

Chapter 2: Introduction to Dataset Sources

Faculty of Digital Engineering
Institute of Digital Technology
2026

Agenda

1. Data Sources for Computer Vision
2. Data Preparation
3. Labelling and Annotation Tool



Data Sources of Computer Vision

Significance of Datasets:

- Datasets provide the input data needed for training, validating, and testing computer vision algorithms.
- High-quality datasets improve the accuracy and generalization of models, while poor datasets can lead to biased or unreliable results.

Data Sources of Computer Vision

Data Sources

- Public datasets
- Proprietary datasets collected for specific projects
- Synthetic data generated using simulation or rendering tools.

Data Source of Computer Vision

Public Datasets

- **Public datasets** are available online through search engines, communities, social media, universities, and data repositories.
- Some public datasets include **labels** that match our purpose, but others may require **relabeling**.
- Public datasets are often published by researchers or universities, which may result in **high-performing benchmarks**.
- Always **cite** public datasets when using them in your research.
- Carefully **verify** dataset specifications and labeling methods before using them in your study.

Data Sources of Computer Vision

Advantages and Disadvantage of using Public Datasets

Advantages	Disadvantages
1. Free access via the Internet 2. Saves time in data collection 3. Useful for research and experimentation 4. Standardized benchmarks for comparison 5. Supports reproducibility and validation	1. Limited customization for specific tasks 2. Requires relabeling or preprocessing 3. Insufficient dataset information at times 4. Hard to outperform prior research results 5. Outdated datasets may not reflect current trends

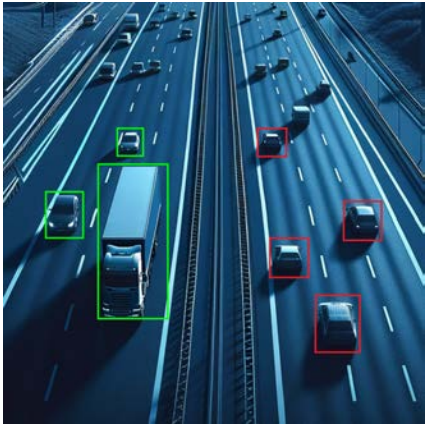
Data Sources of Computer Vision

Examples of public datasets in Computer Vision

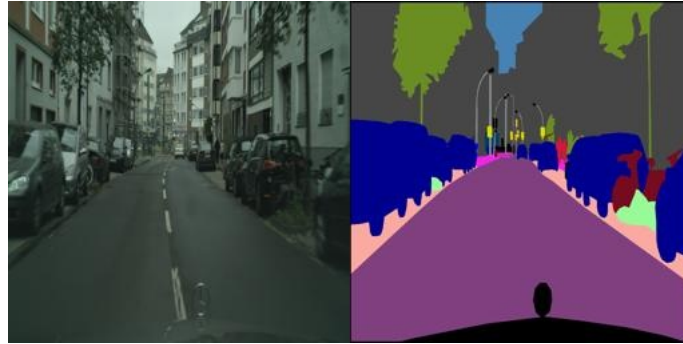
Platform/Dataset	Task	Description
ImageNet	Image Classification	Images across 1,000 classes.
COCO	Object Detection, Segmentation	Common Objects in Context; includes bounding boxes and segmentation masks.
MNIST	Handwritten Digit Recognition	Dataset of 70,000 images of handwritten digits (0 to 9).
KITTI	Autonomous Driving	Dataset for 3D object detection, tracking, and stereo vision.
LFW (Labeled Faces in the Wild)	Face Recognition	Face images for face verification and recognition tasks.
Pascal VOC	Object Detection, Segmentation	Standard dataset for image classification, detection, and segmentation.
Kaggle	All tasks (Classification, Detection, Segmentation, etc.)	A popular platform hosting a wide variety of datasets for competitions and open research.

Data Sources of Computer Vision

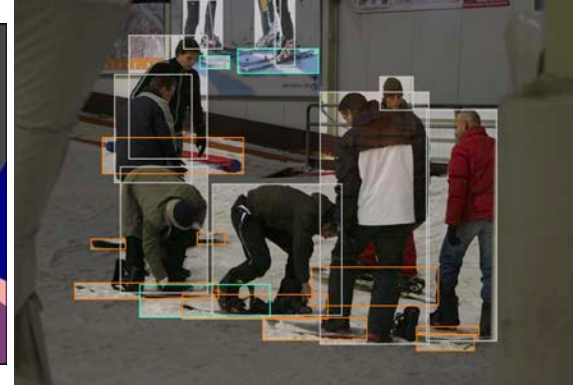
Examples of public datasets in Computer Vision



kaggle Dataset (Vehicle detection)



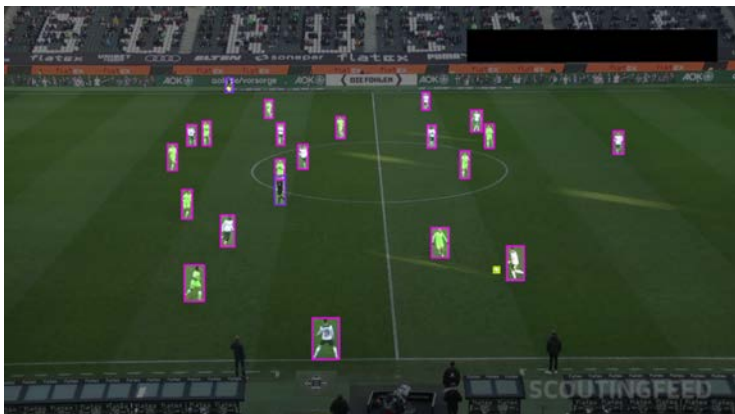
kaggle Dataset (Cityscapes segmentation)



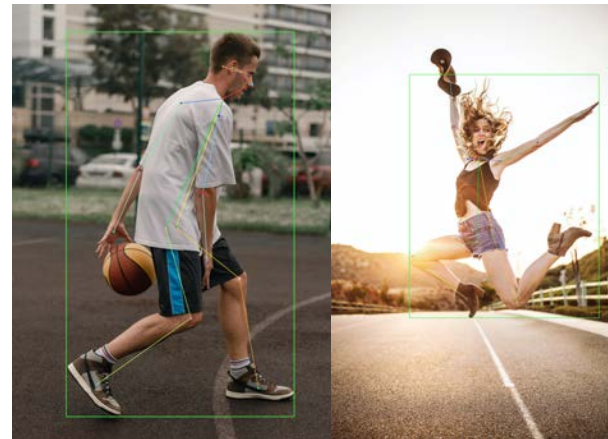
COCO Dataset (Object detection)



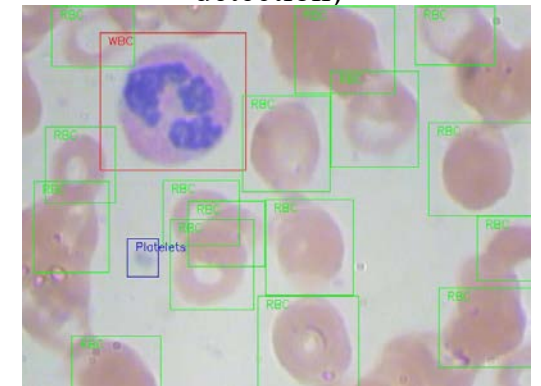
Mask Wearing Dataset (Mask detection)



kaggle Dataset (Football player detection)



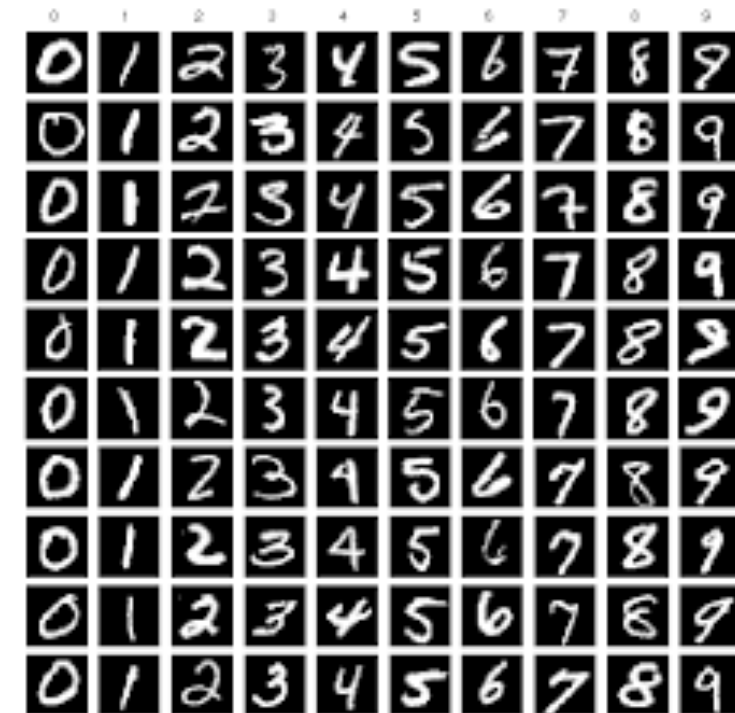
kaggle Dataset (Human pose estimation)



BCCD Dataset (Blood cells detection)

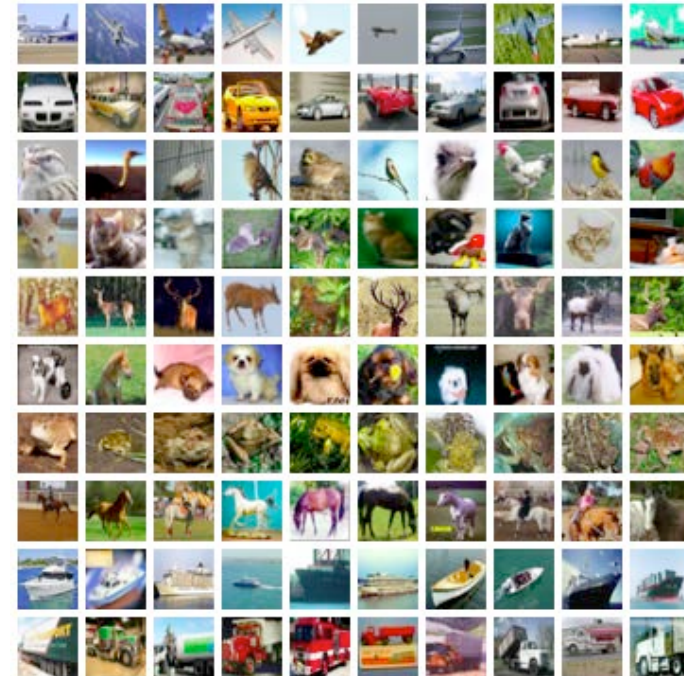
Popular Image Datasets - MNIST

- MNIST: A dataset of handwritten digits (0-9).
- Applications: used for beginner-level image classification tasks and model benchmarking.
- Size: 60,000 training and 10,000 testing images.
- <https://www.kaggle.com/datasets/hojjatk/mnist-dataset>



Popular Image Datasets - CIFAR

- CIFAR: Collections of 32x32 images in 10 or 100 categories, ranging from animals to vehicles.
- Applications: ideal for image classification and object recognition tasks.
- Size: 60,000 images.
- <https://www.kaggle.com/c/cifar-10/>



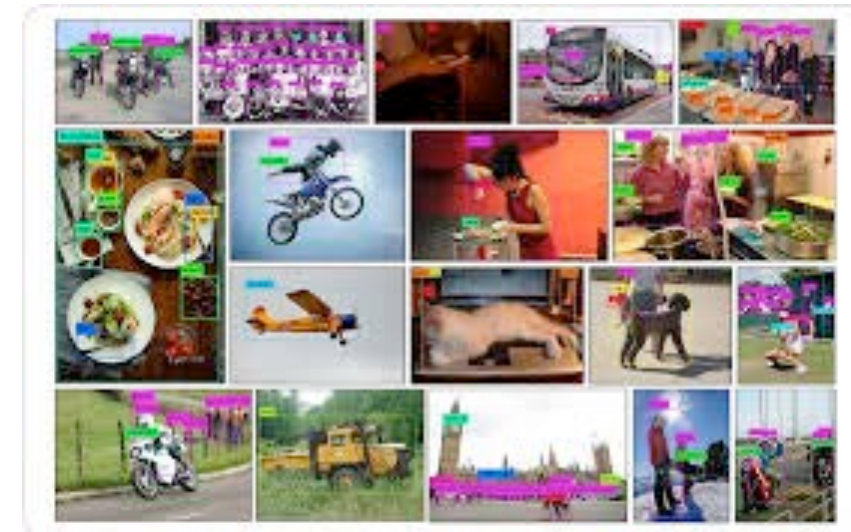
Popular Image Datasets – ImageNet

- ImageNet:
 - Full dataset: Over 14 million images across thousands of categories.
 - Standard subset: 1.2 million images across 1,000 categories.
- Applications: used for large-scale image classification challenges and transfer learning.
- Size: One of the largest datasets for deep learning applications.
- <https://www.image-net.org/>



Popular Image Datasets – COCO

- COCO (Common Objects in COntext): Images annotated with object segmentation, classification, and captions.
- Applications: used in object detection, image segmentation, and image captioning.
- Size: over 330,000 images with 80 object categories.
- <https://cocodataset.org/>



Popular Image Datasets – Open Images

- Open Images: A massive dataset with millions of images annotated for object detection, segmentation, and relationships.
- Applications: suitable for large-scale vision tasks.
- Size: over 9 million images.
- <https://www.kaggle.com/datasets/bigquery/open-images>



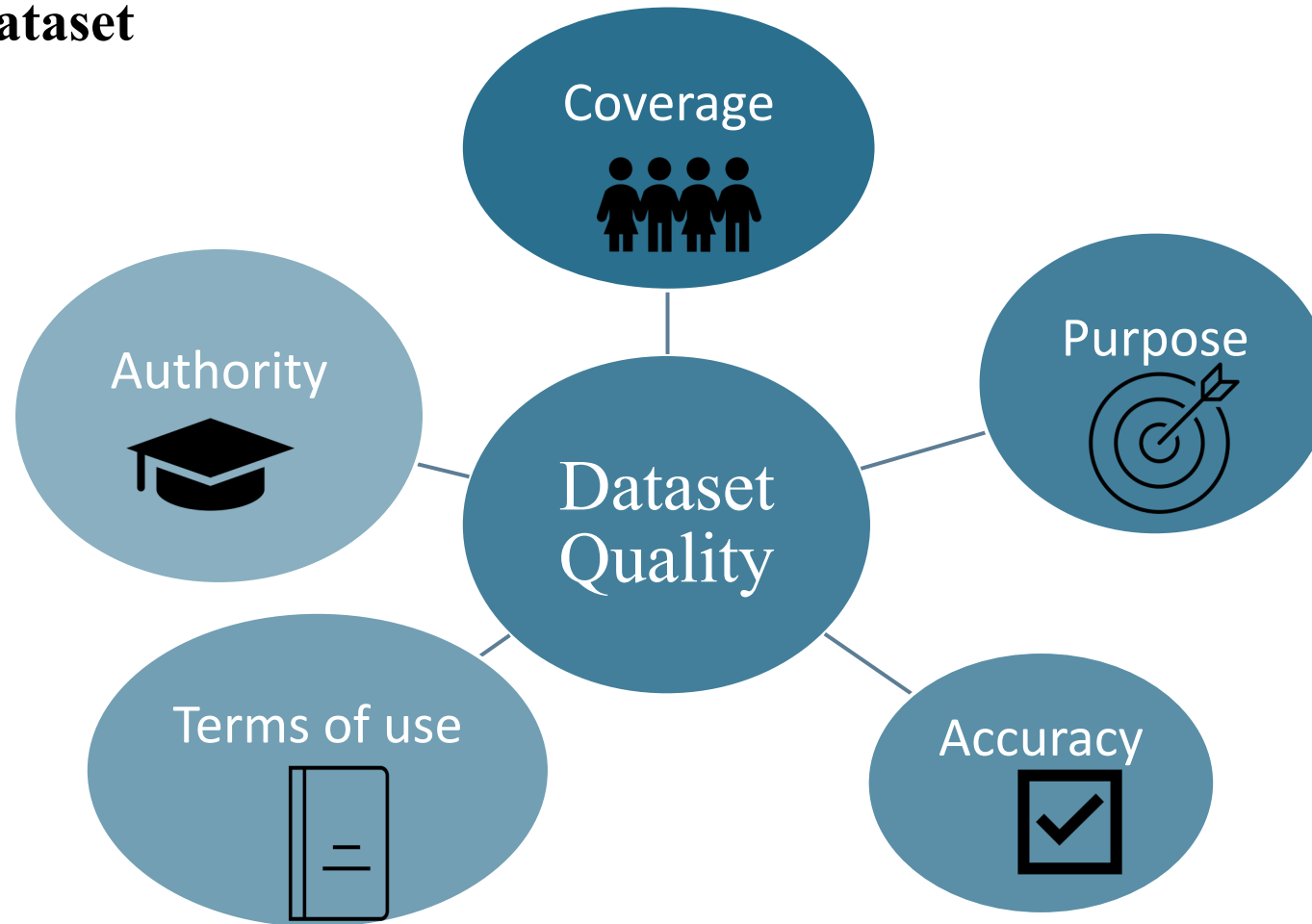
Data Sources of Computer Vision

Build Your Own Dataset

- Collecting your own dataset may involve some or all of the following steps: defining a data-collection procedure, conducting experiments, labeling, annotating, and preparing the data for training the model.
- In some cases, labeling and annotating cannot be done by a non-expert; expertise in a specific field is required.
- A valid dataset yields valid results; therefore, defining a clear procedure for collecting the dataset is crucial.
- To create a generalized dataset, a large amount of data from various conditions is needed to ensure the method can understand all scenarios with minimal error.

Data Sources of Computer Vision

Build Your Own Dataset



Data Sources of Computer Vision

Build Your Own Dataset

The **main significant advantages** and **disadvantages** of data collection are:

Advantages	Disadvantages
Customizability: Tailored data meets the specific needs of the project or research.	Time and Resource Intensive: Data collection is costly, laborious, and can significantly delay project timelines.
High Data Quality: You have control over data accuracy, annotation, and consistency.	Scalability Issues: Building large datasets is challenging due to time, budget, and technical limitations.
Domain-Specific Relevance: Collecting your own data ensures it aligns perfectly with niche or specialized use cases.	Ethical and Privacy Concerns: Collecting sensitive data requires compliance with privacy regulations (e.g., GDPR, HIPAA).
Ownership: Full rights and control over the dataset, avoiding restrictions or licensing issues.	Risk of Bias: Despite careful design, collected data may still introduce sampling or domain-specific biases.

Data Sources of Computer Vision

Generate Synthetic data

- Synthetic data refers to artificially generated data used for training computer vision AI models
- It can be created through various approaches, including rendering engines, simulation software, and generative models
- This method has become a promising solution to address the challenges associated with collecting and labeling real-world data for computer vision tasks.

Data Sources of Computer Vision

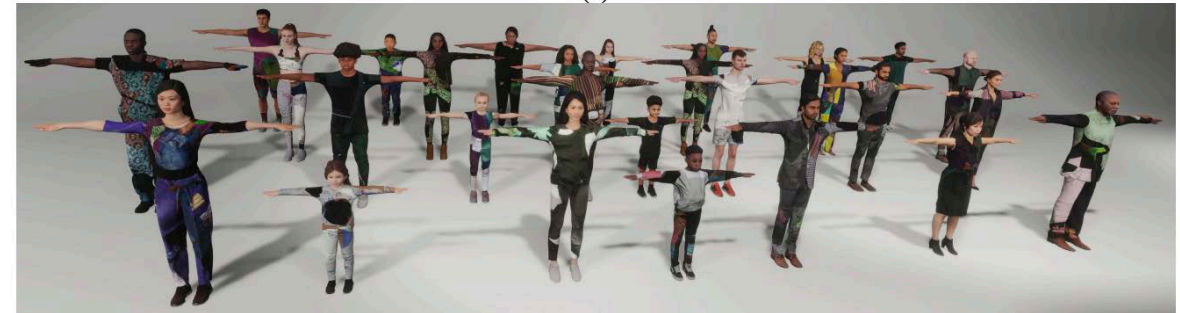
Example of synthetic image



PEOPLESANSPEOPLE generated synthetic
images[<https://arxiv.org/pdf/2112.09290>]



(a)



(b)

PEOPLESANSPEOPLE generated 3D human images
[<https://arxiv.org/pdf/2112.09290>]

Data Sources of Computer Vision

Advantages of Generate Synthetic data



Easy Labelling

Pre-defined labels reduce errors.



Cost-Effective

Generated at a lower cost than real-world data.



Scalable: Quickly produce large amounts of data.



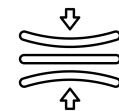
Diverse

Covers rare or dangerous scenarios.



Privacy-Safe

Avoids real-world sensitive information.



Flexible

Simulates difficult or risky situations.

Data Sources of Computer Vision

Limitations of Generate Synthetic Data

Lack of Realism

Struggles to capture real-world complexity and subtle details, leading to less accurate models.

Difficulty with Complex Data

Generating natural language, realistic images, or intricate data requires advanced techniques.

Validation Challenges

Hard to ensure synthetic data accurately reflects real-world trends and anomalies.

Dependency on Real Data

Inaccurate or incomplete real-world data results in unreliable synthetic data.

Bias and Privacy Issues

Existing biases in real data can persist, and clear privacy standards are still evolving.

Data Preparation

- Data preparation is the process of gathering, combining, structuring and organizing data so it can be used in Artificial Intelligence, Business Analytics, Data Science, etc.
- The components of data preparation include **data preprocessing, profiling, cleansing, validation, and transformation.**
- Good data is essential for producing efficient models of any type.

Image datasets

- Image datasets for machine learning are crucial for training models to recognize patterns, classify objects, and make predictions based on visual data.
- ML applications heavily rely on robust and diverse image datasets to deliver accurate and reliable results.

Image datasets

- Image dataset is a collection of images systematically **organized** and **labeled** to support ML tasks.
- The complexity of image datasets can vary widely, ranging from simple sets with a limited number of object categories to highly intricate collections that encompass diverse scenes, and accompany metadata depending on the specific problem they aim to address.

How to choose or prepare an image dataset?

- **Relevance to the task:** ensures the dataset aligns with the application domain.
- **Diversity and Balance:** covers a wide range of scenarios to prevent model bias.
- **Data Quality:** High-resolution images with accurate annotation.
- **Scalability:** Large datasets are necessary for advanced DL models. If an existing dataset is insufficient, data augmentation techniques can help expand it.
- **Privacy and Ethics:** ensures the dataset complies with privacy regulations, especially when working with sensitive data.

Data Annotation

- AI and ML are revolutionizing industries, but how do they understand data?
- Data annotation is a crucial process of labelling data for machine comprehension.
- Data annotation means adding labels or tags to data. This will help systems interpret and process information, making them smarter and more effective.

Data Annotation in Computer Vision Tasks

- Classification tasks: rely on datasets with image labeled by category (e.g., dog, cat, or car)
- Object detection tasks: require datasets with bounding boxes that pinpoint and identify objects within an image.
- Segmentation tasks: involve datasets with pixel-level annotations, providing detailed insights for more precise image analysis.

Why is Data Annotation Important?

- Increase Accuracy: Precise labeling enhances model performance.
- Accelerates Learning: Helps models understand and learn faster.
- Improves Usability: Results in user-friendly AI experiences.
- Boosts Reliability: Ensures models perform consistently and accurately.

Steps in Data Annotation

- Collect Data: Gather text, images, or audio.
- Preprocess Data: Prepare and clean the data.
- Choose Tools: Select suitable annotation software.
- Set Guidelines: Define clear rules for annotators.
- Quality Check: Review annotations for accuracy.
- Export Data: Prepare data for use in AI models.

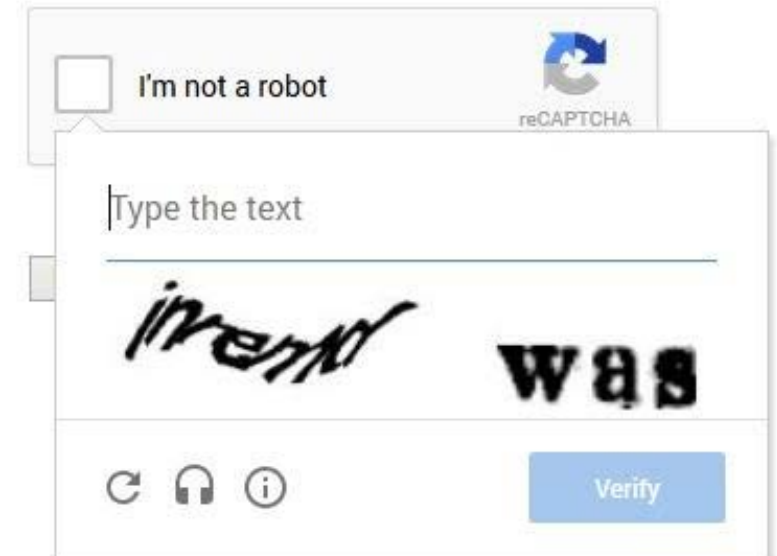
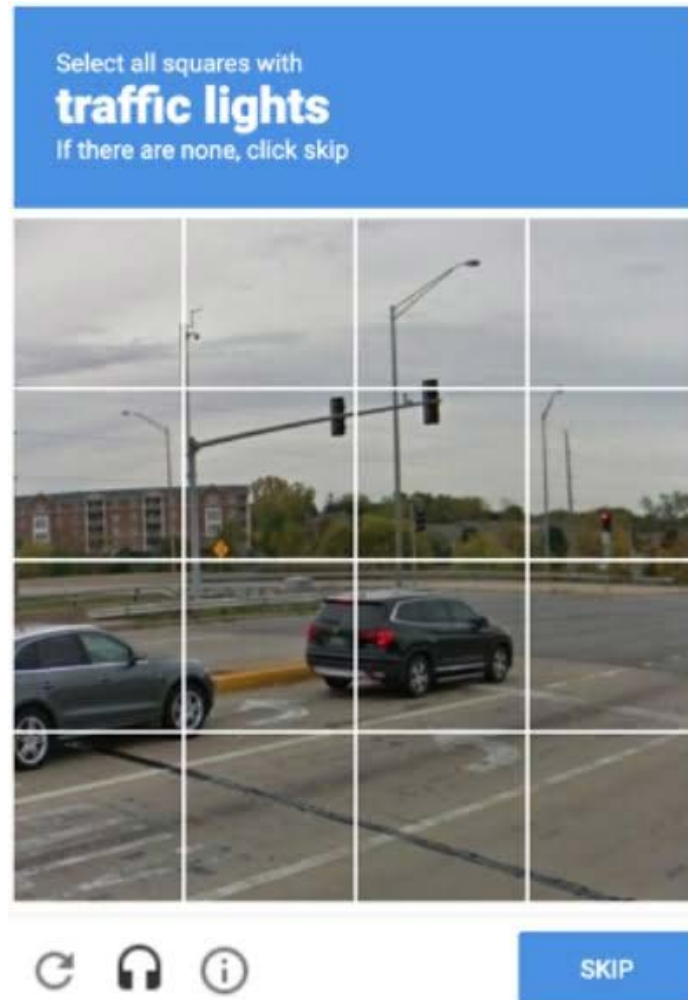
Challenges in Data Annotation

- Time-Consuming: Large datasets require extensive effort.
- Resource Intensive: Needs significant human and technical resources.
- Ensuring Consistency: Maintaining uniform quality across data is challenging.

Overcoming Challenges

- Use Experts: Outsource to professional annotation services.
- Employ Advanced Tools: Utilize tools that automate and streamline the process.
- Regular Quality Reviews: Consistently check and validate annotations.

Captcha



Create your own Dataset

- Custom training sets for machine learning can be created by:
 - Collecting data using web scrapping software tools
 - Cameras and other devices with a sensor
- Third-party dataset service providers can also help with data collection for machine learning tasks.
- You need to annotate the data and label the individual data points appropriately so that it fits well with the type of AI algorithms that you intend to use.

Image Scrapping

- Image scraping is the automated process of extracting and downloading image files from websites using specialized software or code.
- How does it work?
 - ✓ Identify target sources
 - ✓ Extract image URLs
 - ✓ Download images

Image Annotation

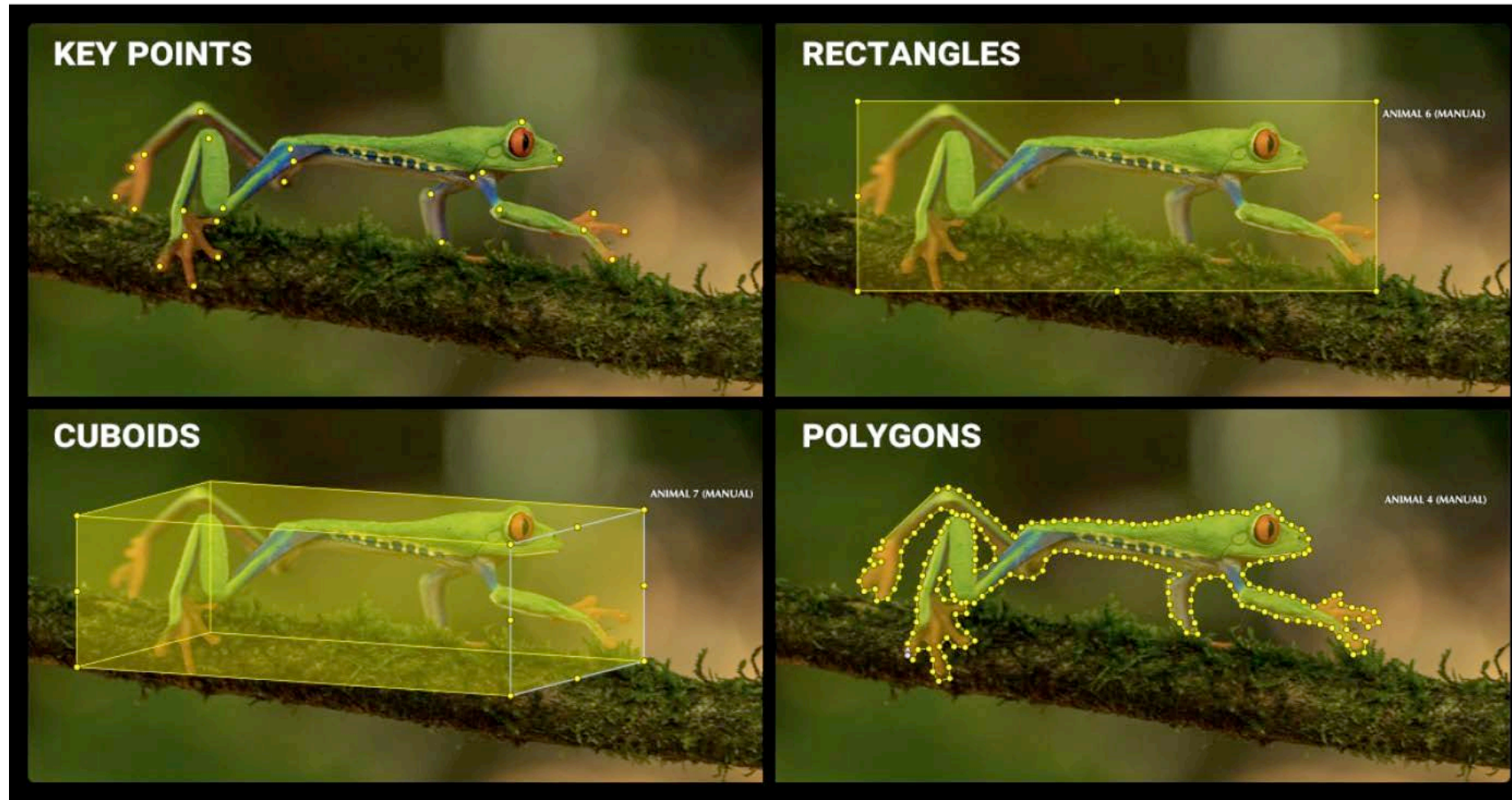


Image Annotation – Common Tasks

- Keypoints: Best for motion tracking and identifying specific points on objects.
- Rectangles (Bounding Boxes): used for object detection by drawing boxes around objects.
- Polygons: Capture precise shapes and boundaries of objects.
- Cuboids (3D Boxes): Annotate objects in three dimensions.

Data Annotation Process



Labelling and Annotation Tools

The common image annotators

Annotator help to prepare dataset for working with training computer vision algorithm.

Annotator	Annotation type	URL
COCO Annotator	Bounding box, polygon, point, freehand	https://github.com/jsbroks/coco-annotator
CVAT	Bounding box, polygon, polyline, points	https://github.com/opencv/cvat
ImageTagger	Bounding box, polygon, line, point	https://github.com/bit-bots/imagetagger
imglab	Bounding box, polygon, circle, ellipse, point	https://github.com/NaturalIntelligence/imglab
labelImg	Bounding box	https://github.com/tzutalin/labelImg
LabelMe	Bounding box, polygon, circle, line, point	https://github.com/wkentaro/labelme
OpenLabeling	Bounding box	https://github.com/Cartucho/OpenLabeling
Yolo_mark	Bounding box	https://github.com/AlexeyAB/Yolo_mark
VIA	Bounding box, polygon, circle, ellipse, polyline, point	http://www.robots.ox.ac.uk/~vgg/software/via/
VoTT	Bounding box, polygon	https://github.com/microsoft/VoTT

