Towards Scene Understanding for Autonomous Operations on Airport Aprons

Daniel Steininger[0000−0003−3810−1803], Andreas Kriegler[0000−0002−5653−5181], Wolfgang Pointner[0000−0002−3141−8405], Verena Widhalm[0000−0001−9318−1418], Julia Simon[0000−0003−0243−9706], and Oliver Zendel[0000−0001−8097−1226]

AIT Austrian Institute of Technology
{daniel.steininger,andreas.kriegler,wolfgang.pointner,verena.widhalm, julia.simon,oliver.zendel}@ait.ac.at

Abstract. Enhancing logistics vehicles on airport aprons with assistant and autonomous capabilities offers the potential to significantly increase safety and efficiency of operations. However, this research area is still underrepresented compared to other automotive domains, especially regarding available image data, which is essential for training and benchmarking AI-based approaches. To mitigate this gap, we introduce a novel dataset specialized on static and dynamic objects commonly encountered while navigating apron areas. We propose an efficient approach for image acquisition as well as annotation of object instances and environmental parameters. Furthermore, we derive multiple dataset variants on which we conduct baseline classification and detection experiments. The resulting models are evaluated with respect to their overall performance and robustness against specific environmental conditions. The results are quite promising for future applications and provide essential insights regarding the selection of aggregation strategies as well as current potentials and limitations of similar approaches in this research domain.

Keywords: dataset design · scene understanding · classification · object detection · airport apron · autonomous vehicles.

1 Introduction

While many research activities in recent years were focused on increasing the autonomy of road vehicles, assistant and autonomy functions for vehicles in off-road domains such as airport environments are still in their infancy. These tasks pose similar requirements regarding safety and robustness aspects, but must be executed in a significantly different domain, which hinders a straight-forward application of existing approaches and datasets. Especially the transition from classic to learning-based computer vision approaches requires high amounts of domain-specific image data for training and testing purposes.

Therefore, the aim of this work is to mitigate this data gap by creating a versatile dataset focusing on apron-specific objects and presenting an efficient approach for data acquisition, sampling and aggregation, which may serve as a

precursor for automating mobile platforms in other challenging domains. Image data was acquired by mounting cameras on multiple transport vehicles, which were operated in apron and logistics areas throughout multiple seasons. The high number of captured sequences covers a wide range of data variability, including variations in environmental conditions such as time-of-day, seasonal and atmospheric effects, lighting conditions, as well as camera-related degradation effects. Furthermore, an efficient sampling and meta-annotation approach was developed to automatically extract a representative set of samples from the extensive amount of recorded image data while minimizing the manual effort required for annotation. We additionally introduce a novel data aggregation strategy and provide preliminary models for detecting and classifying apron-specific objects. To summarize, we propose the following contributions:

- We introduce a novel dataset[1] specialized on objects encountered in apron areas including efficient approaches for image acquisition, dataset design and annotation.
- We train and evaluate baseline models for classification and detection on multiple dataset variants and thoroughly quantify the models' robustness against environmental influences.

Overall, we believe that our contributions provide vital insights in a novel and highly relevant application domain as well as universal strategies for efficient dataset design and experiments setup.

2 Related Work

While detecting and classifying objects encountered on airport aprons represents a novel application domain, there are several links to prior research, especially regarding learning and dataset-design approaches as well as existing datasets containing relevant objects.

Automating apron vehicles requires consistent robustness under a wide range of challenging environmental conditions. While some works focus on specific aspects such as variations in either daytime [4], weather [23] or image degradation [14] for benchmarking models or investigating the variability of existing datasets [1,2], few of them provide a comprehensive analysis regarding the impact of multiple factors on model performance [31,37] and neither includes the classes relevant for the given domain. Capturing the required data for versatile learning experiments can either be accomplished by complex sensor systems providing high-quality multi-modal data, as demonstrated by KITTI [8] and ApolloScape [12] or more portable equipment usually facilitating a significantly higher number of recording sessions and therefore higher data variability, as shown by Mapillary [25]. Objects typically encountered on airport aprons include certain common classes like aircraft or persons, which are part of established datasets, such as MS COCO [19], PASCAL-Context [24], ADE20K [39] and OpenImages [18].

[1] Images and annotations are available at https://github.com/apronai/apron-dataset

However, their coverage in terms of number and variability in these datasets is rather limited. Additionally, a number of dedicated datasets is available for the category of persons [34,38,3,40]. Similarly, different types of aircraft are provided by specialized datasets offering a more fine-grained categorization [22,31], as well as top-view or satellite [21,33,29] imagery.

The majority of relevant classes for the target application, however, is highly specific to the airport domain and rarely occurs outside it. This includes objects such as specialized airport vehicles, traffic signs or containers, which are rarely captured due to safety-related access restrictions. To the best of our knowledge, the presented dataset is the first to put the focus not only on airplanes, but the entire environment of airport aprons.

3 The Apron Dataset

Automating transport vehicles on airport aprons requires a reliable perception of this highly specialized environment. Therefore, the dataset's label specification is focused mainly on multiple types of apron vehicles, but also includes other kinds of static and transient obstacles. Ensuring the necessary relevance and efficiency in creating the dataset requires a consistent strategy across all stages of dataset design, as presented in the following subsections.

3.1 Data Acquisition

To match the requirements of the intended application as closely as possible, data acquisition was conducted in a realistic environment from the transport vehicle's expected point of view. Therefore, cooperating with a commercial airport was essential to gain access to transport vehicles in regular operation. However, this critical infrastructure implies that compliance with safety and legal considerations was required, such as preserving the privacy of passengers and airport staff, as well as ensuring that the recording campaign never interferes with airport and logistics operations. All image data was recorded using Nextbase 612GW dashcams with a resolution of 3860x2160 pixels, providing sufficient image quality combined with low cost and efficient handling. They were mounted on the inside of the windshields of two container-transport vehicles. To ensure that recordings are paused when the vehicle is inactive, their power supply was coupled to the respective engines. To increase the flexibility for later applications, one of them was modified to incorporate a lens with 90° field of view instead of the built-in lens with 150°. Most of the data was captured in time-lapse mode at 5 fps to provide a data variability sufficiently representing the environment. Furthermore, the recordings were complemented by sequences at 30 fps for demonstration purposes as well as future developments such as multi-object tracking.

Over a time period of six months we recorded 1715 image sequences, covering the seasons of spring, summer and autumn. Since transport vehicles typically traverse between multiple locations in logistics and apron areas, these recordings conveniently contain all kinds of objects encountered along their routes. On

the other hand, they include irrelevant sections with low scene activity during parking or moving along monotonous regions, as well as motion blur and noise, which need to be pruned as a first step during annotation.

3.2 Sequence and Image Annotation

While removing highly redundant or irrelevant data, each remaining sequence is additionally assigned a defined set of parameters to specify the environmental factors during recording time.

- *Time of Day* describes the variance between natural and artificial light sources throughout the day.
- *Lighting* specifies sunny and diffuse conditions during daytime based on the appearance of shadows and is undefined for night recordings.
- *Atmosphere* differentiates multiple weather and atmospheric effects.
- *Scene Dynamics* is a measure for the number and activity of dynamic objects in a sequence, as well as variations due to motion of the capturing vehicle.

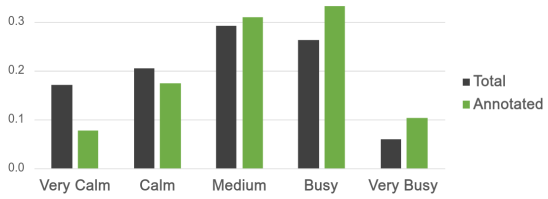

Fig. 1. Distribution of *Scene Dynamics* for total and annotated images

We eventually selected 1209 sequences representing a pool of more than 3.2 million images in total. They were sampled at varying rates proportional to the parameter *Scene Dynamics* aiming to reduce redundancies and extract a representative set of approximately 10k images. As demonstrated in Fig. 1, sequences tagged as *Busy* and *Very Busy* are oversampled by 25% and 50% respectively, whereas *Calm* and *Very Calm* scenes are undersampled analogously. Additionally, the few remaining redundant images, recorded when the vehicle was stopped with the engine running, were manually removed. The final set of 10098 images is annotated with additional per-image parameters. *Degradation* summarizes multiple factors expected to negatively influence model performance as shown in Fig. 2. While these factors typically appear simultaneously, the parameter is set to *High* if any of them significantly influences image quality. Emphasis is placed on near- and mid-range objects in a distance of up to 100 meters.

The distribution of parameters for all annotated frames is displayed in Fig. 3. *Time of Day* variations are well balanced due to the airport operating hours including dawn, dusk and a significant proportion of the night. *Lighting* conditions are annotated only for 68% of the dataset since the parameter does not

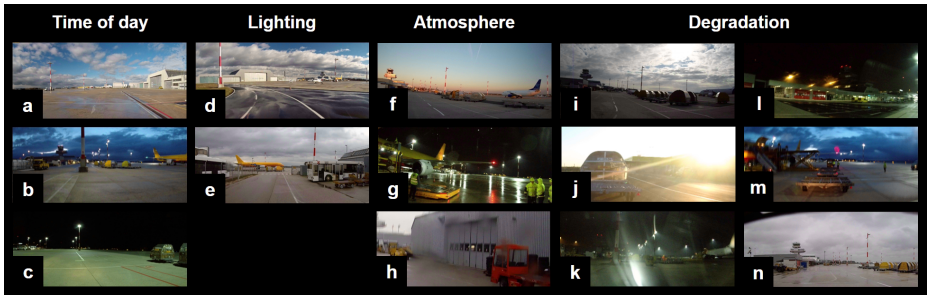

Fig. 2. Representative examples for sequence and image parameters. *Time of day* differentiates between *Day* (a), *Twilight* (b) and *Night* (c), *Lighting* between *Sunny* (d) and *Diffuse* (e) and *Atmosphere* between *Clear* (f), *Rain* (g) and *Heavy Rain* (h), with rain drops significantly impacting perception. The assigned state of *Low* or *High* for *Degradation* typically depends on multiple factors such as under- (i) and overexposure (j), windshield reflections (k, l), motion blur (m) and wiper occlusions (n)

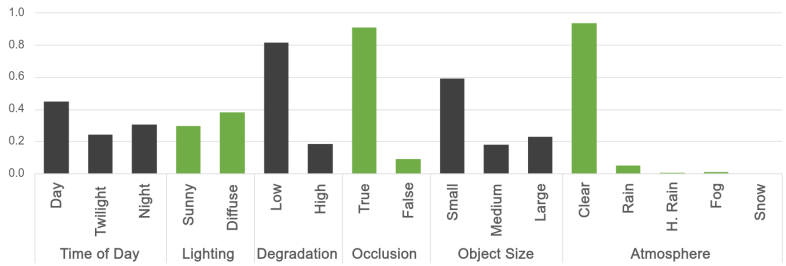

Fig. 3. Per-instance distribution of annotated meta parameters and object sizes (*Small* $< 15k$ pixels $<$ *Medium* $< 45k$ pixels $<$ *Large*)

apply to night recordings and ambiguous sequences containing both *Sunny* and *Diffuse* states. However, the remaining images are roughly evenly distributed, therefore representing a solid basis for evaluating their impact on detection and classification performance. Atmospheric effects, on the other hand, show a strong bias towards *Clear* conditions and thereby represent the environment encountered during the recording time. Nevertheless, the data includes a small number of images of *Rain* and *Heavy Rain* as well as *Fog*, which are useful for preliminary insights regarding the impact of harsh weather conditions as well as the generalization capability of trained models.

3.3 Instance Annotation

Specifying a set of object labels tailored to the target application requires an extensive analysis of sampled images to identify visually distinct categories for frequently appearing types of vehicles and obstacles. Additionally, safety-relevant classes are considered independently of their occurrence frequency. We aim at

annotating a fine-grained definition of classes which can be further condensed for training specialized models on more coarse-grained dataset variants. For this purpose, we define 43 categories, which are fully listed in the supplementary material and visualized in Fig. 4.

Fig. 4. Representative examples of different object categories included in the dataset

For each category we define a detailed textual specification along with multiple selected sample patches to minimize ambiguous assignments and resolve potential corner cases at an early stage. Since the focus is placed on near- and mid-range objects, the minimum size of annotated objects is defined as 28 pixels for vehicle classes and 12 pixels for traffic signs and persons along either dimension. Across the entire set of selected images, this results in a total of more than 169k object instances localized as bounding boxes and assigned one of the defined categories. Additionally, objects are tagged as occluded if they are not fully visible due to other objects or truncated at image borders.

It is well known that object-occurrence frequency in real-world images often follows a long-tailed distribution [20], [27], which is also visible in this case, as demonstrated in Fig. 5. The overabundance of a few head classes with the numerous tail classes collectively still making up a significant portion of the data [41] is challenging for learning systems. The environment on airport aprons is generally relatively structured and controlled but also crowded with an average of 16.8 objects per image. Compared to the total number of samples the object occurrence in terms of images containing a certain category, is more evenly distributed, indicating that especially head classes tend to appear in multitudes within a single image. The distribution of object sizes and its relation to the occurrence frequency is of additional interest, especially for the detection task. The mean of about 80k pixels and a surprisingly low median of 9k pixels indicate that a relatively low number of classes with exceedingly large objects stands in contrast to a large number of classes with small to medium objects.

While the method of image acquisition and the label definition affect these dataset statistics, the long-tailed distribution can be mitigated by exploiting the closed-off, controlled and repetition-driven nature of airport apron processes. Therefore, it might be easier to create accurate and robust models for this domain than it is for other automotive applications and we believe our dataset showcases a promising way to efficiently accumulate data for this purpose.

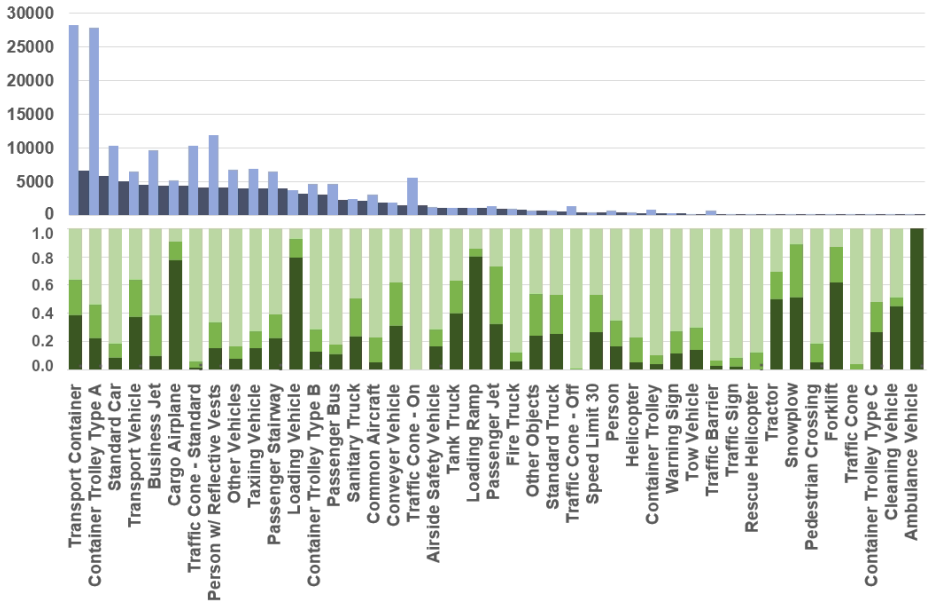

Fig. 5. Top: numbers of samples annotated for each object category (light color) and images containing them (dark color). Bottom: normalized size distribution: (**Large** > 45k pixels > **Medium** > 15k pixels > Small)

3.4 Data Aggregation

By mapping selected classes of the annotated data we define three variants to facilitate a comparative analysis of dataset balancing effects.

- *Fine* is the baseline variant including all 43 annotated labels
- *Top* limits the dataset to the 25 most frequent classes in terms of total object occurrence which contribute roughly 97% of samples
- *Coarse* uses the full set of instances but remaps them to only 23 superclasses based on semantic similarity

For all experiments described below, 68% of each dataset are used for training and 17% for validation, while the remaining 15% of samples are withheld during the experiments and exclusively reserved for testing purposes. This split is applied on a per-sequence basis to reduce the effects of over-fitting.

4 Fine-grained Classification

Fine-grained image classification focuses on correctly identifying differences between hard-to-distinguish object (sub-)classes and predicting the specific variants accordingly. Taking a look at Fig. 4 and the corresponding categories shown in Fig. 5, many visually similar yet distinct classes can be observed in the *Apron*

dataset: multiple types of container trolleys, aircraft, traffic signs and specialized cars and trucks are all commonly encountered on aprons. To correctly differentiate such visually similar classes, feature representations need to be rich in detail. On the other hand, the overall object variety requires a well generalized model, which makes the classification task especially challenging. While classification of object instances also takes place implicitly in detection architectures, a standalone, fine-grained classifier can be tuned and optimized more easily to gain vital insights regarding dataset variability and the final application setup.

Evaluation Metrics For evaluating classification performance a multitude of metrics with different advantages and drawbacks has emerged [30], though literature on classification metrics in the context of computer vision is sparse [9]. Top-1 accuracy (α), defined as the number of correct classifications over the number of ground-truth samples, has been reported on CIFAR [17] and ImageNet [5] and is still widely used [15,6,36,35,28]. However, on datasets with significant class-imbalance or long-tailed characteristics α leads to unintuitive results, since, for example, a model evaluated on a dataset with 90% of samples s belonging to class A and only 10% to class B achieves an α value of 90% by simply always predicting 'A'. This score of exclusively predicting the most frequent class is defined as the null accuracy α_0, where s_i is the number of samples of class i and s is the total number of samples:

$$\alpha_0 = \frac{max\{s_1, \ldots, s_n\}}{s} \tag{1}$$

To evaluate the significance of α, it should be compared to α_0 as well as to the random accuracy α_r, which represents the score if predictions are equally distributed over the number of classes (n).

$$\alpha_r = \frac{1}{n} \tag{2}$$

Moreover, α is prone to be even more biased in the case of top-x accuracy with $x > 1$ where a sample counts as correct if the true label is within the x most-confident model predictions. Therefore, we use the less biased metric of per-class average recall ($\bar{r}$) for selecting the best performing models of each experiment and evaluating all fine-grained classification models. This metric represents the average of class-wise Top-1 accuracies, as used by e.g. [19], rendering each class equally important independent of the number of samples assigned to it.

$$\bar{r} = \frac{1}{n} \sum_{i=1}^{n} \frac{TP_i}{TP_i + FN_i} \tag{3}$$

Additionally, we employ the metric of per-class average precision ($\bar{p}$) calculated analogously as the average ratio between true positives and the total number of predictions for each class:

$$\bar{p} = \frac{1}{n} \sum_{i=1}^{n} \frac{TP_i}{TP_i + FP_i} \tag{4}$$

In formulas above, TP_i and FP_i denote the number true and false positive classifications for class i, respectively, while FN_i is the corresponding number of false negatives. To improve clarity when comparing results, we furthermore report the f_1 score as the harmonic mean of both measures.

$$f_1 = 2 \times \frac{\bar{p} \times \bar{r}}{\bar{p} + \bar{r}} \tag{5}$$

Setup and Optimization We conducted a two-fold validation for comparing multiple ResNet [11] and EfficientNet [32] architectures and tuning their hyperparameters based on $\bar{r}$. Eventually, we chose EfficientNet-B3 since it outperforms the ResNet variants and can conveniently be adapted to specific data using a single scaling coefficient for modifying width, depth and image scale.

Object instances below 30 pixels along both dimensions or an aspect ratio exceeding 10:1 were excluded, slightly reducing the original dataset to roughly 150k samples. After evaluating multiple image-augmentation techniques, the best results were observed using a random horizontal flip before resizing to the required input size of 300 pixels. An additional application of random crop and Gaussian blur did not improve results on the validation set, indicating that the visual variability in the Apron dataset is already significant. It could be observed, however, that predictions on classes with relatively few samples were more accurate using stronger augmentations, since they benefit more from the additional variability.

All models were trained from scratch for 40 epochs using the SGD optimizer with a learning rate of 0.1, which yielded slower but stable convergence unlike Adam [16], as is often the case in PyTorch [26]. We updated the learning rate every 10 epochs, using a step ratio of 0.1 and a weight decay of 0.0005. The experiments were conducted on an NVIDIA RTX 2080 Ti using a batch size of 128. Furthermore, we used the Swish activation function, a CrossEntropy loss with a dropout rate of 0.3 and Kaiming uniform [10] parameter initialization.

4.1 Classification Results

Table 1 shows the results obtained on the respective test sets of all three dataset variants. The baseline variant *Fine* poses a significant challenge, but the corresponding model still obtains an f_1 score of 68.2%. As expected, leaving out low-frequency classes (*Top*) or merging them to superclasses (*Coarse*) significantly improves performance by up to 12.6%. Precision $\bar{p}$ tends to be only slightly higher than recall $\bar{r}$ despite the long-tailed class distribution of the dataset, indicating consistent classification performance across most classes. As expected, obtained α-scores are far above $\bar{r}$, giving a skewed and less distinctive impression of the models' performance and are therefore omitted for more detailed comparisons.

Table 1. Classification results on the *Fine* (F), *Top* (T) and *Coarse* (C) dataset variants as average recall, average precision and f_1 score, as well as top-1, null and random accuracy. The last two columns specify the numbers of samples in the test sets and the entire dataset variants, respectively

	$\bar{r}$	$\bar{p}$	f_1	α	α_0	α_r	s_T	s
F	0.624	0.752	0.682	0.866	0.184	0.023	22.6k	150.6k
T	0.779	0.810	0.794	0.880	0.193	0.040	21.9k	145.9k
C	0.819	0.798	0.808	0.881	0.213	0.044	22.6k	150.6k
⌀	**0.741**	**0.787**	**0.762**	**0.876**	-	-	-	-

4.2 Robustness Analysis

To gain more detailed insights regarding the impact of environmental effects on model performance we filter the test sets by each of the parameters defined in Section 3.2 as well as object size and occlusion. The corresponding evaluations on each resulting set are presented in Tables 2 and 3. Distributions of test sets are similar to those of the overall dataset presented in Fig 3. Note that the samples do not cover the entire test set for the *Lighting* and *Atmosphere* parameters. In the former case this results from the parameter not applying to *Night* settings, while in the latter case all underrepresented conditions were omitted.

Table 2. Impact of environmental conditions on classification performance as deviation from overall f_1 scores (Table 1) for each model on the corresponding test set

	Time of day			**Lighting**		**Degradation**		**Atmosphere**	
	Day	**Twilight**	**Night**	**Sunny**	**Diffuse**	**Low**	**High**	**Clear**	**Rain**
F	-0.020	-0.025	-0.005	0.015	-0.008	-0.005	-0.007	0.005	-0.101
T	0.003	0.006	-0.013	0.006	0.006	0.005	-0.028	0.000	-0.008
C	0.006	-0.012	-0.011	0.018	-0.007	0.008	-0.011	0.002	-0.039
⌀	**-0.004**	**-0.011**	**-0.009**	**0.013**	**-0.003**	**0.003**	**-0.015**	**0.003**	**-0.049**

Overall, as visible in Table 2, both the positive and negative deviations are relatively small across all dataset variants and parameters, indicating that most conditions are sufficiently covered in the dataset. Since the recording vehicles accumulated image data over a long period of time, they encountered a wide variety of conditions expected during long-term autonomous operation. The least deviation is reported between the different times of day. Since all three values are extremely small and recall and precision values are averaged across classes, the offsets of underrepresented classes can even sufficiently distort results for all deviations to be negative in this case. The trends are more clearly visible for the lighting and degradation parameters, where sunny conditions and low image degradation appear to be the least challenging for all models. The strongest negative impact is visible for light rain which reduces the f_1 score by more than

5% compared to clear atmosphere. It is possible that the model learns more ambiguous and blurred filters for rainy samples and as such the predictions are spread more equally across all classes. On the other hand, for samples with clear sight, the learned filters might be very class-specific and overfit on the classes with many samples, since this phenomena is most noticeable for *Fine*.

Table 3. Impact of object size (*Small* $< 15k$ pixels $<$ *Medium* $< 45k$ pixels $<$ *Large*) and occlusion on classification performance as deviation from overall f_1 scores (Table 1) for each model on the corresponding test set

	Object size			**Occlusion**	
	Small	**Medium**	**Large**	**True**	**False**
F	-0.066	0.001	-0.042	0.003	0.004
T	-0.070	0.001	0.013	-0.001	0.033
C	-0.071	0.012	0.033	-0.004	0.025
Ø	**-0.069**	**0.005**	**0.001**	**-0.001**	**0.021**

As expected, Table 3 shows that all models perform significantly better for medium-sized and large objects than for those smaller than 15k pixels. The models seem to be suitable for fine-grained classification tasks where objects are relatively close and largely depicted and therefore most relevant for the intended application scenarios, while tiny and distant objects are more prone to errors. Furthermore, the difference in scores depending on occlusion of objects is relatively low, since more than 90% of all objects in the entire dataset are occluded, providing a rich set of representative training examples.

5 Detection

Based on the insights and promising results gained during our classification experiments, the next step towards the real-world application of autonomous vehicle operation is to analyze entire scenes by localizing and simultaneously classifying objects using an end-to-end detection approach.

Evaluation Metrics We evaluate the results based on the established average-precision (AP) metric defined as the area under the precision-recall curve for each class. But instead of using a single IoU threshold to distinguish between correct and incorrect detections as traditionally used in detection challenges such as Pascal VOC [7], we average the results over 10 IoU thresholds ranging from 0.5 to 0.95, as suggested by the COCO challenge [19]. The reported overall values are subsequently averaged over all available classes of the respective dataset variant.

Setup and Optimization We conducted the experiments using an existing implementation of YOLOv5 (release 6.1) [13] and compared the results of the

small, medium and large architectures on the defined dataset variants. All images were scaled to 1280 pixels along each dimension and subjected to standard data augmentation. We trained on a system with two NVIDIA RTX 3090, using the SGD optimizer with a linear learning rate of 0.01 and a batch size of 16. All models were initialized with the pre-trained weights provided by the authors of [13]. For each combination of architecture and dataset variant we selected the best-performing model out of 50 training epochs based on the validation results.

5.1 Detection Results

As visible in Tab. 4, performance increases with model complexity. Analogous to the original experiments on the COCO dataset [13] the gain is more significant between the small and medium than between the medium and large architecture.

Fig. 6. Representative detection results of the *Coarse* model on the corresponding test set (green: correct detection (*TP*), orange: correct localization, but incorrect class, red: incorrect detection (*FP*), blue: undetected ground-truth object (*FN*))

For each architecture, the dataset variant with the highest granularity (*Fine*) serves as a baseline, since it is the most challenging. As expected, limiting the

Table 4. Detection *APs* on the *Fine* (*F*), *Top* (*T*) and *Coarse* (*C*) datasets variants by model architecture on the corresponding test sets. For comparing the expected computational complexity, the numbers of parameters and floating point operations (FLOPs) for each model are given as stated in [13]

	YOLOv5s6	**YOLOv5m6**	**YOLOv5l6**
Parameters	12.6M	35.7M	76.8M
FLOPs	16.8B	50.0B	111.4B
F	0.364	0.415	0.435
T	0.441	0.480	0.501
C	0.473	0.514	0.526
⌀	**0.426**	**0.470**	**0.487**

label categories to the 25 most frequent ones (*Top*) and thereby reducing the total number of samples improves the score. However, the highest robustness is achieved by the variant combining semantically similar object categories and thereby using all available samples (*Coarse*), which results in a similar number of classes but higher variability. The evaluation therefore indicates that limiting the number of classes by remapping provides a superior alternative to simply omitting under-represented classes regarding accuracy as well as flexibility.

The high performance indicated by the scores using the challenging COCO metric is also noticeable in the qualitative results presented in Fig. 6. The model is well capable of handling occlusions in crowded scenes as well as varying environmental conditions, including different times of day, as well as moderate image degradation, as visible in the upper two rows. However, effects such as strong motion blur and significant underexposure decrease detection performance, as visible in the lower right visualization. Furthermore, even small objects can robustly be localized in most cases. However, they are more prone to being assigned wrong categories, as discussed in section 4.2 and depicted in the lower left image.

5.2 Robustness Analysis

To quantify the influence of environmental effects on detection performance, we evaluate all models based on the filtered test sets analogously to Section 4.2, as shown in Table 5. As an indicator for the significance of each parameter the number of corresponding samples in the test set for the *Fine* and *Coarse* dataset variants are specified, with the number for *Top* being marginally smaller. As described for the classification results in Section 4.2, the total number of samples for each parameter does not necessarily cover the entire dataset.

The slight impacts that can be observed are strongest for changes in lighting conditions as well as image degradation. Since operating areas are well lit at night, the results for the *Time-of-day* parameter confirm that a single model is suitable for operating 24 hours a day. Furthermore, light rain can be handled well with only a slight decrease in performance, while more challenging weather effects require more training and test data.

Table 5. Average impact of environmental conditions as absolute deviation from overall detection *APs* (Table 4) on each test set across all three selected model architectures

	Time of day			**Lighting**		**Degradation**		**Atmosphere**	
	Day	**Twilight**	**Night**	**Sunny**	**Diffuse**	**Low**	**High**	**Clear**	**Rain**
$\mathbf{s_T}$	6.9k	4.5k	5.6k	5.8k	5.4k	13.6k	3.4k	16.1k	0.6k
F	0.039	-0.004	-0.001	0.037	0.027	0.002	0.002	0.002	0.058
T	0.003	0.031	0.005	0.000	0.000	0.002	-0.021	0.000	-0.019
C	0.016	-0.006	-0.003	0.025	-0.010	0.009	-0.030	0.003	-0.041
⌀	**0.019**	**0.007**	**0.000**	**0.021**	**0.006**	**0.004**	**-0.016**	**0.002**	**-0.001**

6 Conclusion

In this work, we demonstrated the process of creating an extensive dataset for the novel application domain of autonomous operation on airport aprons. We introduced efficient concepts for image acquisition and annotation before training and evaluating models for classification and detection based on multiple variants of the dataset. Additionally, we enriched the analysis with annotations of environmental conditions and quantified their impact on model performance.

The results show that our models are already capable of robustly detecting and classifying most relevant near and mid-range objects, rendering them a promising foundation for the further development of assisted and autonomous vehicle operation in this application domain. We achieved our aim of training robust models covering variable conditions at the specific airport used for recording the dataset. While we are aware that the resulting models do not seamlessly generalize to different locations and novel object classes, our dataset and the presented insights represent a valuable basis for significantly reducing the effort and required data to specialize on other airport environments.

We plan to evaluate the resources required for specializing our models and dataset to novel locations by recording additional training and test data at other airports to gain further insights on the re-usability of our concepts and data and their combination with additional approaches. Especially combining the results with multi-object tracking facilitating the propagation of object instances over time holds the potential to further increase detection robustness and facilitate embedded real-time processing on a mobile vehicle.

Acknowledgement. We would like to thank the Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology, and the Austrian Research Promotion Agency (FFG) for co-financing the "ICT of the Future" research project AUTILITY (FFG No. 867556). Additionally, we want to thank our project partner Linz Airport, Quantigo AI and our annotation team consisting of Vanessa Klugsberger, Gulnar Bakytzhan and Marlene Glawischnig.

References

1. Asudeh, A., Jin, Z., Jagadish, H.: Assessing and remedying coverage for a given dataset. In: 2019 IEEE 35th International Conference on Data Engineering (ICDE). pp. 554–565. IEEE (2019)
2. Auer, S., Demter, J., Martin, M., Lehmann, J.: Lodstats–an extensible framework for high-performance dataset analytics. In: International Conference on Knowledge Engineering and Knowledge Management. pp. 353–362. Springer (2012)
3. Braun, M., Krebs, S., Flohr, F.B., Gavrila, D.M.: Eurocity persons: A novel benchmark for person detection in traffic scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2019)
4. Dai, D., Van Gool, L.: Dark model adaptation: Semantic image segmentation from daytime to nighttime. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC). pp. 3819–3824. IEEE (2018)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
8. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
9. Grandini, M., Bagli, E., Visani, G.: Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756 (2020)
10. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Huang, X., Wang, P., Cheng, X., Zhou, D., Geng, Q., Yang, R.: The apolloscape open dataset for autonomous driving and its application. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2702–2719 (2019)
13. Jocher, G., Nishimura, K., Mineeva, T., Vilariño, R.: yolov5. Code repository https://github. com/ultralytics/yolov5 (2020)
14. Kamann, C., Rother, C.: Benchmarking the robustness of semantic segmentation models with respect to common corruptions. International Journal of Computer Vision **129**(2), 462–483 (2021)
15. Khan, A., Sohail, A., Zahoora, U., Qureshi, A.S.: A survey of the recent architectures of deep convolutional neural networks. Artificial intelligence review **53**(8), 5455–5516 (2020)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009)

18. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4. International Journal of Computer Vision pp. 1–26 (2020)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
20. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2537–2546 (2019)
21. Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P.: Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE (2017)
22. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
23. Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A.S., Bethge, M., Brendel, W.: Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv preprint arXiv:1907.07484 (2019)
24. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 891–898 (2014)
25. Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the IEEE international conference on computer vision. pp. 4990–4999 (2017)
26. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in neural information processing systems. pp. 8026–8037 (2019)
27. Salakhutdinov, R., Torralba, A., Tenenbaum, J.: Learning to share visual appearance for multiclass object detection. In: CVPR 2011. pp. 1481–1488. IEEE (2011)
28. Shen, Z., Savvides, M.: Meal v2: Boosting vanilla resnet-50 to 80%+ top-1 accuracy on imagenet without tricks. arXiv preprint arXiv:2009.08453 (2020)
29. Shermeyer, J., Hossler, T., Etten, A.V., Hogan, D., Lewis, R., Kim, D.: Rareplanes: Synthetic data takes flight (2020)
30. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information processing & management **45**(4), 427–437 (2009)
31. Steininger, D., Widhalm, V., Simon, J., Kriegler, A., Sulzbachner, C.: The aircraft context dataset: Understanding and optimizing data variability in aerial domains. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3823–3832 (2021)
32. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2019)
33. Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: Dota: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3974–3983 (2018)
34. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3415–3424 (2017)

35. Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A.L., Le, Q.V.: Adversarial examples improve image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 819–828 (2020)
36. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10687–10698 (2020)
37. Zendel, O., Honauer, K., Murschitz, M., Steininger, D., Dominguez, G.F.: Wilddash-creating hazard-aware benchmarks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 402–416 (2018)
38. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., Tian, Q.: Person re-identification in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1367–1376 (2017)
39. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)
40. Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., Ling, H.: Detection and tracking meet drones challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2021)
41. Zhu, X., Anguelov, D., Ramanan, D.: Capturing long-tail distributions of object subcategories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 915–922 (2014)