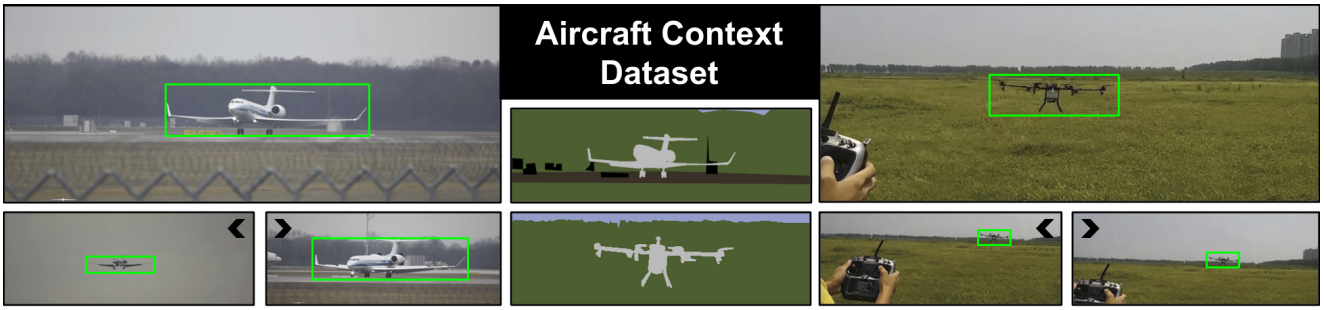
The Aircraft Context Dataset: Understanding and Optimizing Data Variability in Aerial Domains

Daniel Steininger Verena Widhalm Julia Simon Andreas Kriegler Christoph Sulzbacher
AIT Austrian Institute of Technology
{firstname.surname}@ait.ac.at



clear **atmosphere**, runway and vegetation **context,** diffuse **lighting**, not **occluded**, normal **exposure**, medium **size**, medium **weight**, business jet **category**, Gulfstream Aerospace G650 **model type.**

clear **atmosphere**, vegetation **context**, sunny **lighting**, not **occluded**, normal **exposure**, small **size**, miniature **weight**, agriculture **task**, semi-pro copter **category**, Joyance Tech 10L-606 **model type.**

Abstract

Despite their increasing demand for assistant and autonomous systems, the recent shift towards data-driven approaches has hardly reached aerial domains, partly due to a lack of specific training and test data. We introduce the Aircraft Context Dataset, a composition of two intercompatible large-scale and versatile image datasets focusing on manned aircraft and UAVs, respectively. In addition to fine-grained annotations for multiple learning tasks, we define and apply a set of relevant meta-parameters and showcase their potential to quantify dataset variability as well as the impact of environmental conditions on model performance. Baseline experiments are conducted for detection, classification and semantic labeling on multiple dataset variants. Their evaluation clearly shows that our contribution is an essential step towards overcoming the data gap and that the proposed variability concept significantly increases the efficiency of specializing models as well as continuously and purposefully extending the dataset.

1. Introduction

Detection and classification of aerial objects represents an increasingly relevant research area with the aim to extend autonomy in surveillance and airport scenarios involving various aircraft types ranging from large transporters to small-scale unmanned vehicles. The application of learning-based approaches has already been successfully demonstrated in similar areas such as the ADAS domain [4, 21, 36] indicating their potential in the aerial domain. However, these methods require large amounts of training and test data specific to the use case. While extensive datasets are available for other domains as well as arbitrary collections of object categories [15, 8], there are no comparable datasets focusing on aerial vehicles. The few initial datasets existing in this area [45, 34, 6] provide neither sufficient numbers of samples nor the required variability regarding aircraft types and recording conditions. To mitigate this gap, we propose a novel dataset focusing on a wide range of aerial vehicles and their environmental context with the aim of facilitating robust scene understanding. For this purpose, we provide a versatile set of annotations enabling fine-grained classification, detection and semantic labeling, as well as the basis for learning and evaluating pose-estimation and multi-object tracking algorithms.

In short, this work contains the following contributions:

- We provide an extensive novel dataset focusing on aerial objects along with fine-grained annotations for multiple learning tasks. A special emphasis is placed on covering a wide range of aircraft types includ-

ing variants underrepresented in currently available datasets.[1]

- We additionally present and apply a concept for meta-annotation to measure dataset variability and the influence of environmental conditions on model performance. We demonstrate their potential by training and evaluating baseline models on multiple dataset variants.

2. Related work

Aircraft represent a common class in several existing datasets, ranging from the first notable contributions in the form of MSRC [26] and PascalVOC [8] to more recent datasets such as PascalContext [44], MS COCO [15], DOTA [35] and ADE20k [45]. However, since most published datasets typically include only a low number of images containing aerial vehicles mixed with many out-of-context categories and focus their annotation efforts on a specific learning task, they do not provide sufficient image data for training versatile models specialized for aerial applications.

Datasets for manned aircraft. One of the first datasets focusing exclusively on aircraft is the FGVC Aircraft dataset [17] containing 10k extracted patches equally distributed across 102 different aircraft models. Real and semi-synthetic satellite images of aircraft are available in the MTARSI [34] and RarePlanes [25] datasets. However, due to a limited number of annotated models and the inherent top-view perspective, they are not suitable for ground-sensing applications.

Datasets for unmanned aircraft. Contrary to the datasets described above, the term UAV dataset holds a certain degree of ambiguity, since it often refers to images captured by instead of showing the unmanned aerial vehicles and therefore not relevant for this work [18, 3]. Available datasets in the latter category range from small collections focusing on the detection of UAVs as a single class [31, 6] to larger ones also applicable for tracking such as the Amateur Unmanned Air Vehicle Detection Drone Dataset [1]. One of the most recent contributions is made by [28] combining RGB and IR data and focusing on the detection and tracking of UAVs in low-resolution images. Furthermore, UAVData [43] introduces a larger-scale dataset which differentiates between 6 types of UAVs and a balloon class but is not yet publicly available.

While the number of datasets for manned and unmanned aircraft increased over the last years, there are still few combining both. One of the few exceptions is presented by [23] and provides 40 sequences divided equally between the two categories and containing a total of 12k annotated instances.

[1]Annotations are available for academic research at https://github.com/aircraftcontext/aircraft-context-dataset

Meta-annotation and robustness evaluation. Since the application of deep neural networks is moving towards real-time and safety-critical applications, a high amount of research effort is directed at evaluating their robustness and generalization ability under real-world conditions. While many recent publications focus on both the construction of [37, 33] and defense against [16, 41, 22] adversarial attacks, there is also activity focusing on environmental factors such as weather conditions [20, 30] and time-of-day variation [5], as well as various kinds of image corruptions [11, 27]. Whereas these works mainly focus on benchmarking trained models, there are also algorithmic evaluations on the level of dataset coverage [29, 2, 19]. However, their focus is usually on evaluating existing datasets, as opposed to incorporating relevant meta-information into the dataset itself. Datasets that do provide a form of meta-annotation generally focus on attributes of depicted objects [25, 32] applicable for scene-understanding and zero-shot learning, or provide only a limited set of coarse parameters concerning basic properties such as occlusion, truncation or grouping [15, 8, 12]. An incorporation of meta-parameters similar to our approach is used in the WildDash dataset [42]. However, it is focused on semantic labeling in the ADAS domain and mainly intended for benchmarking and testing according to visual hazards.

3. The Aircraft Context Dataset

The introduced Aircraft Context Dataset (*AC*) provides an extensive collection of images and annotations for multiple learning tasks in the aerial domain with a special emphasis on data variability and the inclusion of currently underrepresented aircraft types. Conceptually, it consists of two specialized subsets covering different target objects but sharing a consistent annotation policy and a unified label space. Despite their distinct characteristics, this approach facilitates a seamless integration of arbitrary classes and meta-annotations from both subsets for training and benchmarking. On the one hand, the Manned Aerial Vehicle (*MAV*) subset is focused on a wide selection of aircraft ranging from small utility planes and helicopters to large-scale passenger and transport jets covering both civilian and military, as well as jet-propelled and propeller vehicles. The Unmanned Aerial Vehicle (*UAV*) subset, on the other hand, is specialized in remotely controlled and autonomous aerial vehicles of all scales including copter and fixed-wing, as well as amateur and semi-professional variants. The object instances of each subset are assigned to the categories depicted in Figures 1 and 2, respectively, and enriched with meta-annotations for context and environmental conditions.

Due to the required extent and versatility of the dataset, a well-structured and consistent line of action was essential throughout the entire processing chain, ranging from data-harvesting activities over multiple annotation stages up to

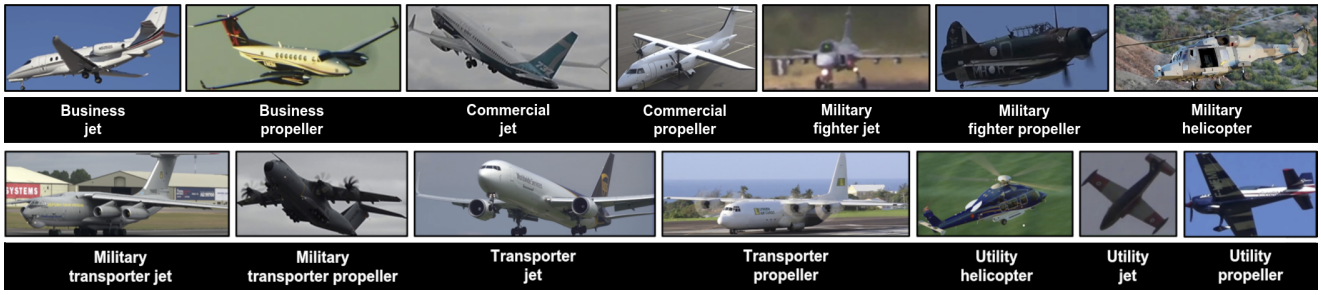


Figure 1. Iconic samples for the categories included in the *MAV* subset.

Figure 2. Iconic samples for the categories included in the *UAV* subset.

the experiments conducted for benchmarking purposes, as detailed in the following subsections.

3.1. Harvesting and meta-annotation

Reaching sufficient data quantity for novel datasets often requires extensive capturing sessions and is typically a resource-intensive and tedious process. However, the aerial domain offers the advantage of a large active community of aircraft-spotting and UAV enthusiasts creating publicly available and often high-quality content suitable for the purpose of this work. By building on this pool of input data, we were able to dedicate our resources mainly to selecting the most relevant samples and creating multi-modal annotations for them, which will be in turn shared with the community for application and further development.

Since some of the learning tasks targeted as future extensions such as multi-object tracking require continuous image sequences as opposed to single frames, we chose YouTube [38] as the primary source of our harvesting activity. From the wide range of available video content for the given domain recorded at diverse geographical locations, the annotation input is carefully selected to cover a high variability regarding aspects such as aircraft types, environmental and lighting conditions, as well as camera and compression effects. A number of sequences depending on video length and relevance is extracted and the visible aircraft are assigned one of a standardized and continuously expanded set of models consisting of manufacturer and type. The resulting list of aircraft models is furthermore enriched with meta-parameters such as length, wingspan and weight.

Finally, the aircraft models are clustered into consistent sets of super-classes. The 14 categories of the *MAV* subset displayed in Figure 1 distinguish between different permutations of application domains (*business*, *commercial*, *military*, *transporter*, *utility*) and propulsion systems (*helicopter*, *jet*, *propeller*). To define the 9 categories of the *UAV* subset visible in Figure 2, the same approach is applied for the respective domains (*amateur*, *military*, *semi-pro*) and available propulsion systems (*copter*, *fixed-wing*, *helicopter*). Additionally, the *UAV* subset includes meta-annotations for application tasks such as *agriculture*, *film/photography* and *racing*.

3.2. Annotation

As shown in Table 1, more than 4k video clips were selected and roughly cut into about 14k sequences based on changes in aircraft type, video transitions or to remove sections which are highly redundant, inappropriate or out-of-scope.

	MAV	UAV	AC
Video clips	2 107	2 153	4 260
Image sequences	8 994	4 843	13 837
Total frames	13.8M	4.6M	18.4M
Extracted frames	32 292	15 078	47 370
Annotated frames	28 788	13 279	42 067
Object instances	37 442	13 834	51 276
Semantic masks	1 669	2 234	3 903

Table 1. Overview of samples and annotations of the *MAV* and *UAV* subsets as well as the entire *AC* dataset.

Of the 18.4M frames extracted of these sequences, we selected 47k containing at least one aerial vehicle. Although the number of input videos is similar for both subsets, twice the number of relevant sequences could be extracted for *MAV* due to their recording characteristics. While sequences provided by aircraft spotters usually focus on recordings of the aircraft itself and often even contain multiple models per video, those containing UAVs are typi-

cally not as structured and include more irrelevant sections. A minimum of two frames per sequence is extracted and increased by additional samples for underrepresented classes to counteract class-imbalance effects. Finally, bounding boxes are created for all visible object instances, as well as semantic masks for a balanced subset of them regarding sequence coverage and class distribution. For the latter annotations, bounding-box sizes are doubled along both dimensions to improve the learning of contextual relationships while keeping the manual annotation effort in a reasonable range. Poly-line masks are drawn for the resulting image patches separating the *aircraft* from the defined context classes *apron/runway*, *building*, *indoor*, *sky*, *vegetation*, *water* and *out-of-scope*.

In total, the final dataset contains 42k images, along with 51k annotated object instances applicable for detection and classification as well as almost 4k semantic masks. Each sample is furthermore enriched with a set of relevant variability parameters to enable a quantitative analysis of dataset properties and benchmark their influence on model performance:

Airborne differentiates between aircraft in flight and on the ground. An aircraft is considered airborne, if none of its parts touches the ground or is held by a person in case of small UAVs.

Atmosphere describes weather conditions influencing object appearance by differentiating between *clear*, *fog*, *rain* and *snow*.

Context is assigned as up to two classes occurring in the immediate background. Available context types are *building*, *indoor*, *sky clear*, *sky cloudy* and *vegetation*, as well as *undefined* for backgrounds not corresponding to any available category. In addition the *MAV* dataset includes the context types *apron* and *runway* to depict specific backgrounds at airports. As opposed to the similar classes used for the more detailed semantic masks, these serve as efficient context cues for dataset analysis and model evaluation.

Degradation refers to the degree of visible image degradation caused by effects such as compression, blur and sensor noise, coarsely categorized as *none*, *low* and *high*.

Exposure separates well-lit object instances from those lacking detail due to the effects of camera *under-* or *overexposure*.

Lighting is used to differentiate between *sunny* and *diffuse* lighting conditions, resulting in either hard or soft shadows and distinct or soft specular highlights, respectively.

Occluded is defined as a binary state indicating if at least 15% of the object is occluded by other objects or truncated at the image border.

Size represents the maximum dimension considering object length and wingspan coarsely clustered into the categories *very small*, *small*, *medium*, *large* and *jumbo* based on commonly used standards [10]. This parameter can be inferred from the model type assigned to the object.

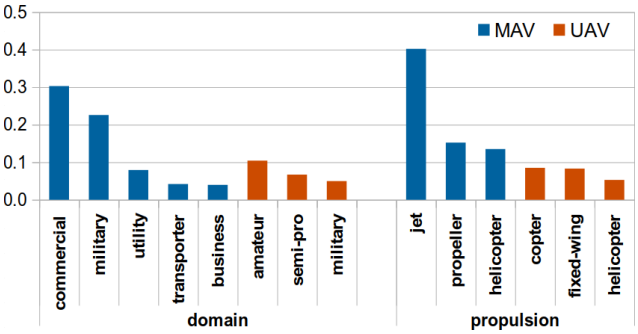


Figure 3. Instance distributions of domains and propulsion types included in the *MAV* and *UAV* subsets.

Weight is defined analogously as one of the categories *light*, *medium*, *heavy* and *super heavy* for *MAV* and *micro*, *miniature* and *heavy* for *UAV*.

Furthermore, the distinctions between aircraft **Type**, **Propulsion** and **Domain** as well as **Task** for UAVs defined during harvesting and meta-annotation (Section 3.1) are incorporated as additional variability parameters.

After image-level meta-annotations are defined during the harvesting stage, all instance annotations are created by our in-house annotation team using the open-source tool Scalabel [24]. As a first step, bounding boxes and variability parameters are assigned for each selected image, which are then used as an input for the creation of semantic masks. Finally, a cross-validation is performed between annotators.

3.3. Dataset analysis

To measure the applicability of our dataset to a broad range of real-world scenarios, an extensive statistical analysis concerning the distribution of both object categories and variability parameters is conducted for both subsets. Figure 3 provides the distribution of instances for each domain and propulsion type. While most categories are sufficiently represented, there is a comparatively high ratio of commercial jets, which directly relates to their overall frequency of occurrence, especially at sites frequented by the aircraft-spotting community. Combinations such as business propeller and military copter, on the other hand, are not yet present in sufficient numbers to robustly classify them as separate categories. This distribution is used to define balanced sets of super-classes for the dataset-ablation experiments presented in Section 4.1. Furthermore, the statistics are iteratively updated and used as a hint for expanding the dataset. For example, in the course of the harvesting stage, we started to specifically include further instances of helicopter variants to counteract their potential under-representation leading to their current frequency in the dataset.

In addition to the target categories, the distribution of

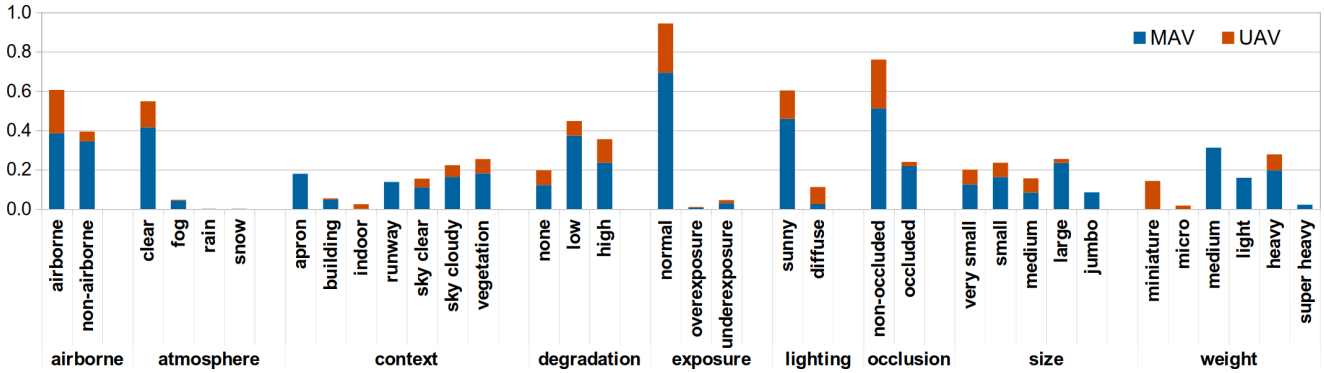


Figure 4. Instance distribution by variability parameter of the *AC* dataset.

variability parameters displayed in Figure 4 gives further insights into the dataset characteristics to be exploited for testing and subsequently expanding the dataset with specific types of image samples. Currently, most parameters are sufficiently represented and balanced to facilitate a robust analysis of their impact on model performance, including a good coverage of the full range of aircraft sizes and weights. One of the exceptions is presented by the atmospheric effects relating to *rain* and *snow*, which are underrepresented and can therefore not yet be used for further analysis. A significant advantage of the current dataset is the overall balancing of relevant context classes allowing a specific evaluation and tailoring of models for a given application scenario.

Furthermore, beside obvious differences in object sizes and weights, *MAV* and *UAV* differ in additional characteristics mainly caused by their typical recording scenarios. Since manned aircraft are more frequently captured in apron areas, they tend to be more often occluded by other aircraft and buildings and less frequently in an airborne state compared to UAVs. Moreover, a slightly higher image quality is visible in the exposure and degradation parameters for the *MAV* dataset, which can be explained by the usually higher quality of the used recording equipment. Furthermore, the object instances of this subset have a strong tendency towards the image center while *UAV* shows a more balanced distribution of both bounding box positions and sizes. As shown in Figure 5, objects depicted in *MAV* furthermore tend to be larger and more rectangular in shape. Accumulations on the right and top borders of both plots indicate the occurrence of aircraft concurrently truncated at opposite image borders.

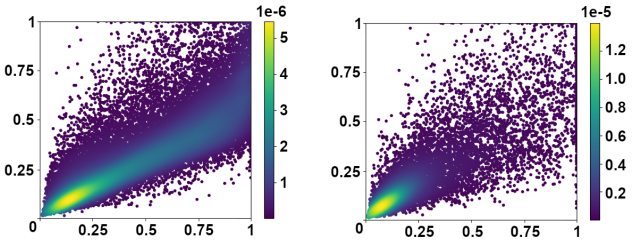


Figure 5. Density distributions of instance widths and heights for *MAV* (left) and *UAV* (right) subsets. Sizes are normalized to the corresponding image resolutions.

4. Methodology

To showcase the relevance and versatility of our proposed dataset, we conduct multiple machine-learning experiments. A selection of dataset variants is derived and used to train models for classification and object detection, which will be evaluated thoroughly and combined accordingly in Section 5. Furthermore, initial experiments are conducted for the task of semantic labeling. All models are trained and evaluated on a system with four NVIDIA RTX 2080 GPUs. To ensure reproducibility, we provide a detailed description of the methodology applied throughout the process in the following subsections.

4.1. Dataset ablation

To demonstrate the flexibility of the Aerial Context Dataset, several dataset variants are derived based on the categorization and variability parameters described in Section 3.2, and their subsequent analysis presented in Section 3.3. Each variant represents a restructured version of the available data, merging defined sets of classes into superclasses according to their common properties. The most fine-grained dataset used in this work (AC_{Fine23}) is structured into the full set of 23 object categories displayed in Figures 1 and 2. Merging the categories results in the most coarse-grained variant ($AC_{Coarse1}$) which simply distinguishes between the presence and absence of visible aerial vehicles of any kind, as well as a slightly more differentiated version distinguishing manned aircraft from UAVs ($AC_{Coarse2}$). Similarly, there are fine-grained variants for the two subsets (MAV_{Fine14}, UAV_{Fine9}) along with their coarse-grained counterparts ($MAV_{Coarse1}$, $UAV_{Coarse1}$). Additionally, to showcase the potential of specializing the dataset for

a specific application, three additional variants are derived for each subset based on selected variability parameters ($MAV_{Domain5}$, $UAV_{Domain3}$, MAV_{Prop3}, UAV_{Prop3}, $MAV_{Airborne2}$, $UAV_{Airborne2}$, $AC_{Airborne2}$) splitting the annotation into super-classes according to the values available for the respective parameter. While these dataset variants are selected due to their relevance for real-world applications, an analogous approach could be applied for any of the presented variability parameters. The resulting variants are used for both the detection and classification experiments described in the following sections, with the addition of a negative class in the latter case. For the semantic labeling experiments, we use a reduced set of the available semantic context categories for both subsets defined in Section 3.2.

4.2. Learning tasks

By defining and consistently using a unified format for all annotations and variability parameters, we ensure maximum data compatibility and facilitate a seamless combination with multiple learning tasks. We selected network architectures for classification, detection and semantic labeling to provide reasonable trade-offs between computational efficiency and accuracy. For all experiments, we use random splits of 80:10:10 between training, validation and test data on a per-sequence level to ensure the frames extracted from a sequence are exclusive to a single set. Furthermore, identical sets are used for classification and detection to preserve independence between training and test data when jointly evaluating both tasks. Standard randomized data augmentation techniques such as horizontal flipping, resizing and cropping are applied to each learning sample. We use stochastic-gradient-descent optimization and adapt the batch size to fit the available GPU memory. Classification and detection experiments typically converge after 15 to 35 training epochs depending on dataset granularity and size, while training semantic labeling takes up to 60 epochs.

Fine-grained classification. For this task, all dataset variants are extended by a negative class consisting of about 7k randomly extracted image regions of the original dataset with a minimum size of 128x128 pixels not containing any visible aerial vehicles. To counter-act an imbalance of object context in the negative samples, the original input images for this process are limited to 100 exclusive samples of the most frequent classes *sky clear* and *sky cloudy*. The trained models are based on Dilated Residual Network [39] with an input size of 256x256 pixels and an initial learning rate of 0.01 reduced by a factor of ten every ten epochs. In addition to the data augmentation techniques described above, we apply Gaussian blur to 10% of the samples, which increases mean precision by up to 0.3% for coarse and 4% for fine-grained dataset variants compared to the application on raw data.

Object detection. Experiments are conducted using a RetinaNet [13] implementation based on Feature Pyramid Network [14] combined with a pre-trained ResNet50 [9] classification backbone. Models are trained for all dataset variants presented in Section 4.1 with an input size of 1280x720 pixels and a fixed learning rate of 10^{-5}.

Semantic labeling. The experiments are based on Deep Layer Aggregation [40]. Since there is an insufficient number of samples for the context classes *water* and *indoor*, we omitted them in the selected dataset variants, thereby reducing the list of target labels for both datasets to *aircraft*, *apron/runway*, *building*, *sky* and *vegetation*. Each input patch is augmented as defined above, but resized to 1344x704 pixels and then randomly cropped to a size of 576x576 pixels. We reduce the initial learning rate of 0.05 by a factor of ten every 20 epochs.

5. Evaluation

This chapter presents a detailed analysis of the experiments described in chapter 4. As a first step, all classification and detection models are evaluated separately using established metrics in order to identify opportunities for efficiently combining tasks and dataset variants. Exemplary combinations are then selected and analyzed in more detail to quantify the influence of variability parameters on model performance. The chapter is concluded by preliminary results of the semantic-segmentation models showcasing the potential of future dataset extensions to facilitate a holistic scene understanding.

5.1. Metrics

Classification experiments are evaluated based on the established metric of F1-Score as the harmonic mean of precision and recall. To mitigate the inherent bias towards strongly represented classes, we compute the mean F1-Score across all available object categories including the negative class. For object detection we use mean Average Precision (mAP) as described by [15] for comparability with the COCO benchmark. Semantic labeling is evaluated using mean Intersection over Union (mIoU) defined by [7].

5.2. Classification and detection results

An overview of evaluation results for classification and detection experiments is presented in Table 2. In addition to standalone classification results, detection is evaluated for pure localization (*LOC*) by treating all available classes as a single target, as well as in combination with the corresponding classification (*CLS*) modules. The subscript *Int* refers to the internal backbone of the detector, while *Ext* combines each externally trained classification model with the instances localized by the according *Coarse1* variant.

	CLS_{Ext}	DET_{Int} LOC	DET_{Int} CLS	DET_{Ext} LOC	DET_{Ext} CLS
MAV_{Fine14}	.739	.864	.520	**.886**	**.622**
$MAV_{Domain5}$	.806	.876	.646	**.882**	**.660**
MAV_{Prop3}	.955	**.897**	.838	.890	**.848**
MAV_{Air2}	.944	.800	.755	**.896**	**.819**
$MAV_{Coarse1}$	.986	**.917**	**.917**	.893	.893
UAV_{Fine9}	.715	.673	.345	**.774**	**.452**
$UAV_{Domain3}$	.773	**.808**	**.506**	.769	.445
UAV_{Prop3}	.925	.762	.703	**.785**	**.703**
UAV_{Air2}	.878	.719	.532	**.779**	**.542**
$UAV_{Coarse1}$	.986	**.804**	**.804**	.777	.777
AC_{Fine23}	.697	.803	.434	**.848**	**.514**
AC_{Air2}	.938	.767	.693	**.842**	**.737**
$AC_{Coarse2}$	.926	.838	.730	**.849**	**.741**
$AC_{Coarse1}$	.996	**.867**	**.867**	.844	.844

Table 2. Results of models trained on all dataset variants for corresponding test sets: standalone classification as *mean F1-Score* (first column) and detection as *mAP*. DET_{Int} denotes detection by internal classification backbone, DET_{Ext} coarse internal localization combined with external classification (*LOC*: class-agnostic localization, *CLS*: classification of localized objects).

Not surprisingly, classification performance significantly increases as the number of target classes decreases throughout the majority of experiments. Pure localization performance, on the other hand, shows less variation since it is, in theory, independent of the number of classes. However, coarser variants still show better localization results as well, which can be explained by their entire capacity being focused on this task instead of additionally learning fine-grained differences between similar objects.

The results confirm that *UAV* represents the more challenging of the two subsets due to the properties discussed in Section 3.3 including smaller object sizes, higher data variability and lower image quality. Furthermore, classifying the airborne state proves to be a more difficult task than distinguishing propulsion types since the latter is based on distinct object appearances as opposed to background variations with a higher number of border cases.

Overall, the best results are achieved by building upon the localization of a coarse detection module and combining it with a more fine-grained classification. This furthermore presents the opportunity to apply a single detection model and switch or even simultaneously combine multiple classification models according to the use case, as shown in Figure 6. Moreover, by using an external classifier trained with a negative class, false positives generated by the localization step can be mitigated using this setup, which can be exploited by using a more sensitive detector to achieve a higher overall recall without sacrificing precision.

5.3. Impact analysis of data variability

In addition to the overall evaluation presented in the previous section, a more thorough understanding of model robustness under varying environmental conditions can be achieved by incorporating the variability parameters described in section 3.2. Filtering the test sets to exclusively include instances matching the evaluated criteria, yields parameter-specific *mAP* values. Table 3 summarizes the results of the configuration denoted as DET_{Ext} in the previous section. It includes all variability parameters sufficiently represented and not inherently biased by the definition of target classes. The depicted values are offsets in *mAP* to the overall performance per dataset variant presented in Table 2 and averaged per subset in the case of classification.

Some of the presented parameters can be influenced by modifying the physical setup of the recording system, while others are caused by environmental conditions and can only be tackled by adapting the models or training data. The strongest impact is visible for parameters in the former category including image degradation and occlusion which can be at least partially influenced by re-positioning the camera viewpoint to avoid static objects and buildings in the line of sight. By quantifying the model's sensitivity to these parameters, we can derive a strong priority for using high-quality image input in potential applications, since model performance significantly improves with lower degradation. The influence analysis of atmospheric effects and lighting expectably shows best performance in clear and sunny conditions. The positive impact of the *airborne* state, on the other hand, is not as intuitive, but can be explained by the predominance of more distinct backgrounds such as sky and vegetation. For the parameter of object context, values are generally relatively high partially resulting from the dominant ratio of samples in the *apron* and *undefined* categories not included in the analysis. Nevertheless, a clear trend is visible towards sky and vegetation backgrounds, while buildings represents the most challenging context due to their occasionally similar appearance to the target classes.

The impact analysis quantifies model robustness against specific environmental factors and recording setups and therefore provides a valuable basis for selecting models according to application requirements and deriving boundaries for the conditions under which they can be expected to work reliably. Furthermore, the results provide additional cues for specifically sampling training data, as well as defining the focus of future dataset extensions. For example, while models can always be expected to perform worse under foggy conditions than in clear atmosphere, the impact can be mitigated by either oversampling the according instances during training or increasing their ratio in the overall dataset by specific harvesting and annotation.

	State		Atmo		Object context				Degradation			Lighting		Occlusion	
	ar	*nar*	*cla*	*fog*	*clr*	*cld*	*veg*	*bld*	*ndg*	*ldg*	*hdg*	*sun*	*dif*	*noc*	*oc*
MAV_{LOC}	4.2	-4.3	3.2	-4.8	**6.3**	5.2	2.2	0.9	6.1	2.1	**-13.6**	1.1	0.0	5.1	-3.0
UAV_{LOC}	2.1	-10.7	1.0	7.9	7.1	**8.8**	0.2	-0.9	-2.8	7.8	-4.1	1.2	-0.4	2.0	**-18.3**
AC_{LOC}	5.0	-5.1	4.0	-0.8	**10.1**	8.1	1.2	-0.8	8.1	5.9	**-10.3**	3.0	0.0	5.1	-3.1
MAV_{CLS}	1.6	-2.4	3.8	-6.8	2.1	0.6	0.9	2.2	**5.4**	1.6	**-17.0**	1.5	0.3	5.2	-8.2
UAV_{CLS}	1.7	-7.9	0.0	-5.3	3.6	5.8	-2.5	-4.0	-2.1	**5.9**	-3.4	0.4	-0.2	1.8	**-21.8**
AC_{CLS}	3.3	-4.9	4.8	-4.9	3.0	2.1	1.0	-5.1	**7.6**	3.5	**-14.1**	3.1	-0.8	4.6	-9.0

Table 3. Influence of variability parameters on model performance as absolute *mAP* variation for localization (*LOC*) and averaged over external classification experiments (*CLS*): airborne (*ar*) and non-airborne (*nar*) state, clear (*cla*) and foggy (*fog*) atmosphere, sky-clear (*clr*), sky-cloudy (*cld*), vegetation (*veg*) and building (*bld*) object context, no (*ndg*), low (*ldg*) and high (*hdg*) image degradation, sunny (*sun*) and diffuse (*dif*) lighting, non-occluded (*noc*) and occluded (*oc*) object.

5.4. Semantic labeling

We initially trained models including the five classes described in section 4.2, achieving acceptable results for the classes *aircraft*, *sky* and *vegetation*, while the mIoU of *building* and *apron/runway* reached averages of 0.160 and 0.574, respectively. Therefore, we combined all categories except the former three into a background class, which can additionally be used as a fallback class for image areas not assignable to one of the dataset categories. Evaluation results for experiments conducted on these dataset variants (MAV_{Seg3} *and* UAV_{Seg3}) are summarized in Table 4.

	Aircraft	Sky	Veg	Bg	Overall
MAV_{Seg3}	.750	.948	.798	.350	**.712**
UAV_{Seg3}	.669	.878	.673	.527	**.687**

Table 4. Per-class and overall semantic labeling results (*mIoU*) on selected dataset variants.

The current state of the dataset already provides a sufficient basis for robustly differentiating between aircraft, sky and vegetation for common use cases, while the remaining classes should be specifically targeted during future extensions. The slightly lower labeling quality of the *UAV* variant is mainly due to a higher inherent diversity of appearance regarding aircraft and their surroundings, whereas MAVs are more often captured in the structured environment of airports. However, both subsets show overall promising initial results, which are qualitatively depicted in Figure 6.

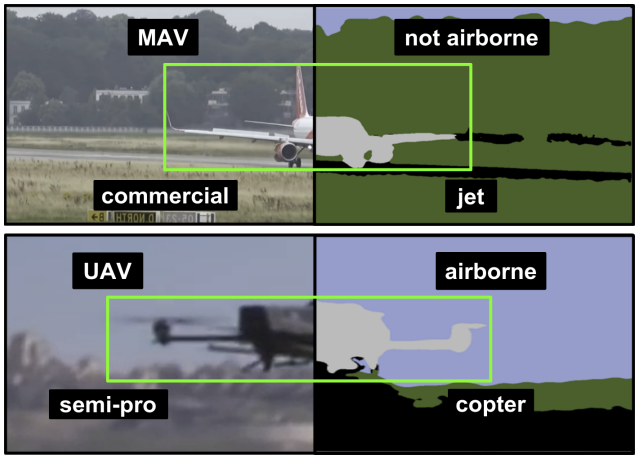


Figure 6. Qualitative results for classification, detection and semantic labeling on *MAV* (top) and *UAV* (bottom) subsets.

6. Conclusion

We introduced the Aerial Context Dataset, an extensive collection of image data and multi-modal annotations for manned aircraft and UAVs along with a rich set of variability parameters. To demonstrate the potential of our dataset and variability concept, baseline models were trained on multiple dataset variants and thoroughly evaluated. Special emphasis was directed at exploiting the variability annotations to evaluate model performance under varying environmental conditions. Thereby, we were able to demonstrate how useful insights for both optimizing training input and further extending the dataset can be extracted with comparatively little additional annotation effort. Overall, we conclude that the dataset represents a vital step towards closing the data gap in the aerial domain. As a next step, we plan to use the insights gained during this work to coherently extend the dataset and mitigate any identified gaps, as well as to apply the variability concept to other domains. In the long term, we plan to complement the dataset with annotations for pose estimation and multi-object tracking, which are already considered in the design. In addition to exploiting the variability parameters for dataset and model analysis, another promising research direction would be their reformulation as a separate classification task to increase efficiency in dataset creation and improve failure awareness.

Acknowledgement. We would like to thank our annotation team consisting of Vanessa Klugsberger, Marlene Glawischnig and Gulnar Bakytzhan.

References

[1] Mehmet Çağrı Aksoy, Alp Sezer Orak, Hasan Mertcan Özkan, and Bilgin Selimoğlu. Drone dataset: Amateur unmanned air vehicle detection. 10.17632/zcsj2g2m4c.4, 2019. Accessed: 2020-11-18. 2

[2] Abolfazl Asudeh, Zhongjun Jin, and HV Jagadish. Assessing and remedying coverage for a given dataset. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 554–565. IEEE, 2019. 2

[3] Ilker Bozcan and Erdal Kayacan. Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8504–8510. IEEE, 2020. 2

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[5] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3819–3824. IEEE, 2018. 2

[6] Drone dataset (uav). https://www.kaggle.com/dasmehdixtr/drone-dataset-uav, 2019. Accessed: 2020-11-18. 1, 2

[7] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 6

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1, 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[10] ICAO. Doc 4444–procedures for air navigation services: Air traffic management, 2007. 4

[11] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8828–8838, 2020. 2

[12] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, pages 1–26, 2020. 2

[13] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 6

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2, 6

[16] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2

[17] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 2

[18] Murari Mandal, Lav Kush Kumar, and Santosh Kumar Vipparthi. Mor-uav: A benchmark dataset and baselines for moving object recognition in uav videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2626–2635, 2020. 2

[19] Senthil Mani, Anush Sankaran, Srikanth Tamilselvam, and Akshay Sethi. Coverage testing of deep learning models using dataset characterization. *arXiv preprint arXiv:1911.07309*, 2019. 2

[20] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 2

[21] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4990–4999, 2017. 1

[22] Arash Rahnama, Andre T Nguyen, and Edward Raff. Robust design of deep neural networks against adversarial attacks based on lyapunov theory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8178–8187, 2020. 2

[23] Artem Rozantsev, Vincent Lepetit, and Pascal Fua. Flying objects detection from a single moving camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4128–4136, 2015. 2

[24] Scalabel open-source web annotation tool. https://scalabel.ai. Accessed: 2020-11-10. 4

[25] Jacob Shermeyer, Thomas Hossler, Adam Van Etten, Daniel Hogan, Ryan Lewis, and Daeil Kim. Rareplanes: Synthetic data takes flight. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 207–217, 2021. 2

[26] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *European conference on computer vision*, pages 1–15. Springer, 2006. 2

[27] Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin T Vechev. Fast and effective robustness certification. *NeurIPS*, 1(4):6, 2018. 2

[28] Fredrik Svanström, Cristofer Englund, and Fernando Alonso-Fernandez. Real-time drone detection and tracking with visible, thermal and acoustic sensors. *arXiv preprint arXiv:2007.07396*, 2020. 2

[29] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. 2

[30] Georg Volk, Stefan Müller, Alexander von Bernuth, Dennis Hospach, and Oliver Bringmann. Towards robust cnn-based object detection through augmentation with synthetic rain variations. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 285–292. IEEE, 2019. 2

[31] Matouš Vrba, Daniel Heřt, and Martin Saska. Onboard marker-less detection and localization of non-cooperating drones for their safe interception by an autonomous aerial system. *IEEE Robotics and Automation Letters*, 4(4):3402–3409, 2019. 2

[32] Peng Wang, Bingliang Jiao, Lu Yang, Yifei Yang, Shizhou Zhang, Wei Wei, and Yanning Zhang. Vehicle re-identification in aerial imagery: Dataset and approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[33] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020. 2

[34] Zhi-Ze Wu, Shou-Hong Wan, Xiao-Feng Wang, Ming Tan, Le Zou, Xin-Lu Li, and Yan Chen. A benchmark data set for aircraft type recognition from remote sensing images. *Applied Soft Computing*, 89:106132, 2020. 1, 2

[35] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018. 2

[36] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2174–2182, 2017. 1

[37] Bin Yan, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Cooling-shrinking attack: Blinding the tracker with imperceptible noises. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 990–999, 2020. 2

[38] Youtube. https://www.youtube.com, 2020. Accessed: 2020-11-30. 3

[39] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017. 6

[40] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018. 6

[41] Valentina Zantedeschi, Maria-Irina Nicolae, and Ambrish Rawat. Efficient defenses against adversarial attacks. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 39–49, 2017. 2

[42] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernandez Dominguez. Wilddash - creating hazard-aware benchmarks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[43] Yuni Zeng, Qianwen Duan, Xiangru Chen, Dezhong Peng, Yao Mao, and Ke Yang. Uavdata: A dataset for unmanned aerial vehicle detection. *Soft Computing*, pages 1–9, 2021. 2

[44] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018. 2

[45] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1, 2