Edge AnnotationZ Challenge

Zenseact is at the frontier of autonomous vehicle development, and we want you to take part in pushing what is perceived to be possible. Together with you, we hope to develop solutions that can enable vehicles to improve their perception systems without any human supervision. This alleviates the burden of transmitting massive amounts of data to a central compute cluster, which ensures that the integrity of all information captured remains intact. By creating such a solution, we contribute to a better, safer, and integrity-preserving traffic environment while also advancing what is state-of-the-art in autonomous systems.

Introduction

There is currently a race within the automotive industry to provide ever-higher levels of automated driver assistance systems in their vehicles, where a safe, fully autonomous solution could be considered the holy grail. As these systems become more advanced and the Operational Design Domain (ODD) expands, the amount of data needed to train the perception systems increases rapidly. Not only will the labeling of the training data become prohibitively expensive, but also finding, recording, and storing all necessary rare cases with test vehicles will be extremely time-consuming. Further, transferring data from vehicles must be done in a cost-effective manner, while being legally compliant and keeping sensitive information safe.

Edge-learning, or federated learning, poses an attractive alternative to traditional methods. As capabilities for storage, compute and connectivity increase in today's vehicles, this could enable using the customer fleet for learning at massive scale. However, edge-learning is very much an active area of research, and more work is needed to understand this new learning paradigm and how to use it to improve our safety critical systems.

In the context of autonomous drive, there are many different types of tasks to solve. While some of them can be solved using methods that can be trained using only data, most of them require supervision in the form of annotations. For instance, in the common perception task of object detection, bounding boxes are needed to train an object detector. Traditionally, these bounding boxes are provided by human annotators. But, in the edge case, scenarios exist where the data cannot leave the vehicle and thus cannot be manually annotated. This requires new methods that can generate the annotations automatically.

We want you to help us create such methods. Together we can advance the state-of-the-art and further improve our autonomous driving systems. This will not just make it more comfortable and safer for the driver but has the potential to reduce accidents and fatalities in traffic and improve traffic safety in general.

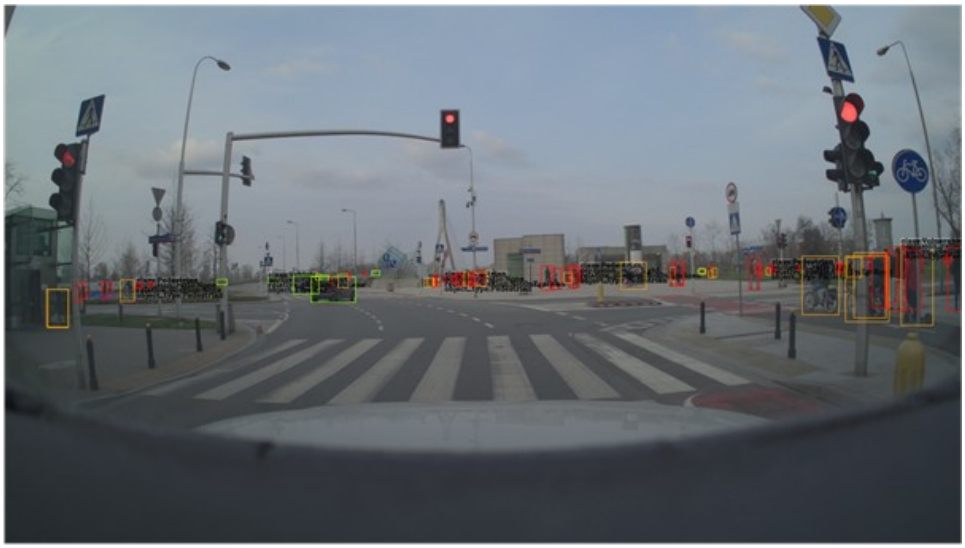

Figure 1: Sample image with objects annotated.

Problem description

In this challenge we want you to generate edge-annotations that would allow us to train our single-frame 3d object detection networks in a supervised manner. Specifically, this requires pairs of 2d and 3d bounding boxes, as well as the correct class assignment, for each dynamic object that is visible in each image.

We are trying to simulate a scenario where customer cars can save a limited amount of data each day, which is automatically annotated when the car is parked. Exactly how the auto-annotation should work is up to you. However, since this process occurs when the car is stationary and plugged in, you do not need to consider latency requirements associated with real-time systems. Further, unlike the online networks, you have access to information from both the past and the future and can use the benefit of hindsight.

We have no specific approach in mind for this challenge and therefore attempt to provide you with all the information that could be available in the edge scenario. That is, sequences of sensor data from cameras, LiDARs, high-precision GPS, vehicle bus data, as well as the output of the perception system that runs in the car. The perception system will be emulated by simple single frame detectors. We will also provide these detectors as an example and potential starting point for improvements. This way you get to choose if you want to train your own networks or focus on fusing/tracking/smoothing the existing detections.

Even though this challenge envisions a future where all training happens at the edge, we assume access to some annotated data. This means that you will be able to train/tune various learning-based models as part of your edge-annotation system. This is simulated by the training set that you will receive.

Your performance will be evaluated based on how well your edge-annotations correspond to human annotations on a small, hidden, test set. Furthermore, we require you to submit code, and/or models that enables us to fully reproduce your results, see the evaluation sections for more details. Moreover, a document explaining your solution in English, and how to run the code is needed.

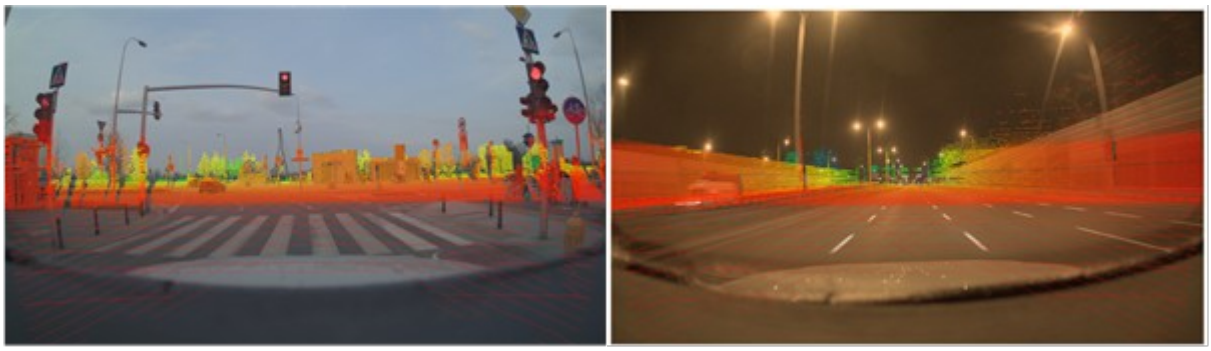

Figure 2: LiDAR detections overlayed in image.

Input-output example

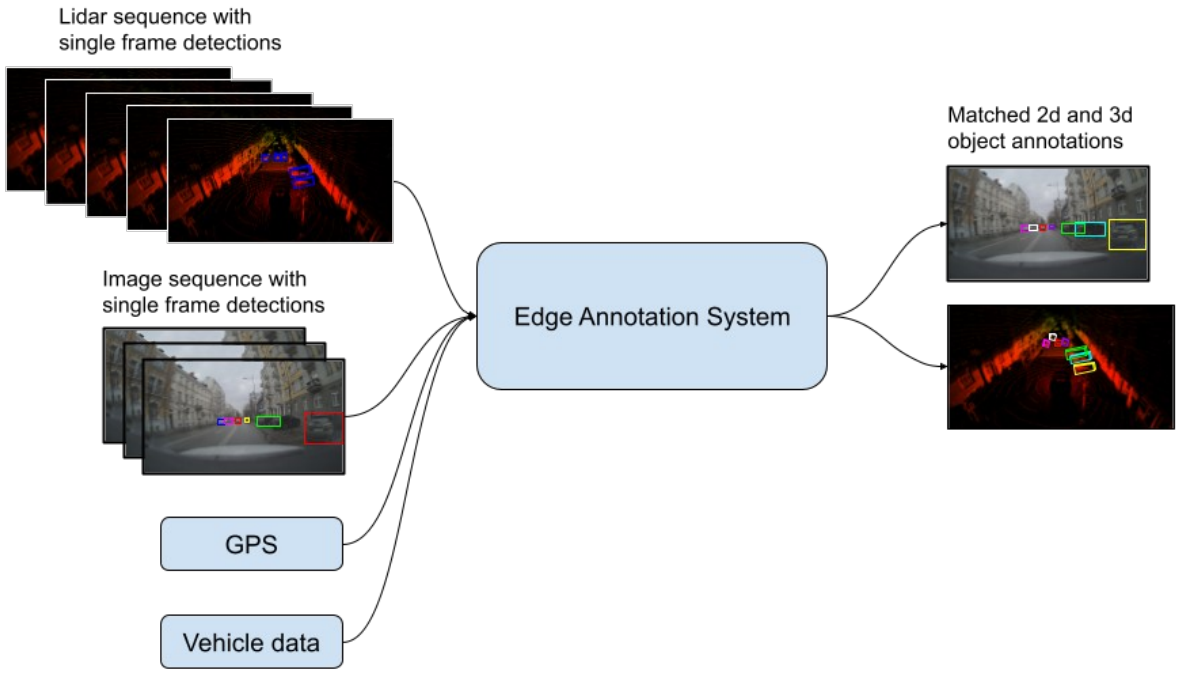

Figure 3: Schematic for expected I/O for your system.

Dataset description

The dataset we provide for this challenge is a multimodal dataset consisting of 6666 unique sequences, captured by Zenseacts' development vehicles on highway, country, and urban roads in and around Warsaw, Poland. Out of these, 5238 sequences contain annotations for dynamic objects and from these we withhold 788 sequences for our hidden dataset, resulting in 4450 annotated sequences available during development. Each sequence is composed of sensor data from LiDAR, camera, high-precision GPS and the vehicle bus (e.g., IMU, lateral and longitudinal velocity).

What is unique about this challenge is the fact that we want to leverage both past and future information to make an accurate annotation for a specific target frame. Therefore, we provide data around this target frame for all sensors. More specifically, for each sequence we provide

- 3 camera frames, captured at 30 Hz
- 21 LiDAR point clouds, captured at 10 Hz
- 1s of high-precision GPS, before the target frame, and up to 300 meters of ego-vehicle travel past the target frame (equivalent to about 10s on average) at 100 Hz
- Annotations for the target frame
- 2D and 3D predictions from pretrained single-frame detectors (vision and LiDAR).

Note here that the middle-frame is what we consider to be the target frame, i.e., we have 1 camera frame before and after the target frame and 10 LiDAR point clouds before and after the target frame.

While the dataset provides annotations for static objects, road-edges etc., we are in this challenge focusing solely on dynamic objects. We are even limiting the scope of the challenge even further by only considering vehicles (cars, trucks etc.) and vulnerable vehicles (motorcycles and bikes).

The annotations we provide are quite extensive. They include properties that are not strictly necessary to perform the task at hand. To be clear, we do not want you to create these properties in your edge-annotations. For each predicted object, we want you to provide:

- 2D bounding box
- 3D center point (x, y, and z coordinates)
- 3D extent (height, width, and length)
- 3D rotation (yaw angle)
- Object class

For a more detailed description of the dataset, we refer to the dataset document attached to this challenge.

Limitations

The single frame detectors, whose detections we provide you with, have been pretrained on the KITTI dataset, and fine-tuned on the training dataset for this challenge. To not limit your solution space, you are allowed to train your own detectors. However, these should only be pretrained using the KITTI dataset. For vision, we also allow pretraining on ImageNet.

Each team will be provided with a single GPU at the AI Sweden infrastructure to use during the challenge. The GPU is a NVIDIA A100 40GB. The dataset is only available on the provided computer node.

While you are allowed to train on any hardware of your choice, your code for inference must be compatible with the hardware provided. This is such that we can reproduce your results and verify your solution.

Lastly, you are not allowed to manually annotate any of the data. Only the annotations provided in the dataset should be used for supervised learning, while the remaining data can be used in a self-supervised manner.

Evaluation & Metrics

Your solution will be evaluated based on how good edge-annotations you can provide at the target frame. The metric for determining the quality of your annotations is mean average precision. Note also that the mean average precision will be a (weighted) average over the 2D and 3D properties.

As mentioned previously, your provided code must be runnable on the AI Sweden infrastructure environment to be considered for evaluation. Failing to comply with this will lead to an invalid result. Moreover, the solution must comply with the spirit of the challenge for eligibility.

How to enter

The number of teams that can enter the challenge will be limited due to infrastructure limitations. To enter the challenge, please send us a written application containing a proposed solution method, including identified risks, and mitigation strategies. Also provide a resume describing the relevant background and experience of the team (see the application form). Based on the written applications, eight teams will be chosen to participate in the challenge.

We acknowledge that to construct a solution strategy, it is often necessary to have access to the data. Hence, a smaller, representative data set will be made available before the competition through AI Sweden, in addition to a development kit.