

Closing the Loop: Combined Edge-Cloud Approach to Minimize Imbalance in Datasets

Maximiliano Cuevas, Nenad Ocelic, Torsten Stiehm (DXC-Luxoft)



Introduction

Collecting high-quality and well-balanced training data remains a key challenge in the development of reliable AI-based perception systems. In particular, object-specific data collection often suffers from imbalanced distributions across relevant object properties (see Fig. 1). Logging-based data acquisition strategies can therefore benefit from property-related data filtering, where specific object properties act as trigger conditions for selective data logging. Such targeted logging is especially important in scenarios with limited storage or bandwidth, where only a subset of data can be retained.

To achieve balanced datasets across multiple parameters, it is not sufficient to simply increase the overall amount of collected data. For example, in object detection datasets, collecting an excessive number of samples at similar object scales or aspect ratios can lead to biased models and reduced generalization performance. Instead, data collection should aim for a well-distributed coverage of relevant geometric and positional properties.

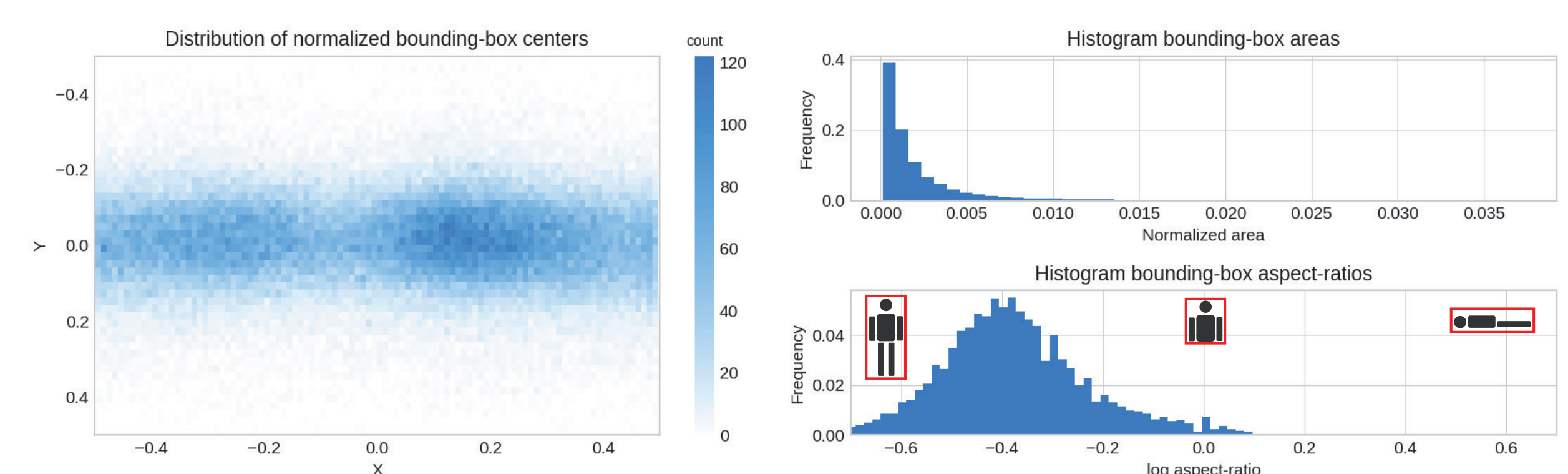


Figure 1: Distribution of person bounding box properties in BDD100k [3] (© DXC-Luxoft)

Approach

This work presents a simple and effective approach to property aware data filtering. Autoencoders are employed as a lightweight mechanism to characterize objects based on geometric properties such as size, aspect ratio, and position within the image. By leveraging these features, data can be selectively filtered or prioritized during logging, enabling a more balanced and informative dataset without requiring complex manual heuristics.

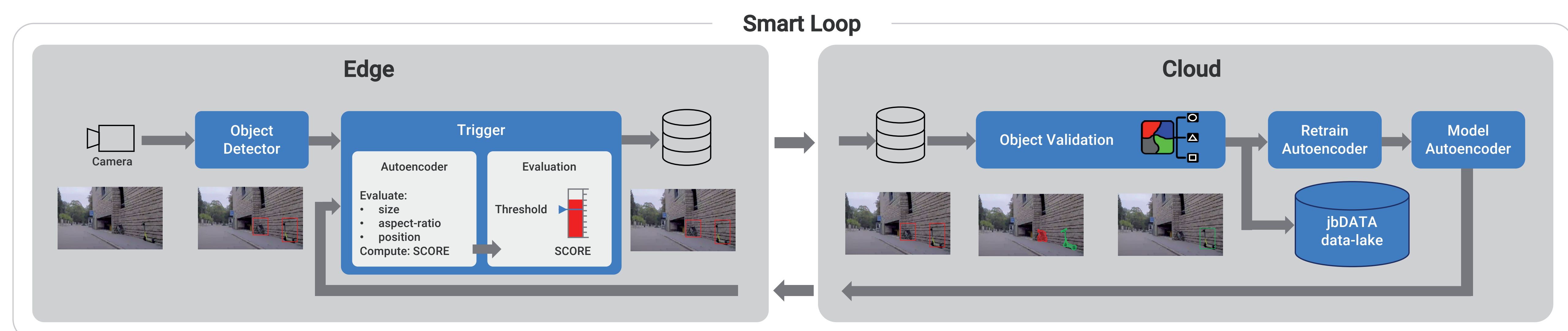


Figure 2: Dual validation of anomaly detection (© DXC-Luxoft, front camera images by b-plus)

Anomaly detection in the edge

Autoencoders learn to reconstruct their input data. Consequently, data drawn from the same distribution as the training data set typically produce a low reconstruction error, i.e.: mean square error (MSE) between input and output, while out-of-distribution data results in significantly higher reconstruction errors (see fig. 3).

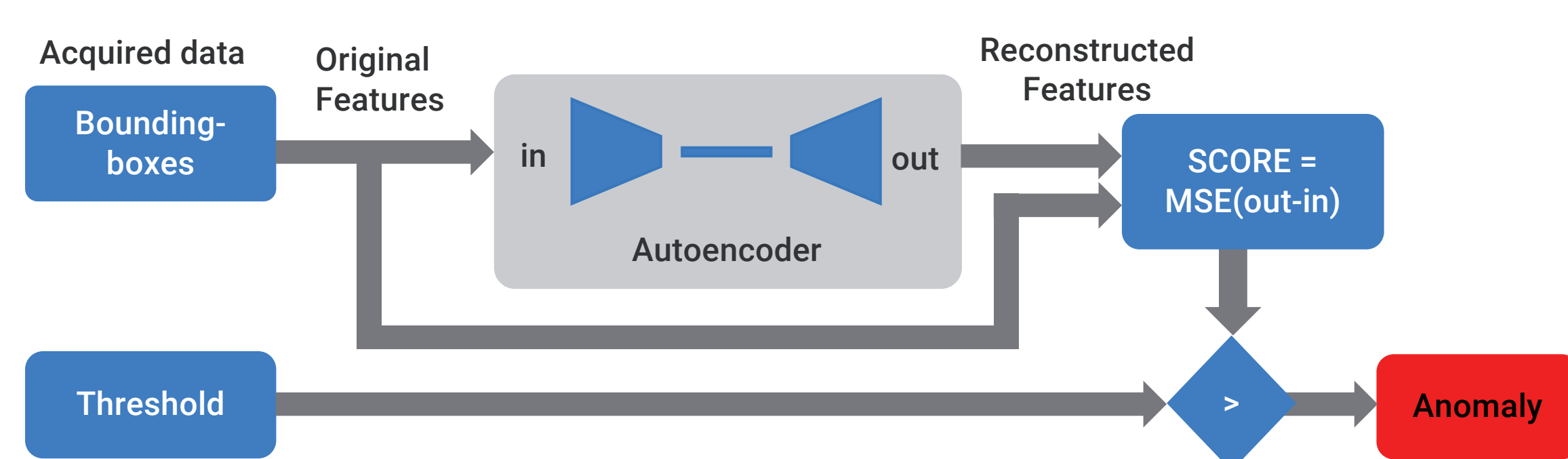


Figure 3: Autoencoder for atypical bounding-box detection (© DXC-Luxoft GmbH)

We have explored the use of autoencoders to find anomalies related to the position, size and shape of bounding-boxes. In this context, a bounding-box is considered an anomaly when some of the mentioned features are not frequent. For example, the position of the bounding-box lies in a region of the image which is unusual, or its size and aspect-ratio are not typical.

This property of autoencoders has been integrated into the trigger algorithm, which helps to discard those objects that are very common in a dataset, in terms of position, size and shape, promoting object diversity within a class.

Anomaly validation in the cloud

The anomalies found at the edge by the autoencoder are validated in the cloud by a dedicated validation component. For this purpose, we developed Classify Everything, a pipeline based on a novel prompting approach by Luxoft which integrates SAM [1] and Alpha-CLIP [2]. It enables precise zero-shot segmentation and object detection for arbitrary objects. As a result, objects that do not belong to the target class are rejected and valid bounding-boxes are refined to tightly fit the object. This set of curated data is aggregated to the data-lake and employed to retrain autoencoder in the next loop.

References:

- [1] Nikhila Ravi et al. SAM 2: Segment Anything in Images and Videos, <https://arxiv.org/abs/2408.00714>, 2024
- [2] Zeyi Sun et al. Alpha-CLIP: A CLIP Model Focusing on Wherever You Want. <https://arxiv.org/abs/2312.03818>, 2022
- [3] Fisher Yu, et al. BDD100K. <https://arxiv.org/abs/1805.04687>

For more information regarding the poster, contact:

torsten.stiehm@dxo.com

l.cuevaseiroa@dxo.com

[linkedin.com/company/ki-familie](https://www.linkedin.com/company/ki-familie)
www.justbetterdata.de

just better DATA is a project of the KI Familie. It was initiated and developed by the VDA Leitinitiative autonomous and connected driving and is funded by the Federal Ministry for Economic Affairs and Energy.

