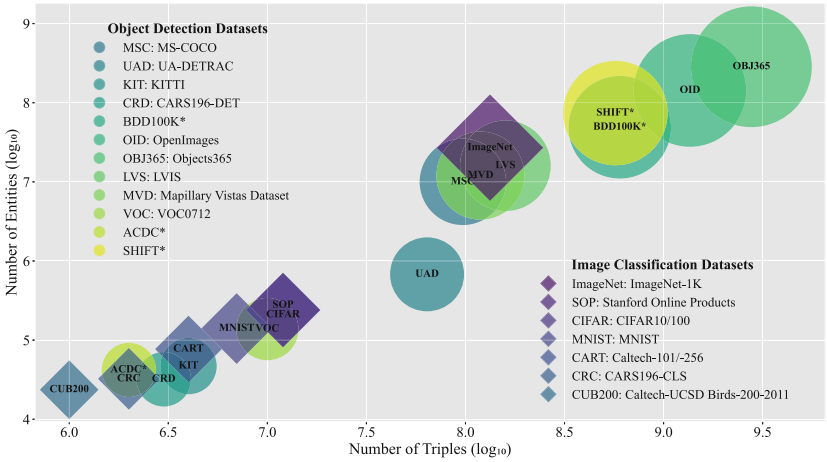

Fig. 4. Statistics of Triples and Entities in VisionKG for Object Detection and Image Classification Datasets. ImageNet [9], SOP [37], CIFAR [24], CART [16], CUB200 [47], MSC [31], UAD [49], KIT [15], CRC [22], BDD100K [55], OID [26], OBJ365 [42], LVS [17], MVD [35], VOC [12], ACDC [40], SHIFT [44]. $*$ presents that synthetic data generated by VisionKG is included.

the distribution of the classification task in VisionKG. Thanks to the interlinked datasets and semantic-rich relationships across visual tasks, users can query images with desired category distributions and contexts to tailor training pipelines for specific scenarios. For object detection, Table 1 reveals that VisionKG includes roughly 576 million triples for 50.8 million box-level annotations primarily derived from large-scale datasets like OpenImages [26] and Objects365 [42]. This variety of visual features enables users to create diverse composite datasets according to specific requirements, such as the size or the density of bounding boxes, which can contribute to mitigating biases inherent in datasets obtained under specific conditions and scenarios, e.g., to enhance the model performance in detecting densely distributed small objects, which are typically challenging to localize and recognize [29,30]. For visual relationship detection, which aims at recognizing object relationships in images, we have integrated datasets such as Visual Genome [23] and Spatial Scene [54], encompassing over 2.1 million triples across 1.2 million annotations for both bounding boxes and object-level relationships. Additionally, VisionKG includes 22.4 million triples for instance segmentation, enabling users to retrieve and reuse masks of all instance-level objects in downstream scenarios.

Table 1. Statistics across various Visual Tasks in VisionKG.

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	19	4.3M	50.8M	576.2M
Total	37	6.4M	57.6M	617.3M

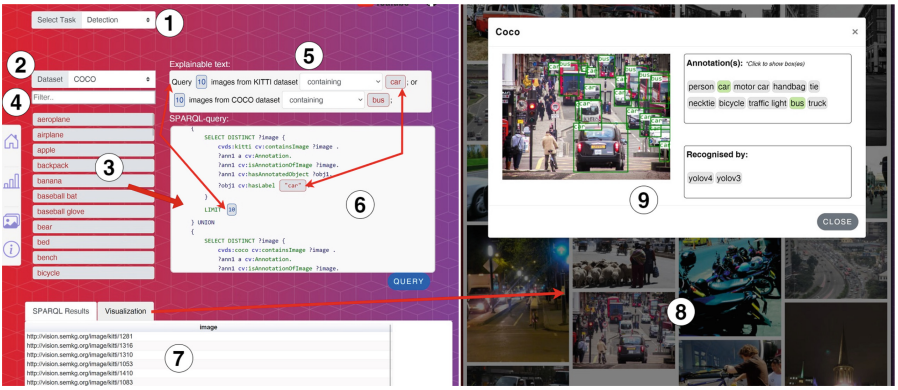

Fig. 5. VisionKG Web Interface

3.3 Visual Dataset Explorer Powered by SPARQL

Organizing multiple datasets, often in heterogeneous formats and with different taxonomies, into a single pipeline is a time-consuming task. To streamline this process, VisionKG offers a GUI for our visual SPARQL frontend, that enables users to access, explore, and efficiently combine data using either the SPARQL query language or natural language, utilizing the text-to-SPARQL parser in VisionKG. This empowers users to specify their requirements or criteria using graph query patterns in an interactive fashion. Figure 5 shows the live-interactive visual datasets explorer in VisionKG: Users can initiate their exploration by selecting a desired task, such as Detection and Classification from a drop-down menu in Fig. 5 ①. Upon task selection, the system will promptly generate a list of all compatible datasets that support the chosen task, as Fig. 5 ② illustrates.

Next, users may choose a dataset, such as MS-COCO [31] or KITTI [15], from the list. This will prompt the system to display all available categories within that dataset in Fig. 5 ③. To filter or select specific categories, users can simply enter a keyword into the text box depicted in Fig. 5 ④. This process is further facilitated by allowing users to drag and drop a category from Fig. 5 ③ to the query box in Fig. 5 ⑥. The system will then auto-generate a SPARQL query, accompanied by an explainable text in Fig. 5 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 the boxes shown in Fig. 5 ⑤ or Fig. 5 ⑥. Additionally, users can also adjust the number of images to be retrieved.

Once the query is finalized, the user can click the `QUERY` button, and the results are displayed in a table format as seen in Fig. 5 ⑦. Additionally, users can select the `Visualization` tab to view the results graphically, as shown in Fig. 5 ⑧. By clicking on an image, users can access additional information, such as meta-data or annotations generated by popular DNNs shown in Fig. 5 ⑨. Overall, VisionKG offers an intuitive and efficient method for users to explore and programmatically query images across diverse datasets, thereby accelerating the data flow in CV pipelines.

4 Enforcing FAIR Principles for Visual Datasets

4.1 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 unique and persistent identifiers for each resource, making it easy to find and access specific images or sets of images. Figure 2 ① illustrates one RDF data snippet linking images and their annotations in MS-COCO [31], KITTI [15], and Visual Genome [23]. This enables the use of standardized or popular vocabularies/ontologies, such as schema.org to enrich metadata associated with the content and context of image data as described in Sect. 4.2. These metadata can be used to facilitate searching, filtering, and discovery of visual content based on specific criteria, such as object category, weather condition, or image resolution as demonstrated in Sect. 3.3. In particular, VisionKG links each piece of metadata to a URI for the corresponding image to ensure that metadata clearly and explicitly describes the image they refer to, e.g., `containing` bounding boxes of a `person`, a `pedestrian` or a `man` in Fig. 2 ①. This not only enables easy retrieval and exploration of target images and their related ones based on their metadata but also ensures that more metadata can be enriched incrementally by simply adding more RDF triples linked to the corresponding image. Such desired features are powered by a triple store for storing, indexing, and querying (cf. Sect. 3).

In this context, VisionKG can significantly improve the **accessibility** of data and metadata by using standardized communication protocols and supporting the decoupling of metadata and data. Its publication practice makes it easier for the targeted user groups to access and reuse relevant data and metadata, even when the original data is no longer available. For instance, several images of ImageNet or MS-COCO were downloaded or extracted from web sources and the metadata will provide alternative sources even if the original sources are no longer accessible.

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 allow users to explore different aspects of VisionKG, such as the included tasks, images, and annotations with diverse semantic attributes. Additionally, many query examples[4,5] enable users to explore the functionalities of VisionKG in detail and specify queries in SPARQL patterns.

4.2 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 RDFS ontology as shown in Fig. 6. The schema captures the semantics of the properties of visual data related to computer vision tasks. Our approach makes use of existing and well-established vocabularies such as schema.org. This ensures interoperability and backward compatibility with other systems using these vocabularies and reduces the need for customized schema development.

The key concepts in the computer vision 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 the `<cv:Image>` class, `<schema:Dataset>` to 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 (Fig. 6 ⓐ). Our newly created vocabulary offers the descriptors to capture the attributes of images that are relevant for training a computer vision (CV) model (Sect. 3.3), such as the image dimensions, illumination conditions, or weather patterns as depicted in Fig. 6 ⓑ.

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 does not require bounding boxes. To enable interoperability of annotations across different computer vision tasks, we developed a taxonomy for them using an RDFS ontology as illustrated in Fig. 6 ⓒ. In particular, defining the object detection annotation class as a subclass 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 preprocessing step of the training pipeline.

Annotations are associated with labels that define the object or relationship between two objects (visual relationship). However, labels use heterogeneous formats, and their semantics are not consistent across datasets. For instance, as shown in Fig. 2 ①, the `pedestrian` in the KITTI dataset or the `man` in the Visual Genome (VG) dataset are annotated as `person` in the MS-COCO dataset.

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

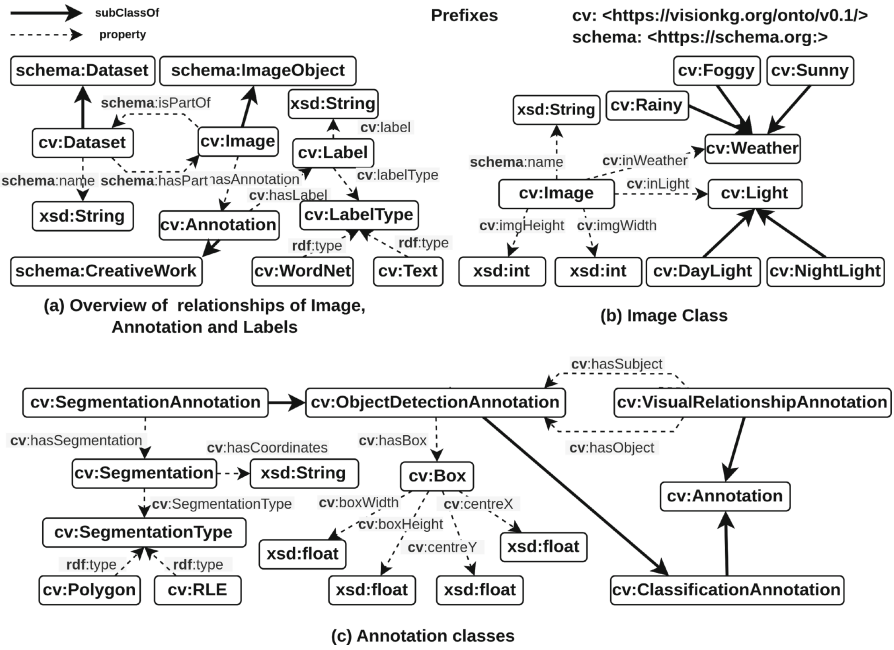

Fig. 6. VisionKG Data Schema

Furthermore, in the VG dataset, WordNet [32] 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 a dataset with other existing knowledge graphs to facilitate the **semantic interoperability** across datasets. Figure 2 ② and Fig. 2 ③ exemplify how inconsistent labels from three datasets can be aligned using the RDFS taxonomies from WikiData.

4.3 Enhance *Reusability* Through a SPARQL Endpoint

To enhance the reusability of visual data assets, VisionKG provides a SPARQL endpoint[6] to enable users to programmatically discover, combine and integrate visual data assets along with semantic-rich metadata with the vocabularies provided in Sect. 4.2. In particular, users can specify SPARQL queries to automatically retrieve desired data across datasets for various computer vision tasks. More exemplary queries are provided at http://vision.semkg.org/.

Moreover, we annotated VisionKG with licensing information for more than ten types[7] of licenses associated with datasets listed in Sect. 3.2, so that users can filter datasets by their licenses to build their own custom datasets. For example,

[6] https://vision.semkg.org/sparql.
[7] List of dataset licenses in VisionKG: http://vision.semkg.org/licences.html.

a user can issue a single query to retrieve approximately 0.8 million classification training samples 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.

5 VisionKG: Facilitating Visual Tasks Towards MLOps

Talk about how to use SPARQL to access images, remove the part of 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]. This requirement underscores VisionKG's significant potential to boost the development of MLOps, including data integration, semantic alignment, and unified access to data across heterogeneous formats and sources. We demonstrate these advantages based on three use cases. More details and use cases are available in our GitHub repository[9].

5.1 Composing Visual Datasets in a Unified Taxonomy

Open visual data [9,26], along with their corresponding annotations, often come in a variety of structures and taxonomies. Considering that efficient data management with unified access and expressive query functionalities plays a pivotal role in MLOps, this necessitates a unified approach to organizing those visual corpora. To ensure the efficiency and reliability of developed ML models and the quality and consistency of visual content, as outlined in Sect. 4 and 3, VisionKG, equipped with a SPARQL engine, enables researchers and developers to build composite datasets in graph queries, regardless of the diversity of sources or heterogeneity of annotated formats.

Initially, users can perform queries for a subset or specific categories within a single dataset, such as images containing `car` and `van` from KITTI [15] (Fig. 7 ①). Additionally, benefiting from interlinked datasets in a unified schema, VisionKG supports users to query images from diverse sources with heterogeneous formats, such as images containing `car` from the MS-COCO [31] and `sedan` from UA-DETRAC [49] datasets, despite their differing annotation formats (Fig. 7 ③).

Additionally, inconsistent taxonomies and a vast number of categories, e.g., 1,000 categories in ImageNet-1K [9] based on WordNet [32] and 500 classes in OpenImages [26] using Freebase [4], cause a challenge to construct composite datasets across multiple sources. For instance, in developing a `person` recognition system, diverse features are required to ensure the robust performance

[8] https://creativecommons.org/licenses/by/4.0/.
[9] https://github.com/cqels/vision.

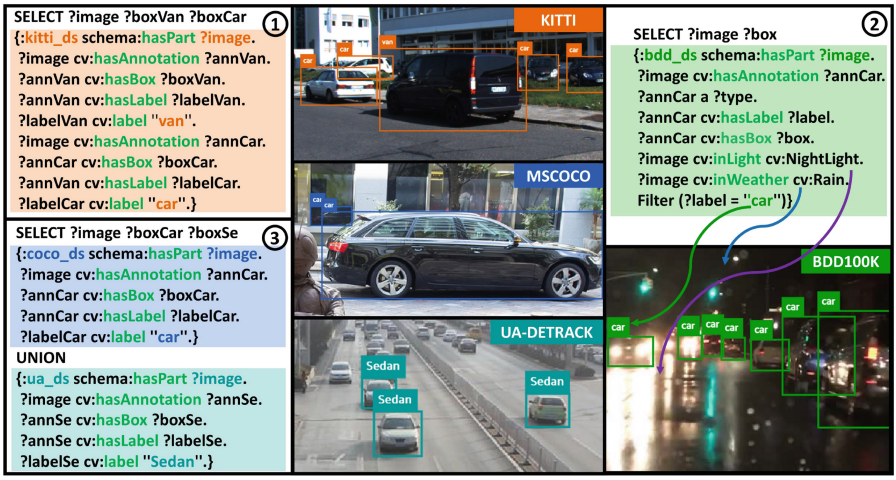

Fig. 7. Retrieving data under various conditions in VisionKG using SPARQL

of learning models. However, there are distinct definitions for the class `person` across different sources and taxonomies: It may defined either as `person` in MS-COCO [31] or `pedestrian` in KITTI [15], or even `man` in Visual Genome [23] (Fig. 2). Hence, unifying labels from distinct taxonomies is not only a time-consuming but also a labor-intensive process. To alleviate it, one approach is to construct a unified taxonomy for these datasets and mitigate the bias introduced by specific domains or similar categories. With VisionKG, we are one step closer to achieving this. Users can carry out this process with the assistance of external knowledge bases, thereby leveraging external knowledge and facts to retrieve interlinked visual content. 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. In this way, VisionKG enables users to query images containing the desired target categories (such as `person`) conveniently (Fig. 2 ③) rather than a more complex query (Fig. 2 ①) that covers all possible cases.

Thanks to the semantic interoperability (Sect. 4.2) of interlinked annotations across diverse label spaces, users can construct 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 supporting the data flow in MLOps.

5.2 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 model development. However, despite the advancements of MLOps, there are some limitations 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, such as Deep Lake [18], can be challenging and requires significant effort.

As described in Sect. 3.3, VisionKG boosts 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 Fig. 3, users can query images and annotations with a few lines of SPARQL code to use RDF-based descriptions to get desired data, such as images containing box-level annotations of `car` and `person` from interlinked datasets in VisionKG. In combination with popular frameworks, e.g., PyTorch, TensorFlow, or toolboxes, e.g., MMDetection [6], Detectron2 [52], users can further utilize the retrieved data to build their learning pipelines with minimal Python code without extra effort. Benefiting from the interlinked datasets and existing model zoo, users solely need to define the model they want to use and the hyperparameters they want to set. Listing 1.1 demonstrates a simplified example code to query VisionKG data and perform training and testing for object detection. More features of automated pipelines using VisionKG are available in our GitHub repository[10].

```
# Import VisionKG utilities and integrated training pipeline
from vision_utils import semkg_api
from torch_model_zoo import utils
from torch_model_zoo.train_eval import train_eval_pipeline

def prepare_vkg_pipeline(query_string):
    # Execute the query
    rels = semkg_api.query(query_string)
    params = utils.prepare_for_training(rels)
    return params

# SPARQL query to VisionKG for images with specific objects
query_string = ''' SPARQL query for object detection '''
params = prepare_vkg_pipeline(query_string)
params['MODEL'] = 'fasterrcnn_resnet50_fpn'
train_eval_pipeline(params)
```

Listing 1.1. A simplified example of VisionKG pipeline for MLOps

5.3 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

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

particular tasks, often resulting in a limited distribution of image data applicable only in narrowly defined situations [38]. This not only imposes unnecessary burdens but also introduces biases within learning systems and constrains the robustness of VRSs.

As discussed in Sect. 5.1, VisionKG enables users to compose datasets across interlinked data sources and semantic-rich knowledge bases, automating 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, in developing a robust object detector, besides bounding boxes and annotated categories, other environmental situations should also be considered and incorporated as prior knowledge to improve the robustness of trained detectors, such as weather and illumination conditions. Using VisionKG, as shown in Fig. 7 ②, users can 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 retrieve useful visual features and build models that cater to various scenarios robustly, e.g., images captured during adverse weather conditions or at different times of the day. This functionality can assist users in evaluating the capability of models in domain transfer, e.g., if a detector trained on KITTI [15] is robust enough to detect `cars` in `snowy` weather conditions or handle rare categories and long-tail phenomena [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 enable users to construct scenario-specific datasets covering rich semantics in a convenient fashion. In this way, VisionKG enables users to build robust visual learning systems in a data-centric manner.

6 Conclusions and Future Works

VisionKG **removes** data integration, heterogeneity and format inconsistency problems from datasets intended for vision computing tasks. It **enhances** the integrated datasets with a host of semantic annotations from knowledge bases. It **provides** a unified SPARQL query interface to this integrated dataset and **offers** powerful exploration tools including a human language interface. All these three functionalities significantly **improve** access, efficiency, and usability to visual training data based on the power of semantic technologies and demonstrating how important these technologies are for any kind of data-intensive research and industrial development. The current version of VisionKG includes 617 million RDF triples describing approximately 61 million entities from 37 datasets and four popular computer vision tasks. VisonKG is easily **extensible** and will empower communities to grow around the provided resources. It can serve as a **blueprint** for many digital data resources as the functionalities provided are generic and reusable.

Acknowledgements. This work is supported by the Deutsche Forschungsgemeinschaft, German Research Foundation under grant number 453130567 (COSMO), by the Horizon Europe Research and Innovation Actions under grant number 101092908 (SmartEdge), by the Federal Ministry for Education and Research, Germany under grant number 01IS18037A (BIFOLD) and by the Horizon Europe Research and Innovation programme under grant agreement number 101079214 (AIoTwin).

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)
2. Bisong, E., Bisong, E.: Kubeflow and kubeflow pipelines. In: Bisong, E. (ed.) Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners, pp. 671–685. Apress, Berkeley (2019). https://doi.org/10.1007/978-1-4842-4470-8_46
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)
4. Bollacker, K., Cook, R., Tufts, P.: Freebase: a shared database of structured general human knowledge. In: AAAI, vol. 7, pp. 1962–1963 (2007)
5. Budroni, P., Claude-Burgelman, J., Schouppe, M.: Architectures of knowledge: the European open science cloud. ABI Tech. **39**(2), 130–141 (2019)
6. Chen, K., et al.: MMDetection: Open MMLab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
7. Cordts, M., et al.: The cityscapes dataset. In: CVPR Workshop on the Future of Datasets in Vision, vol. 2. sn (2015)
8. Cui, P., Liu, S., Zhu, W.: General knowledge embedded image representation learning. IEEE Trans. Multimed. **20**(1), 198–207 (2017)
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)
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)
11. Ebert, C., Gallardo, G., Hernantes, J., Serrano, N.: DevOps. IEEE Softw. **33**(3), 94–100 (2016)
12. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vision **88**(2), 303–338 (2010)
13. Fang, Y., Kuan, K., Lin, J., Tan, C., Chandrasekhar, V.: Object detection meets knowledge graphs. In: International Joint Conferences on Artificial Intelligence (2017)
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)
15. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. Int. J. Robot. Res. **32**(11), 1231–1237 (2013)
16. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset (2007)

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)
18. Hambardzumyan, S., et al.: Deep lake: a lakehouse for deep learning (2023)
19. 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)
20. Koeva, S.: Multilingual image corpus: annotation protocol. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), pp. 701–707 (2021)
21. Koeva, S.: Ontology of visual objects. In: Proceedings of the 5th International Conference on Computational Linguistics in Bulgaria (CLIB 2022), pp. 120–129. Department of Computational Linguistics, IBL – BAS, Sofia (2022). https://aclanthology.org/2022.clib-1.14
22. 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)
23. Krishna, R., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vision **123**, 32–73 (2017)
24. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NeurIPS (2012)
26. Kuznetsova, A., et al.: The open images dataset v4: unified image classification, object detection, and visual relationship detection at scale. Int. J. Comput. Vision **128**(7), 1956–1981 (2020)
27. 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)
28. 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)
29. 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)
30. 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)
31. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
32. Miller, G.A.: Wordnet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
33. Monka, S., Halilaj, L., Schmid, S., Rettinger, A.: Learning visual models using a knowledge graph as a trainer. In: Hotho, A., et al. (eds.) ISWC 2021. LNCS, vol. 12922, pp. 357–373. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88361-4_21
34. Moore, B.E., Corso, J.J.: Fiftyone. GitHub (2020). https://github.com/voxel51/fiftyone

35. 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)
36. Nielsen, F.Å.: Linking ImageNet WordNet synsets with Wikidata. In: Companion Proceedings of the the Web Conference 2018, pp. 1809–1814 (2018)
37. 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)
38. 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)
39. Qin, A., Xiao, M., Wu, Y., Huang, X., Zhang, X.: Mixer: efficiently understanding and retrieving visual content at web-scale. Proc. VLDB Endow. **14**(12), 2906–2917 (2021)
40. Sakaridis, C., Dai, D., Van Gool, L.: ACDC: the adverse conditions dataset with correspondences for semantic driving scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10765–10775 (2021)
41. 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)
42. Shao, S., et al.: 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)
43. Speer, R., Chin, J., Havasi, C.: ConceptNet 5.5: an open multilingual graph of general knowledge. In: AAAI (2017)
44. Sun, T., et al.: SHIFT: a synthetic driving dataset for continuous multi-task domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21371–21382 (2022)
45. 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)
46. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)
47. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-UCSD birds-200-2011 dataset (2011)
48. 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) (2019)
49. Wen, L., et al.: UA-DETRAC: a new benchmark and protocol for multi-object detection and tracking. Comput. Vis. Image Underst. **193**, 102907 (2020)
50. Whang, S.E., Roh, Y., Song, H., Lee, J.G.: Data collection and quality challenges in deep learning: a data-centric AI perspective. VLDB J. 1–23 (2023)
51. Wilkinson, M.D., et al.: The fair guiding principles for scientific data management and stewardship. Sci. Data **3** (2016)
52. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2 (2019). https://github.com/facebookresearch/detectron2
53. Yamamoto, Y., Egami, S., Yoshikawa, Y., Fukuda, K.: Towards semantic data management of visual computing datasets: increasing usability of MetaVD. In: Proceedings of the ISWC 2023 Posters, Demos and Industry Tracks (2023)
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)

55. Yu, F., et al.: 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)
56. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: a survey. IEEE Trans. Pattern Anal. Mach. Intell. (2023)
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)
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)
59. Zhu, X., Vondrick, C., Fowlkes, C.C., Ramanan, D.: Do we need more training data? Int. J. Comput. Vision **119**(1), 76–92 (2016)