A Data Annotation Process for Human Activity Recognition in Public Places

Mickael Cormier

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
mickael.cormier@kit.edu

Abstract

Behavior analysis of individuals in crowds or groups of people in public places through surveillance cameras gains importance for several different actors. Automatically detecting and understanding pedestrians in real-world uncooperative scenarios is very challenging. Common issues such as limited annotated data, unreliable data and annotation quality, and appropriate use of this data for supervised learning often originate in steps preceding the modeling of specialized neural network architectures. In this report, the necessity and requirements for designing a reliable data annotation process are presented. Some precise ideas for automation through neural networks are discussed in a conceptual manner.

1 Introduction

Automated analysis of video content through deep learning algorithms offers tremendous potential for autonomous driving, health care, agriculture, surveillance for both home and public places. In the last years most annotated datasets required by such algorithms have been labeled manually. ImageNet [6], the first major publicly available large scale dataset was published in 2009 with more than 14 millions hand-annotated images, for which 20,000 to 30,000 people a year worked for several years through crowd-sourcing [14]. Systems which could potentially threaten human lives, such as assisting systems for surgery or autonomous driving require reliable, high quality data, with strict and explainable validation measures. Furthermore the data need to cover rare events and scenarios which could represent critical safety issues. For instance, to ensure the correct functioning of a sensor system for autonomous driving, it must be demonstrated over a distance of 300,000 km within a given scenario, meaning over 240 millions frames and 3.6 billions objects are to be annotated accurately [20]. Considering an average of 60 seconds for the thorough annotation of an object, the sole creation of a single validation dataset for a given scenario represents a several-year project with hundreds of full-time annotators.

Similarly, safety related surveillance projects not only require detection of people but also an estimation of their body posture as well as their activity. Given a public place covered by several surveillance cameras, scenes of interest need to be annotated frame by frame, with an average of 50 persons in sequences recorded with 30 frames per seconds for several minutes. Considering several camera clusters, manual annotation appears almost not affordable regarding the rigorous and transparent validation of the system which are required to fulfill legal and ethical requirements. Therefore, the average annotation time per object must be reduced. To this aim a highly automated data annotation process and data system is enquired.

The remainder of this report is oriented toward the design of such a process and, thus, organized as follows: the concept of data annotation and automation are presented in Section 2, requirements of a reliable annotation process are introduced in Section 3, while the steps of the data annotation process and the data quality assurance are discussed in Section 3.2 and Section 4 respectively. A conclusion is given in Section 5.

Figure 2.1: Manual keypoint and bounding box annotations for a crowded scene in [5].

2 Data Annotation and Automation

In this work we consider the annotation process for human activity recognition in public places such as a public transportation hub. Cameras are placed strategically in order to monitor traffic, assure the traveler's safety and security. Therefore, large areas coverage using clusters of cameras with wide field of views offering multiple views of given hotspots is required. Thus, data for raw scenes including multiple views and dozens up to larger crowds of hundreds of pedestrians result from this setup, as illustrated in 2.1.

An important part of use cases in the field of human activity recognition is based on supervised learning, which means training data with target annotation is required in order to update a prediction model. The process of collecting annotation is often much more expensive than the process of collecting the data itself. This annotated data, the ground truth, is however limited by the expertise of the person annotating. Depending on the complexity of the task and its specifications, a label may require on the one hand pixel-wise precision, e.g. semantic segmentation, keypoint detection, precise 2d and 3d bounding boxes. On the other hand, tasks such as person tracking, re-identification and temporal action localization in video require a more profound understanding of the scene developing. Furthermore, several different annotation formats may internally define the same type of annotation differently [13, 2, 8, 10, 22, 4, 17] resulting in a supplementary layer of complexity for the annotator, if not intuitive or usual. For instance a 2d bounding box in COCO [13] is defined as a (x-top left, y-top left, width, height) tuple, while a Pascal VOC [8] bounding box is defined by the tuple (x-top left, y-top left,x-bottom right, y-bottom right) and YOLO [19] defines a bounding box by it's center. Annotation software such as [22, 8, 3] support multiple formats and offer different tools in order to partly provide automation for specific tasks. However, those are not specifically designed for multiple-views use cases and provide automation only to a limited level.

In the following, we formally define different levels of automation for multiple-view data annotation for human activity recognition in public places.

- **Level 0: No Automation**
  The whole annotation process is done manually. The annotation tools provide simple functions to produce annotation.
- **Level 1: Tool Assistance**
  The annotator is assisted by different tools which minimizes the effort of annotating video frames. Provided multiple views of a the same place, the annotator may switch between views of a sequence while annotating. The annotations are converted to the views of this sequence. Furthermore, given a partly annotated sequence, the tool is able to interpolate the movement of the objects into subsequent frames.
- **Level 2: Partly Automated Annotation**
  Using a point, scribbles or a polygon the annotator defines a region of interest within a frame, the desired annotations are returned by the tool and manually corrected if necessary.
- **Level 3: Highly Automated Annotation**
  The whole (multi-view) sequence is automatically annotated for the chosen

type of labels. The human annotator mainly conducts quality checks and corrections.

- **Level 4: Highly Automated Annotation Pre-Checking**
  The whole (multi-view) sequence is automatically annotated for the chosen type of labels. Quality checks are automatically generated for a human validator to review and extend.

Data annotation is a complex process which comes at a great cost. Designing specific tools which enable automation is an important step in order to produce data annotation at scale. Nevertheless, the quality of the data produced requires constant control and validation. Therefore, a transparent and reliable annotation process is necessary.

3 Designing a Reliable Annotation Process

A reliable annotation process requires clear steps for which different actors, e.g. person or entity, have well defined responsibilities and are held accountable. In this section, a reliable annotation process for human activity recognition in public places is described. The overall process is illustrated in a activity-based flowchart diagram in Figure 3.1. A recent white-paper [7] published during the writing of this work presents similar views regarding an abstract annotation process for supervised learning. In contrast, this report focuses on the concrete case of multi-view annotation processes of crowds and discusses concepts regarding data privacy and ethics. Furthermore, concrete strategies for the use of automatic annotation proposals are defined.

3.1 Participants in the Annotation Process

We identify four entities as participants in the annotation process and define those as follow.

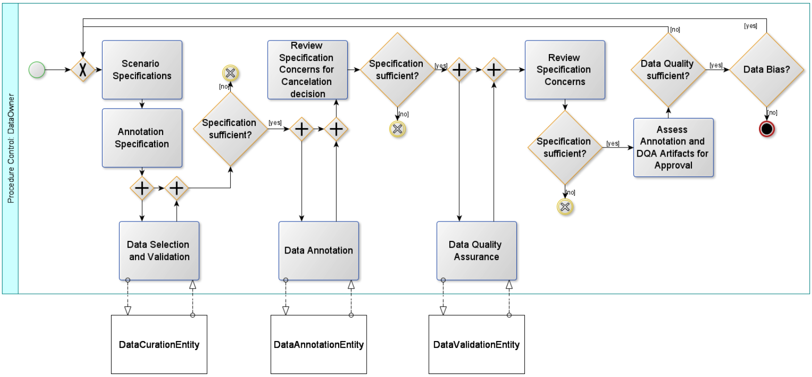


Figure 3.1: Annotation Process Diagram

3.1.1 Data Owner Entity

Ahead of the annotation process the data owner entity is in charge of the data acquisition and identification for the use case, as well as cleaning and preprocessing. The data owner is responsible for the data and the complete annotation process at any time of the process. This entity is in charge of initiating sequence annotation and validation, and finally approve or cancel the result of the process. They are responsible for formulating scenario specifications, quality requirements and identifying scenarios risks and requirements such as anonymization, potential bias, legal issues or underrepresented aspects.

3.1.2 Data Curation Entity

The data curation entity is responsible for assuring compatibility of the data with laws, ethic, potential bias and anonymization requirements. This entity is in charge of analyzing the provided data and scenario specifications, formulate necessary data curation and transformation in order to meet specifications. They extend the scenario specifications with concrete fail condition regarding annotations, such as label balancing or systematic bias.

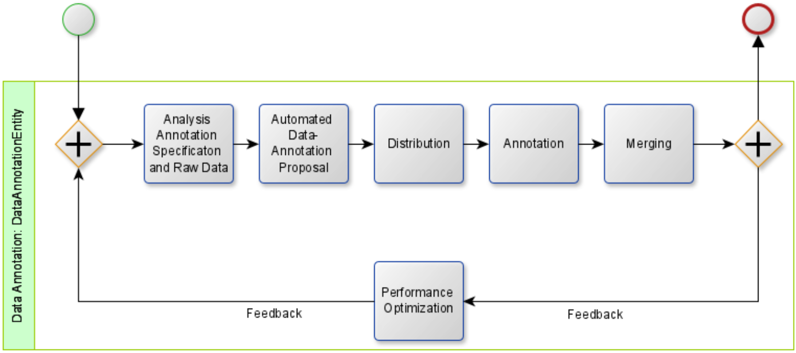


Figure 3.2: Data Annotation Process Diagram

3.1.3 Data Annotation Entity

The data annotation entity produces the annotation and report potential issues either to the curating entity or the data owner. This entity delivers the annotations to the data owner.

3.1.4 Data Validation Entity

The data validation entity fulfills quality assessment of the annotations and reports its results to the data owner or if necessary to the data curation entity.

3.2 Steps of the Annotation Process

In the following the different steps of the annotation process starting from scenario specification to data splitting for training and validating are described. The preceding steps of physical data acquisition and preprocessing are not further discussed here, since those are not directly part of the annotation process itself.

3.2.1 Scenario Specifications

The data owner entity is responsible for the dataset and the annotation project and thus is required to analyse the use case to be covered, its environmental and technical conditions as well as their implication for the machine learning specifications.

Firstly, the conditions of the environment such as lighting, surrounding area variety, weather, time of day of the target scenario have to be defined and openly compared to the provided raw data. The available data should represent the target use case. In case of potential parts missing from the use case, decisions about the addition of supplementary real or synthetic data are required.

Secondly, the technical conditions such as media format, sampling rate, resolution, camera position, potential camera cluster focusing on different places and their implication are analyzed. The aim is to comprehend the feasibility of particular annotation types on specific views, e.g. not enough pixels of the person of interest are provided in order to estimate their pose, too much natural occlusion through fixed construction or vegetation. In case of a camera cluster, it is essential to determine whether the cameras are calibrated and their parameters are known in order to reconstruct 3d representations or convert annotations from one view to another.

Thirdly, while working with data focusing on person in public places important issues emerge concerning privacy, currently applicable laws as well as ethical considerations. Whereas deployed models probably will work on a raw data feed, training and validation data should be anonymized and bio-metrical information of person shouldn't be used implicitly, if not targeting a specific problem related to bio-metrics. Furthermore, the anonymized person should be considered as an avatar of a real person and shouldn't provide unattended long-time information on real-person, e.g. a person may be re-identified in a different view of the same scene whithin the same camera cluster, however this person shouldn't be re-identified in another cluster or in later recordings. Besides data privacy specifications, the data owner entity is expected to address potential bias and ethical issues and proceed carefully during data selection. It is proven that non representative data produce biased results which may impair the quality of the model or, worse, accentuate social inequalities and present bias against ones ethnicity or gender [1, 15, 18]. Those issues are not only dangerous and noxious, but also damaging to the trustworthiness of the model [11].

Finally the annotation task to be fulfilled is described along it's aim. The intended effect is to increase the implication and engagement of the labeling entity to the task ahead. For instance, the annotation of keypoints for a person may imply that the model to be trained will perform human pose estimation. In this case, the temporal margin of error regarding specific keypoints offers a greater tolerance. However considering pose based action recognition, the lack of precision over time may cause spatial noise over time, which in turn impacts the results of the model.

3.2.2 Annotation Specifications

After analysis of the raw data and the scenario specifications, the data owner entity is expected to define clear and unambiguous labeling specifications regarding which kind of annotation is to be produced, e.g. person detection, instance segmentation, pose estimation. Quality tolerances are precisely defined, e.g. size of bounding box, margin of error tolerated. In dialog with the labeling entity, concrete annotation instructions are derived from these general specifications and must prevent potential issues. Those instructions aim to clearly define solutions for edge cases, clarify unspecific labels and identify potential conflicting instructions. Furthermore, they represent the basis for the annotation validation process and may support the initial configuration of annotation proposal tools. Finally, the specifications and instructions support the potential adaptation of annotation tools if required.

At any time in this step specification concerns may be raised which need to be immediately reviewed, therefore halting the whole process. If the specifications reveal weakness, the process should be aborted and the specifications fixed before starting a new iteration. The earlier such issues are identified, the more efficient the whole process becomes. Despite potential concerns for slowing the process down in its early stages, openly assessing annotation specification is cost efficient and prevents repeated annotation of the same sequences.

3.2.3 Data Selection and Verification for Annotation

The selection of representative data for annotation is carried out according to the annotation specifications. On one hand, depending on the target scenario, the available data is usually subject to subsampling: Spatial sampling, i.e. limit the number of sample per spatial region or completely reject specific spatial regions due to annotation concerns, and/or temporal subsampling, i.e. extract every n-th frame.

On the other hand, a common scenario in the active learning literature represents this data selection step perfectly, which is the unlabeled pool scenario. Considering the yet to be annotated raw data, a round base game is defined. In every round and active learning model ranks the data in the unlabeled pool, and the k bests are selected for annotation, given a fix annotation budget. The selected data is annotated and added to the training set of the active learning model, which is re-trained on the new dataset. These rounds are repeated until the annotation budget is fully drained. Recent methods for CNN not only select the best data for annotation based on efficiency and diversity, but also consider that model training mainly uses batches of data [21, 12, 23].

Both aspects could also be sequentially combined. However, even careful data selection may lead involuntarily to strong biases. This is where the data curation entity is required to formally analyze the selected data and may raise empiric concerns, which need to be reviewed and validated at the end of each annotation process.

3.2.4 Coordination and Distribution

Depending on the volume of selected data, annotation specification, personal availability and prior experiences, the data annotation entity decides on the data and/or label slicing and form of distribution. Overlapping subsets should always be considered for validating individual accuracy as well as for training purposes for new annotators. Different slicing options depending on the data are available. For instance, given multiple labels, it is possible to distribute the target labels between available annotators, e.g. pedestrians, babys, objects and cars for instance segmentation. For motion tracking, sequences could be distributed per camera, cluster of cameras for scenes. Overall for a dataset composed of multi-view sequences, these may be split in subsequences along the temporal axis and / or the spatial axis. Single subsequences might be shared between the annotator for validation purposes and / or for particular crowded sequences, or different labeling tasks. Eventually, those subsequences are required to be merged in order to finalize the annotation process. In this case merging heuristics are required, e.g. voting, average of all annotation, cherry picking. The merging step may raise issues concerning the annotation specifications, and therefore requiring a new annotation iteration to correct these issues.

3.2.5 Annotation

The data annotation entity performs the labeling and delivers the resulting annotations as illustrated in Figure 3.2. They raise issues immediately during the process and provide feedback on the tool at their disposal for the task.

3.3 Use of Pre-Annotation

Considering the annotation of human poses for crowds in public places with multiple cameras, the annotation effort required increases exponentially. For this reason, the annotation entity relay strongly on automated tools to improve and accelerate the annotation process. As shown in Figure 3.3, efficient tools may improve the annotation greatly.

3.3.1 Use of existing Annotations

Sometimes part of the dataset of an annotation project had been annotated earlier. In this case automated import tools are required to import and reuse these preexisting annotations. However, these annotation are considered (human-) generated pre-annotations pending for review, since they probably do not fully comply the annotation specifications.

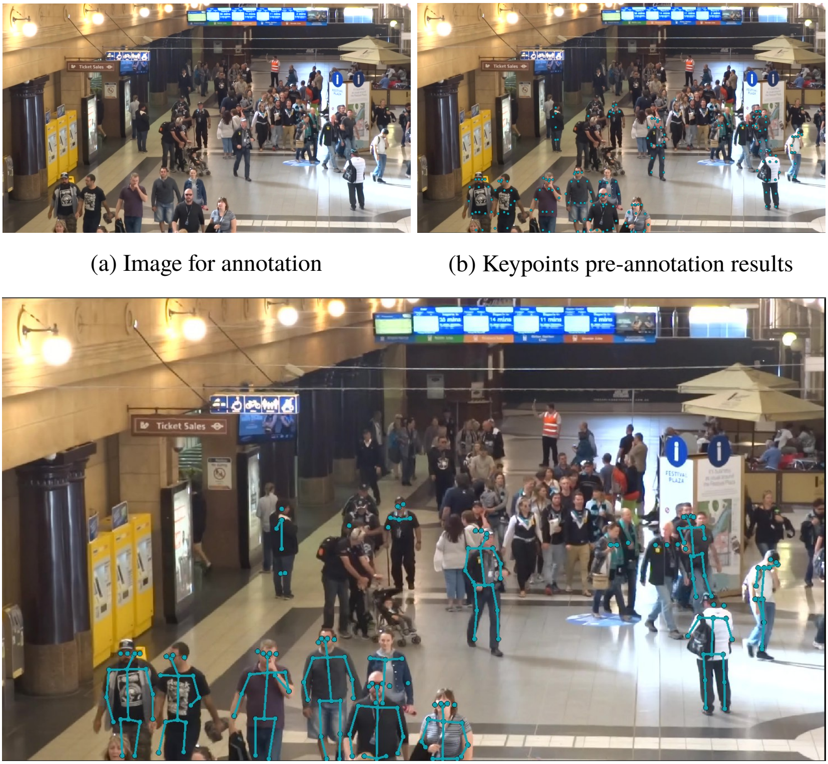
(a) Input Bild

(b) Pre-Annotations

Figure 3.3: Pre-Annotations for the domain application greatly improve the quality and speed of the annotations. In this case most of the individuals are already annotated.

3.3.2 Transfer and Curriculum Learning

The annotation entity may use pre-annotation tools to generate annotation proposals. However, since at the beginning of the annotation process none or only scarce annotations are available, the tool requires prior training on similar data. If annotations are available, the tool can be fine-tuned on the target domain using transfer learning. In the case when no annotations are available, the tool can be used to generate a prliminary annotation proposal which then will be reviewed manually and progressively be retrained on the new available data. For instance, we consider the task of annotating the pose of all pedestrians in the dataset illustrated in Figure 2.1. At first, the annotation proposal tool has only been trained on COCO, an other target domain shown in Figure 3.3(a), where the ratio pixels per person is much higher. As expected, the results in Figure 3.4(b) and Figure 3.4(c) are acceptable on the first row. However, the persons behind them aren't recognized at all. After some frames have been manually annotated the pre-annotation tool can be trained using these annotations to create better key-point predictions. Ideally images of a lower difficulty level are first annotated and used for training. The difficulty is then progressively increased in accordance with the performance limits of the actual



(a) Image for annotation

(b) Keypoints pre-annotation results

(c) Keypoints pre-annotation results as skeleton representation

Figure 3.4: Keypoints pre-Annotation on MOT20 [5] using a pre-annotation tool trained on COCO [13]. Only the first row is detected which corresponds to the training domain.

instance of the tool. After a few iterations, this curriculum learning approach greatly reduces the effort for long and / or similar sequences.

3.3.3 Online Learning

Theoretically, considering the annotation process as the training phase of a model. Annotation proposals would represent a training iteration. The manually corrected annotation is thus the ground truth. Therefore, the correction itself, e.g. the distance between the location of a proposed bounding box and the corrected one can be used to calculate an error which in turns may be used to update the parameter of the model. Consequently the resulting online learning enables immediate performance increase, without requiring complete retraining on the model and a long wait before the new deployment of the tool. Furthermore, the learned model should already conform to the annotation requirements.

3.3.4 Continuous Learning

Considering long time annotation projects with potentially continuous flows of new data to be annotated, different strategies may be required. Given a satisfactory annotation proposal tool trained with curriculum learning and / or online learning, we define an annotated subset, which is manually approved, as a validation set. The tool automatically annotates new incoming data and is periodically retrained over the ever growing dataset. The non-changing validation subset is then used to regularly assess the performance of the tool and thus prevent catastrophic forgetting or a negative feedback loop.

3.4 Discussion

A reliable, efficient and therefore highly automated annotation process is a complex and difficult process to model with need of constant improvement. Several interests and requirements are to be acknowledged and dealt with. Nevertheless, we identified great sources of improvement which can be addressed with efficient automation covering several aspects of machine learning which are currently topics of active and ever evolving research.

4 Data Quality Assurance

The annotation of data is a long, complex and repetitive process which requires intense concentration. Multiple potential issues may appear during annotation. The annotation may be incomplete, e.g. missing frames in a sequence or classes of objects, or annotations in a wrong format, e.g. the pose has been annotated with 14 keypoints, but 17 were required. Different mistakes can be made such as misclassification, false positive, false negative or incorrect re-idenfication. Furthermore, annotations may lack precision, e.g. a keypoint lies a few pixels besides the intended point, a bounding box excludes the feet of a person. Throughout a longer sequence annotations maybe inconsistent, e.g. bounding box around the full body of a person instead of placing the bounding box around the visible body. Nevertheless, such issues are reliably detectable and therefore it is possible to address them within a short time. The data validation entity is responsible for methodically finding and reporting these issues. Following the annotation specifications, they perform a quality check on the annotation delivered by the data annotation entity. They provide detailed feedback on possible annotation issues to the data annotation entity and raise specification issues immediately. Specialized tools for review are used, thus the data validation entity itself is not permitted to perform the correction of the annotations.

In the remainder of this section the steps of the data quality assurance process is shortly reviewed, then the final step of data selection and validation is described.

4.1 Steps of the Data Quality Assurance Process

The steps of the data quality assurance process, as illustrated in Figure 4.1, are principally identical to the annotation steps. First the data validation entity analyzes the annotation specification and may use automation for annotation pre-checking. Then the validation task is distributed along the available personnel, performed and finally merged. Lastly the feedback and assigned issues, the validation artifacts, are reported to the data owner entity.

4.2 Data Selection and Validation

After ensuring sufficient quality of annotations, the dataset is split into three parts (with a possible fourth part of non assigned data): a training, validation and test subset.

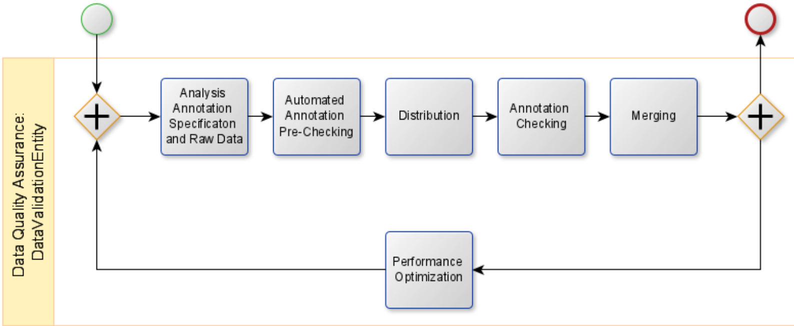


Figure 4.1: Data Quality Assurance Process Diagram

4.2.1 Training Subset

This subset is used to train the learned parameters for a supervised model. It should be carefully selected according to target use case and the scenario specifications, e.g. edge cases, ethic, law. The dataset should be representative, hence severe imbalance may result in discarding part of the annotated data.

4.2.2 Validation Subset

The validation subset is distinct from the training data. It is used to measure the performance of models during prototyping and training. It should reflect edge cases and offer sufficient diversity in regard to the target use case.

4.2.3 Test Subset

The test subset is used to measure the performance of the model after training in order to detect potential overfitting against the validation data. Therefore, the test subset is distinct from the training and validation subsets. Furthermore, it is used to asses the performance of a model against the target use case.

Each subset should be subject to review from the data curation entity against the target use case and scenario specifications. Finally, the whole process, intents and specifications should be rigorously documented to facilitate intern and extern audits, e.g. model cards [16] and datasheets [9].

5 Conclusion

Neural network aided data pre-annotation is a very promising approach. Yet not new. It is well-known for person detection or pose estimation in surveillance scenarios that specific problems such as false classification, person occlusion, split or merged detection often result from automatic predictions. A reliable and partly neural network aided annotation process should technically profit from human intelligence through specialized human operators and improve their abilities. Several methods were presented and discussed in this report with the focus on human activity in public place surveillance. Furthermore, concepts were presented for a reliable and transparent data annotation process, which naturally includes computer assisted steps through tailored neural networks. Future work will include the implementation and evaluation of these methods and also further investigations on a reliable and efficient annotation process for scenes including multiple cameras oriented on one and a same place.

References

[1] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on fairness, accountability and transparency*. 2018, pp. 77–91.

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. "COCO-Stuff: Thing and stuff classes in context". In: *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE. 2018.

[3] *Computer Vision Annotation Tool (CVAT)*. https://github.com/openvinotoolkit/cvat/.

[4] Marius Cordts et al. “The Cityscapes Dataset for Semantic Urban Scene Understanding”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[5] P. Dendorfer et al. “MOT20: A benchmark for multi object tracking in crowded scenes”. In: *arXiv:2003.09003[cs]* (Mar. 2020). arXiv: 2003.09003. URL: http://arxiv.org/abs/1906.04567.

[6] J. Deng et al. “ImageNet: A Large-Scale Hierarchical Image Database”. In: *CVPR09*. 2009.

[7] Matthis Eicher et al. *Whitepaper Process considerations: A reliable AI data labeling process*. Aug. 2020. URL: https://wiki.eclipse.org/images/0/0e/WhitePaper_Process_considerations-A_reliable_AI_data_labeling_process.pdf.

[8] M. Everingham et al. “The Pascal Visual Object Classes (VOC) Challenge”. In: *International Journal of Computer Vision* 88.2 (June 2010), pp. 303–338.

[9] Timnit Gebru et al. “Datasheets for datasets”. In: *arXiv:1803.09010* (2018).

[10] Andreas Geiger et al. “Vision meets Robotics: The KITTI Dataset”. In: *International Journal of Robotics Research (IJRR)* (2013).

[11] HEG-KI. *Ethics guidelines for trustworthy AI*. 2019. URL: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[12] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. “Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning”. In: *Advances in neural information processing systems*. 2019, pp. 7026–7037.

[13] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context”. In: *CoRR* abs/1405.0312 (2014). arXiv: 1405.0312. URL: http://arxiv.org/abs/1405.0312.

[14] John Markoff. *Seeking a Better Way to Find Web Images*. Nov. 2012. URL: https://www.nytimes.com/2012/11/20/science/for-web-images-creating-new-technology-to-seek-and-find.html.

[15] Margaret Mitchell et al. “Diversity and inclusion metrics in subset selection”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 117–123.

[16] Margaret Mitchell et al. “Model cards for model reporting”. In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 220–229.

[17] Gerhard Neuhold et al. “The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes”. In: *International Conference on Computer Vision (ICCV)*. 2017. URL: https://www.mapillary.com/dataset/vistas.

[18] Inioluwa Deborah Raji et al. “Saving face: Investigating the ethical concerns of facial recognition auditing”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 145–151.

[19] Joseph Redmon et al. “You Only Look Once: Unified, Real-Time Object Detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.

[20] Daniel Rödler. *A Feast of Sensor Data: Feeding Self-driving Algorithms*. Nov. 2020. URL: https://understand.ai/blog/annotation/autonomous-driving/machine-learning/2020/11/12/feast-of-sensor-data-feeding-self-driving-algorithm.html.

[21] Ozan Sener and Silvio Savarese. “Active Learning for Convolutional Neural Networks: A Core-Set Approach”. In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=H1aIuk-RW.

[22] *Sloth*. https://github.com/cvhciKIT/sloth/.

[23] Fedor Zhdanov. *Diverse mini-batch Active Learning*. 2019. arXiv: 1901.05954 [cs.LG].