Qually: A Quality Validation Toolbox for Automotive Perception Data Towards Trustworthy AI

Public report

A data quality validation toolbox for AI

Author: Yinan Yu, Samuel Scheidegger, Jörg Bakker

Date: 2022-08-30

Project within: FFI, Traffic safety and automated vehicles

Contents

1 Sammanfattning på Svenska

Datadrivna tekniker som artificiell intelligens (AI) och djup maskininlärning är ofta förekommande som en del av perceptionssystem för fordon. Då dessa system är starkt beroende av data är datakvalitet av yttersta vikt. I ett perceptionssystem för fordon samlas data in av sensorer och transformeras till olika format beroende på var i AI-systemets pipeline för databehandlingen den befinner sig. Även om data i olika stadier delar liknande attribut, skiljer sig effekterna av deras egenskaper i varje enskilt steg väsentligt från varandra. Därför måste krav på datakvalitet definieras specifikt för varje steg.

I det här projektet är målet att utveckla ett verktyg för end-to-end kvalitetskontroll för att upptäcka fel och anomalier genom hela AI-pipelinen. För att uppnå detta mål delar vi upp projekt i tre arbetspaket, där det första steget är att designa en uppsättning dataegenskaper och deras motsvarande krav som kvalitetsspecifikationer för data i varje steg. Som ett andra steg, givet dessa specifikationer, har vi utvecklat en verktygslåda, Qually, för att utvärdera datakvalitetsmått och upptäcka fel och anomalier i hela AI-pipelinen. I det sista arbetspaket, som en demonstrator, används Qually för att förbättra automatiserade dataannoteringar. Detta implementeras i tre steg: 1) fel identifieras av Qually med hjälp av de kvalitetsmåtten; 2) Qually föreslår en automatisk korrigering med ensembletekniker; 3) den korrigerade annoteringen utvärderas av Qually för att bekräfta förbättringen i kvalitet. Feldetekteringen och föreslagna korrigeringar inspekteras manuellt för att statistiskt validera resultatet av Qually.

Som nästa steg, förutom vidareutveckling av Qually som mjukvara för att förbättra dess robusthet, kapacitet, skalbarhet och fullständighet, planerar vi att fokusera på att utöka uppsättningen av dataegenskaper och kvalitetsspecifikationer, särskilt genom att inkludera tekniska och affärsmässiga krav från olika fordonsintressenter. Vi planerar också att undersöka möjligheten och skalbarheten av att integrera formella verifieringstekniker för kvalitetskontroll.

2 Executive summary

Data-driven techniques such as artificial intelligence (AI) and deep learning are frequently deployed as part of automotive perception systems. Due to their heavy dependency on data, data quality is at the essence. In particular, in an automotive perception system, data is captured by sensors and transformed into different formats depending on where it is in the AI data processing pipeline. Although data at different stages share similar attributes, the impact of their properties at each individual stage differ significantly from one another. Therefore, data quality requirements need to be defined specifically at each stage.

In this project, the objective is to develop an end-to-end quality control toolbox to detect errors and anomalies throughout the entire pipeline. To achieve this objective, we divide the

project into three work packages, where the first step is to design a set of data properties and their corresponding requirements as quality specifications for data at each stage. Given these specifications, as a second step, we have developed a toolbox, Qually, to evaluate data quality metrics and detect errors and anomalies throughout the AI pipeline. In the last work package, as a demonstrator, Qually is applied to improve automated annotations. This is implemented in three steps: 1) errors are identified using the quality metrics evaluated by Qually; 2) Qually suggests an automatic correction using ensemble techniques; 3) the corrected annotations are evaluated by Qually to confirm the improvement in quality. The error detection and suggested corrections are manually inspected to statistically validate the outcome of Qually.

As the next step, besides further developing Qually as a software to improve its robustness, capacity, scalability and completeness, we plan to focus on enriching the set of data properties and quality specifications, especially by including technical and business requirements from various automotive stakeholders. We also plan to investigate the possibility and scalability of integrating formal verification techniques for quality control.

3 Background

Artificial Intelligence (AI) and deep learning has become ubiquitous in automotive perception systems over the last decade. The performance of deep learning algorithms has surpassed traditional techniques in many aspects, especially for computer vision based systems. However, the impressive capacity of deep learning is a double-edged sword. The interpretability is compromised due to the high complexity of these algorithms. In practice, one frequent question we encounter is "how do you know when an AI system would fail?" and a subsequent question is "why does it fail in this scenario?" They are especially important for automotive applications since safety and lives are at stake.

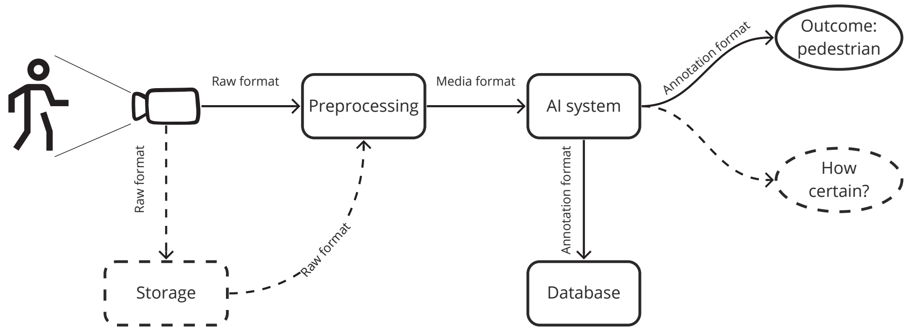


Figure 1: The deployment pipeline.

To investigate the answers to these questions, one needs to trace back to their origin - data. As data-driven techniques, properties and the quality of data plays a crucial role in an

AI system. To be able to better understand and predict failures of AI systems, in this project, we first characterize data using its *format* in the pipeline depending on the interface and information it carries. We categorize data into four data formats: raw format, media format, annotation format and meta format. The raw format refers to the raw log data format that carries compact binary information. The media format is the most familiar human readable format such as PNG for images and PCD for 3D point clouds. The annotation format refers to the schema of data annotations, such as bounding boxes for object detection. AI system predictions also share this data format. The meta format contains meta information such as ownership and statistics of other data. At each stage throughout the AI pipeline, data is transformed into a specific format, consumed and passed on to the next building block.

To illustrate this transformation process, we show two types of AI pipelines, deployment and development, in Fig. 1 and Fig. 2, respectively. The deployment pipeline is the installation of the AI system, where its predictions are being consumed by the end users, whereas the development pipeline refers to the training and validation processes during AI system development.

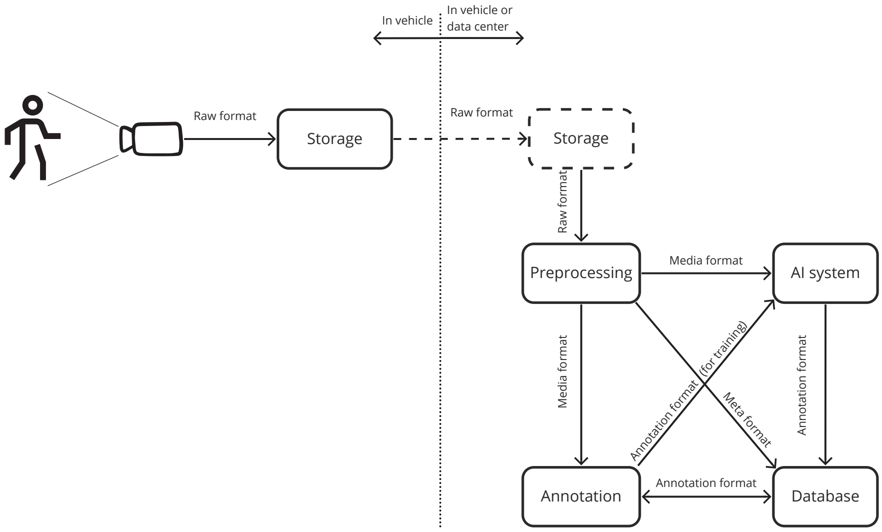


Figure 2: The development pipeline.

When undetected errors are present at one stage, they will propagate to the subsequent modules in the pipeline. In this project, we aim to define and detect such abnormalities as early as possible to avoid potentially catastrophic system failures cause by errors in data.

4 Purpose, research questions and method

The purpose of this project is to enable anomaly and error detection at each stage in the data pipeline to eliminate the propagation of errors that cause unwanted behaviors. Data quality control and anomaly detection is important for improving system interpretability and transparency.

Concretely, we focus on the following research questions in this project:

RQ1. How can we systematically define and quantify data quality in the context of automotive perception systems?

RQ2. What types of noise and errors are there in data throughout the pipeline?

RQ3. Can we intervene and correct these errors to improve data quality?

4.1 Data description

4.1.1 Sensors

In this study, the system is developed based on data collected using the following sensors:

- LiDAR: An active optical sensor that uses laser range measurements to construct a 3D point cloud of the surrounding environment.
- Camera: A passive optical sensor that captures light from the surroundings to create a 2D image.
- IMU: A set of electromechanical sensors that are typically able to measure the direction, angular speed and acceleration of the body that it is attached to.
- Wheel speed sensor: A sensor attached to the wheels of a vehicle to measure the speed.
- GNSS sensor: A satellite based positioning system that outputs a global position, speed and heading.
- Radar sensor: An active electromagnetic sensor measuring the reflectivity of the surroundings and outputting the angle (azimuth, elevation), range and relative speed (Doppler) to the reflection.

The main data collection activity was funded by a previous FFI project in collaboration with Chalmers Resource for Vehicle Research (the REVERE lab).

4.1.2 Formats

A *data format* refers to the abstraction we work with when defining and measuring data quality. Briefly speaking, a format reflects the processing stage in the data pipeline. Each format has its own properties and quality requirements.

Raw format The raw data format refers to the container format used by the data logger. This is the most compact data format in the data pipeline. Briefly speaking, information is captured by the sensor, encoded as binary streams and sent with protocols such as TCP/IP over Ethernet networks. One data log typically contains multiplexed data streams from asynchronous sensors. For this data format, we may encounter errors such as stream corruption due to sensor malfunctioning, packet loss, incorrect timestamping, etc.

Media format This format refers to data that computer programs can read and render for humans, which makes it the most frequently used data format by engineers, data scientists and product stakeholders. This format contains, for example, PNG, JPEG, MP4, PCD, etc. Quality measures for this format includes errors and noises that hinder human interpretation or algorithm accuracy. Examples of such obstruction are image corruption, overexposure, low sensor resolution, etc.

Due to distinct behaviors, data in the media format is divided into two categories: data frame and sequential measurements.

- A *data frame* refers to one measurement unit around one time instance. It can be further categorized into two different types.
  - Single sensor: Measurements that are captured and grouped as one entity in the data stream, e.g. an image captured by a 2D camera sensor;
  - Sensor fusion: Measurements captured by multiple sensors that are grouped around one timestamp;
- *Sequential measurements* contain continuous recordings, e.g., a video sequence.

Annotation format There are two types of data in this format.

- Annotations used for AI system development;
- Predictive outcomes from a deployed AI system.

Similar to the media format, in this project, we split the annotation format into two corresponding categories:

- Frame-based: Annotations attached to one measurement unit (e.g., one image), for example, image classification, 2D and 3D bounding box detection, object instance segmentation, semantic segmentation, etc.
- Sequence-based: Annotations attached to a sequence of measurements (e.g., a video sequence), for example, object dynamics (e.g., tracking each object within the sensors' field-of-view), ego dynamics (e.g., the ego motion and inertial measurements), etc.

Meta format This is the format that stores the meta data and statistics, such as distributions of pixel values in an image. It can also include statistics of the annotation format, such as class labels, object sizes, etc. Errors that arise in this format are primarily programmatic since they are meta data acquired or computed from other formats. Moreover, data coverage plays a crucial role in measuring the quality of the meta format.

4.2 Quality measures

There are many ways to define data quality. In this project, we choose one of the most commonly applied definitions with the following six dimensions: consistency, accuracy, completeness, validity, uniqueness and timeliness. This set of quality measures are often applied in the context of business data governance. We find it applicable to data quality of automotive perception systems. In this section, we give a brief description of each data quality measure to illustrate the concept.

Consistency Is there a contradiction or conflict in data? To test consistency, one needs to define data logic and specifications within each format and cross validate between different formats.

Accuracy Does data represent reality accurately? This quality is validated by comparing data to sources of ground truth. When ground truth is not available, expected values are synthesized to estimate data accuracy.

Completeness Are there missing values in the data collection? In order to measure the completeness, expected scopes and data fields need to be defined.

Validity Does data have a valid schema with correct data fields, data types, range, etc? This requires a well defined data structure and interface.

Uniqueness Are there duplicate records? Is data properly version controlled? For each data format, the definition of uniqueness is different. A version control system is typically required for high level data formats. For instance, if two versions of annotations exist on the same object without a unique identifier for each version, these records will then be considered duplicates and therefore not unique.

Timeliness Is data updated at the required frequency and available within a required delay? Note that for high level data formats, such as annotation and meta, the required frequency refer to both the currentness of data, and the fact that outdated historical data needs to be omitted in order to maintain an up-to-date and relevant description.

4.3 Data properties and quality specifications

Table 1: Examples of data properties developed in this work.

	Raw format	Media format	Meta format	Annotation format
Consistency	Monotonically increasing timestamping and colinearity between different timestamps within the multiplexed data streams	Visibility of objects that occur in different sensors around the same timestamp	Distributions of, e.g., class labels, 1) overall 2) consecutive data frames, 3) between sensors	The correlation between the size and class label of an object (e.g., a human cannot be 5 meters tall)
Accuracy	LiDAR detection count (may decrease significantly in foggy weather)	Camera exposure (over-exposure results in inaccurate measurements)	Code coverage of tests for statistics computations	Class label accuracy
Completeness	Packet loss rate	Sensor's field-of-view	Meta data coverage	Object occlusion
Validity	Checksum	Sensor visible range compared to the theoretical dynamic range	Confidence interval	Annotation schema
Uniqueness	Packets count at each unique timestamp	Undesired reflections and multipath interference	Reproducibility of statistics (i.e., two statistics of the same random variable on the same sample is considered a duplicate)	Unique version of annotations on the same object
Timeliness	Data logging latency	Decoding time	The time interval from which statistics are computed (i.e., samples shall be update-to-date, meanwhile outdated historical samples shall be discarded in order to avoid skewness)	Data processing speed (e.g., frames per second)

Given the abstraction of data formats and six dimensions of data quality, in this project, we have developed a set of systematic data properties and their corresponding quality specifications, which can be expressed as a table with data formats as its columns and quality measures as its rows. Some example properties can be found in Table 1, where each cell contains one example property for the corresponding cross section. Quality specifications are defined as use case specific requirements on each data property.

5 Objective

The objective of this project is to develop a data quality control toolbox for anomaly detection and error correction for improving reliability of AI systems for both development and deployment pipelines. This toolbox is designed to scale to large automotive data sets with high data coverage and processing speed. As a demonstrator of the toolbox, it is used for improving the accuracy of automated annotations, which is an important topic for automotive perception systems.

6 Results and deliverables

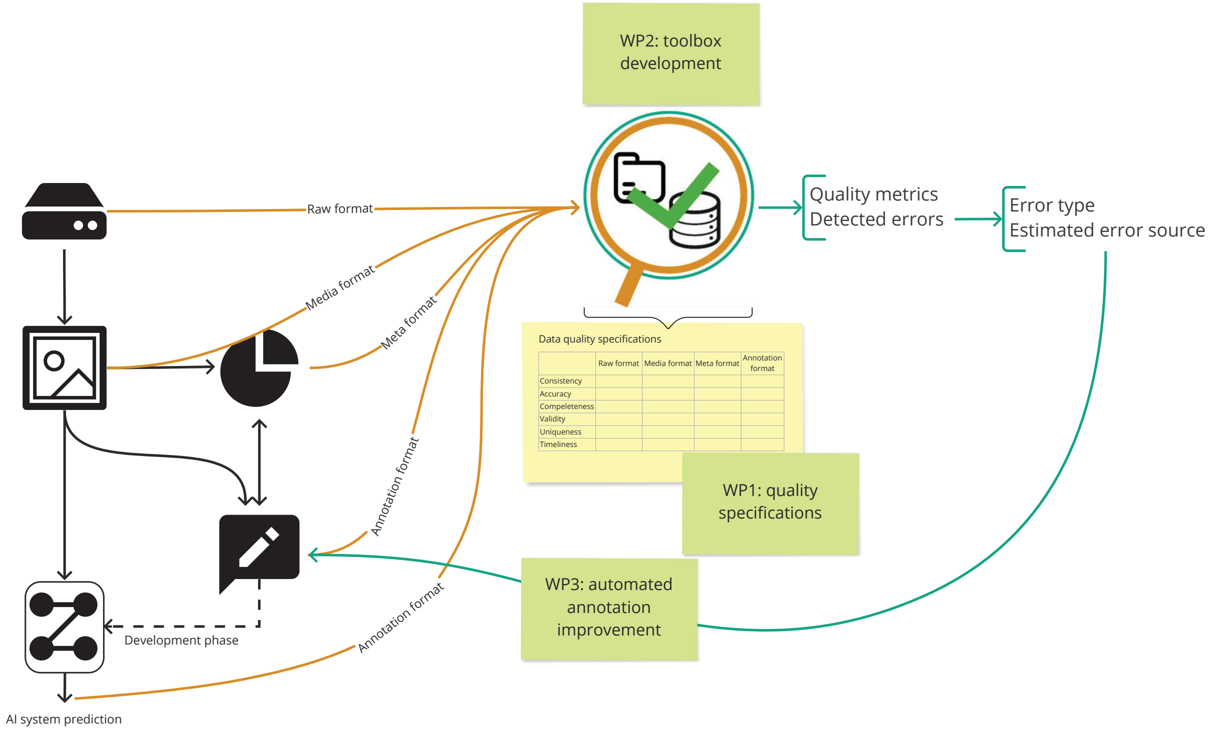
The implementation of this project is divided into three work packages. The relation between theses work packages is illustrated in Fig. 3.

6.1 WP1. Specifications of data quality metrics

In this work package, a set of specifications are defined for each data format. We have defined approximately 20 specifications for each format within the six data quality dimensions stated in Sec. 4.2. Note that these specifications are defined on a rather technical level. To meet business requirements, discussions with automotive stakeholders are planned as the next step. We plan to present the outcome of this work package as a scientific publication.

6.2 WP2. Qually: a validation toolbox for end-to-end data quality control

The outcome of this work package is a data quality control toolbox named *Qually*. The backend of this software is written primarily in Python. Qually is end-to-end in the sense that it detects anomaly and reports errors in data for all aforementioned data formats and their specifications developed in WP1.



Data quality specifications

	Raw format	Media format	Meta format	Annotation format
Consistency				
Accuracy				
Compeleteness				
Validity				
Uniqueness				
Timeliness				

Figure 3: Data transformation between different formats and the role of Qually.

6.3 WP3. Automated annotation improvement

To validate the applicability and effectiveness, in the last work package, Qually is applied to improve our automated annotation process using ensemble learning and uncertainty measures to evaluate the annotation quality and make suggested corrections. For instance, for object detections, such as instance segmentation and bounding box detection, a set of characteristics called *object signatures* are developed as a consistency measure. Note that the AI platform at Asymptotic AI mainly provides automated annotations. Errors that are specific to manual annotations are not handled explicitly by Qually. We plan to publish scientific papers and technical reports to disclose part of our findings.

Illustrative examples Example errors of different types are visually illustrated in Fig. 4, 5, 7, 6 and 8. Qually has successfully detected and corrected these errors by applying object logics and ensemble learning techniques. In this set of images, the figure to the left is the erroneous annotation identified by Qually, and the figure to the right is the version corrected by Qually.

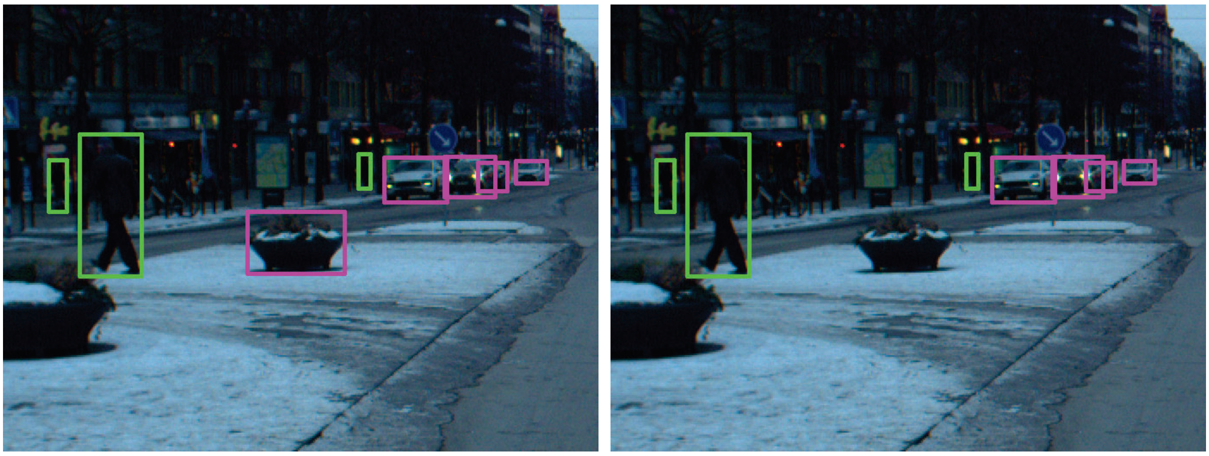


Figure 4: In this example, a *false positive* detection is present in the middle of the image to the left, where a plant pot is recognized as a car. In this case, Qually has issued a FAL_POS warning on the object by a weighted voting mechanism using multiple pattern recognition algorithms.

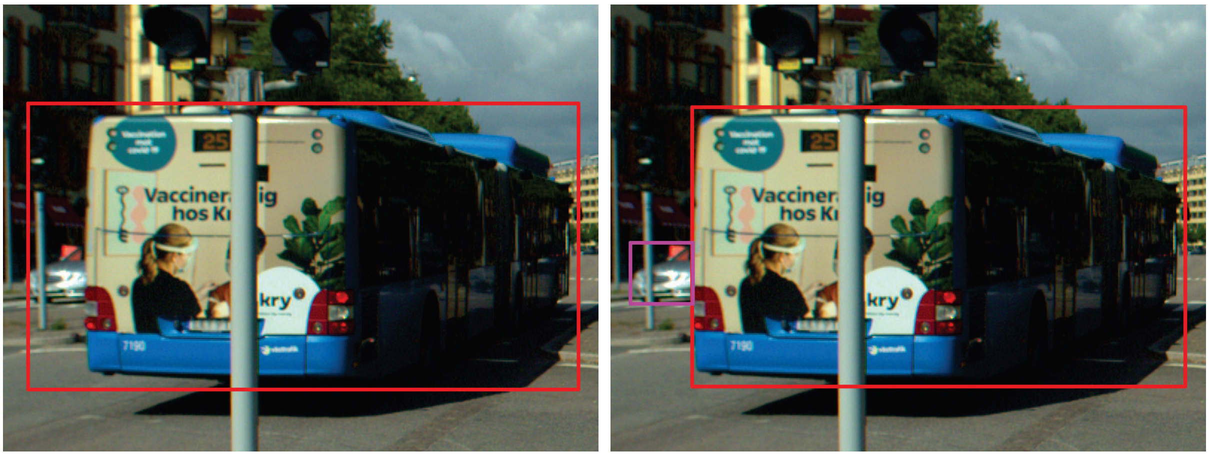


Figure 5: An example of frame-based 2D bounding box for object detection. This is a typical scenario, where the annotation of a bus has included an adjacent object (i.e., a car) to its left. This example illustrates two types of errors that Qually has detected and corrected: *false negative* on the car and *regression error* on the box.

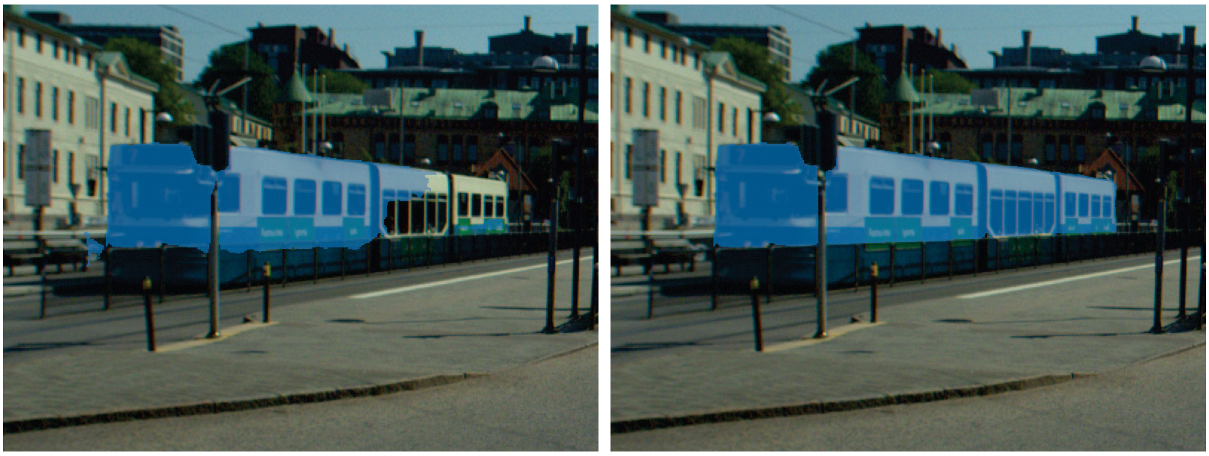


Figure 6: An example of frame-based instance segmentation. The tram class is underrepresented in the data set. Especially due to their unusually large spatial occupancy, instance segmentation is often inaccurate when they are close to the ego vehicle.

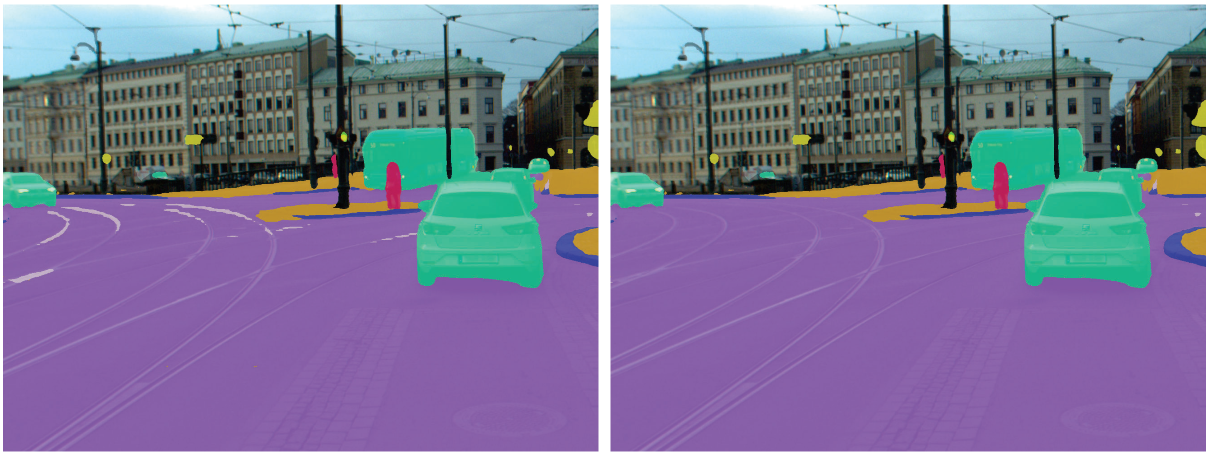


Figure 7: An example of frame-based semantic segmentation, where tram tracks are recognized as lane markings in this annotation. This is identified by ensemble techniques in combination with the assumptions on the geometry of lane markings.

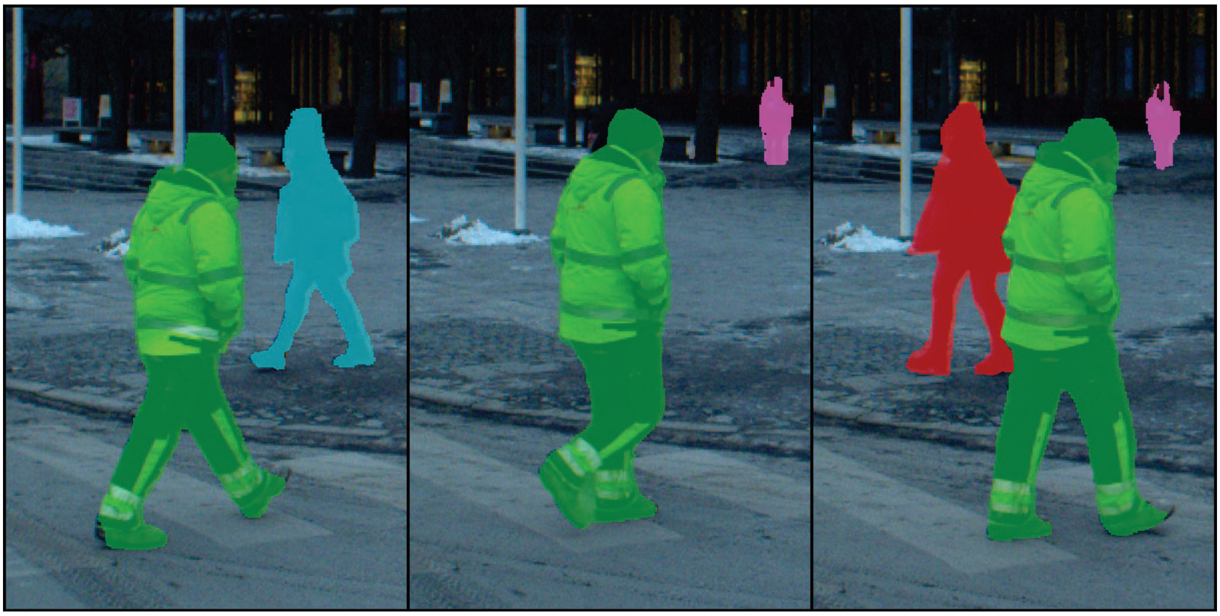


Figure 8: An example of sequence-based automated annotation error. This is typically referred to as the *association error*. In the image to the left, a pedestrian is detected and segmented as an object instance with a unique ID (blue mask). As time evolves, in the second image, the pedestrian walks behind another pedestrian in the foreground (green mask) and becomes completely occluded. In the image to the right, the pedestrian appears again and becomes visible. However, due to the temporary occlusion, the ID of the pedestrian is no longer persistent (blue → red mask). In Qually, this type of issues are mitigated by computing an *object signature* to maintain consistency between data frames in sequence-based annotations.

7 Dissemination and publication

The dissemination of this project is shown in Tab. 2.

Table 2: Dissemination

How has / is the project result to be used and disseminated?	Mark with x	Comment
Increase knowledge in the field	x	Data quality control plays a crucial role in AI. By investigating the state-of-the-art research and developing the quality control toolbox, we have gained essential knowledge in this field.
Be passed on to other advanced technological development projects	x	The developed quality control toolbox is applied to providing data to the EU project SHAPE-IT https://www.shape-it.eu/. It will also be used in various future projects for automotive research and development.
Be passed on to product development projects	x	By the end of the project, we have started the integration of the toolbox into our inhouse AI platform SnapXS to enrich its functionality.
Introduced to the market	x	As a result of the project, the toolbox is used to provide automated annotations for one of our automotive customers.
Used in investigations / regulatory / licensing / political decisions	N/A	N/A

8 Conclusion and further research

In this project, we have implemented a first version of the data quality control toolbox, Qually . To validate the applicability and effectiveness of Qually, we apply the technology to an important use case - error detection and correction for automated annotations - with satisfactory results. As a toolbox, Qually will be further developed after the project, where we will focus on improving properties such as robustness and scalability. There are limitations in this study. For instance, the data quality specifications studied in this project are rather on a low technical level. As the next step, the plan is to include more business requirements by having discussions with automotive stakeholders.

9 Participants and contact persons

Yinan Yu - Project leader
yinan.yu@asymptotic.ai

Samuel Scheidegger
samuel.scheidegger@asymptotic.ai

Jörg Bakker
jorg.bakker@asymptotic.ai