FOUNDATION MODEL-BASED DATA SELECTION FOR DENSE PREDICTION TASKS

Niclas Popp [1, 2] ✉, **Dan Zhang**[1], **Jan Hendrik Metzen**[3, *], **Matthias Hein**[2], **Lukas Schott**[1]
[1]Bosch Center for Artificial Intelligence
[2]University of Tübingen
[3]IPAI Aleph Alpha Research
[*] Work done while at Bosch Center for Artificial Intelligence
{niclas.popp@de.bosch.com}

ABSTRACT

Data selection, the problem of selecting a small dataset to be labeled from a large unlabeled pool is an important practical problem. In particular, dense prediction tasks such as object detection and segmentation require high-quality labels at pixel level, which are particularly costly to obtain. We propose object-focused data selection (OFDS), which leverages object-level representations from foundation models to ensure that the selected image subsets semantically cover the target classes, including rare ones. We show that OFDS achieves state-of-the-art performance both for object detection and image segmentation with substantial improvements over all baselines in scenarios with imbalanced class distributions. Moreover, we demonstrate that pre-training with autolabels from foundation models on the full datasets before fine-tuning on human-labeled subsets selected by OFDS further enhances the final performance. Finally, OFDS consistently improves active learning methods when replacing the random selection of the initial labeled dataset, the so-called "cold start problem" of active learning, with OFDS.

1 INTRODUCTION

The performance of machine learning systems critically depends on the availability and quality of training data (Zha et al., 2023; Agnew et al., 2024). In vision tasks such as object detection or semantic segmentation, where pixel-level annotations are required, producing high-quality labels is time-consuming and costly. Per image, dense labeling can take between a few seconds for simple cases to over 90 minutes for complex scenes Lin et al. (2019). These demands make optimizing the annotation process under a limited budget a longstanding challenge in computer vision. Recent advances in open-world foundation models Liu et al. (2024); Ren et al. (2024); Ravi et al. (2024) have opened up new annotation and data selection possibilities. These models demonstrate strong zero-shot capabilities, allowing them to generalize across tasks without further fine-tuning. However, directly deploying these large models in resource-constrained applications, such as autonomous driving, is often impractical due to their size and computational demands. In this work, we focus on training small, task-specific models and leverage foundation models to guide data selection and reduce annotation costs for dense prediction tasks. This is also an important problem for active learning where the initial data selection is typically done randomly Chen et al. (2024); Nath et al. (2022); Samet et al. (2023).

Without any cost for human labeling, foundation models can provide machine generated annotations, so-called autolabels. This raises the question: *Can open-world foundation models eliminate the need for dense human annotations?* To investigate this, we introduce a calibration strategy for open-world object detection and segmentation models (Liu et al., 2024; Ravi et al., 2024) to generate reliable autolabels and examine the results on different datasets (Everingham et al., 2010; Cordts et al., 2016). We find that training on the full dataset with autolabels can outperform training on human-annotated subsets in highly budget-constrained scenarios for simple tasks. However, as annotation budgets or task complexity increase, human annotations are essential.

This leads to our second question. If the quality of autolabels is not sufficient to fully replace human labels, *can foundation models be used to improve the data selection for dense prediction tasks?* A reliable data selection method for real-world applications must effectively handle long-tailed class distributions, where rare classes only have a few instances in the unlabeled dataset. Selecting images

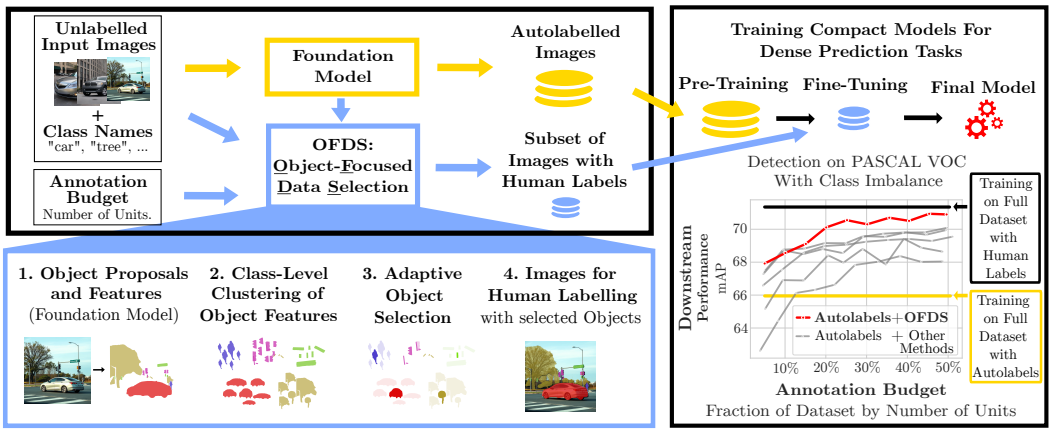

Figure 1: **Illustration of the Setup for Object-Focused Data Selection.** The figure describes our holistic setup for dealing with constrained annotation budgets for dense prediction tasks. We first use a foundation model to generate autolabels for the entire dataset and pre-train a model using them. Subsequently, we use OFDS to select a subset of the unlabeled pool of images to be annotated by humans. The blue box in the lower left summarizes the four main steps of OFDS. Its core advancement is to use object-level features to ensure a semantic covering of the target classes through the selected objects. The subset selected by OFDS is then used to fine-tune the model.

with objects from rare classes is essential to ensure that downstream models perform well on them. For this purpose, we propose object-focused data selection (OFDS). Our method assumes knowledge of the target class names and selects an initial subset of images from an entirely unannotated pool under the constraint of a fixed annotation budget. OFDS operates under a budget defined at the level of *objects* rather than *images*. This is motivated by the fact that the cost for annotations are typically charged by the number of annotated instances (for object detection) or masks (for semantic segmentation) also referred to as *units*. Unlike existing methods Xie et al. (2023); Li et al. (2023b); Sorscher et al. (2022); Chen et al. (2024), which rely on a fixed set of image-level feature vectors, OFDS uses foundation models to propose object-level feature vectors. This enables choosing representative objects and ensure that the selected subset semantically covers all classes, including rare ones. We illustrate data selection with OFDS in Figure 1.

To the best of our knowledge, data selection and autolabeling have so far only been considered independently. We argue that the most effective approach to utilize a constrained annotation budget is a combination of both. Therefore, we first pre-train the downstream models on the full dataset with autolabels and then fine-tune on human-labeled subsets constrained by a fixed annotation budget. Figure 1 illustrates the training setup and shows the performance for object detection on PASCAL VOC with class imbalance. The depicted results highlight that pre-training on the autolabeled dataset combined with fine-tuning on a human-annotated subset selected by OFDS leads to the best results.

Our main contributions summarize as follows:

1. **Introducing Object-Focused Data Selection (OFDS).** We present OFDS, a method that leverages object-level representations provided by foundation models to guide data selection. The core advancement of OFDS is to perform data selection with respect to semantic similarity on the level of objects rather than entire images like existing methods Xie et al. (2023); Li et al. (2023b). OFDS consistently improves the performance over all existing baselines due to its class-aware selection, in particular in scenarios with imbalanced class distributions.
2. **Holistic Training Strategy for Dealing With Constrained Annotation Budgets.** We rethink training compact models for dense downstream tasks under constrained annotation budgets by combining data selection with autolabels. Therefore, we pre-train the model on the entire dataset with autolabels and fine-tune on subsets with human labels selected through OFDS. This approach enhances performance compared to training solely with autolabels or only on human-labeled subsets as done in prior works on data selection.
3. **Improving Cold Start of Active Learning Methods.** Active learning methods commonly require a small initial labeled dataset (the "cold start problem"). We show that using OFDS instead of a random selection consistently improves active learning methods for semantic segmentation and object detection.

2 Related Work

Active Learning. Active learning is a long standing approach in machine learning that targets the problem of reducing the labeling cost Ren et al. (2021). There are three key distinctions between our approach and classical active learning. First, active data selection is specific to a single model training and the data selection process is carried out iteratively while training. OFDS is independent of the downstream model being trained and the selection is performed "passively" before training. Second, most common active learning methods for dense prediction tasks are specific to single tasks such as object detection Yang et al. (2024) *or* semantic segmentation Mittal et al. (2023); Hwang et al. (2023); Kim et al. (2024) while OFDS is agnostic to dense tasks. The third distinction is the so-called *cold start problem.* Active learning frameworks typically assume the presence of an initial labeled subset of data before selecting additional datapoints Nath et al. (2022); Samet et al. (2023); Chen et al. (2024). Our work focuses on data selection which can be carried out without any labeled images. This problem has been addressed specifically for 3D semantic segmentation by Nath et al. (2022) and Samet et al. (2023). To the best of our knowledge, HaCON (Chen et al., 2024) is the only task-agnostic method for the cold start problem which can be used comparably to OFDS.

Dataset Pre-Selection. Dataset pre-selection Li et al. (2023b) is a specific setting for dataset selection. In contrast to our approach, dataset pre-selection is performed without any knowledge about the downstream task or associated target classes. UP-DP Li et al. (2023b) uses a prompt learning approach together with multi-modal clustering to select images. In contrast, we argue that for dense prediction tasks the classes must be known for the annotation process to conform the selection to the annotation budget. Thus, we base our selection on the classes for the downstream task.

Coreset Selection. The goal of coreset selection is to select a subset of a large dataset to approximate the entire dataset. It is commonly used to reduce the training cost of a model by training on a subset but achieve performance as close to the full dataset as possible. This differs from data selection where the main goal is reducing the labeling cost. Furthermore, approaches for coreset selection Mirzasoleiman et al. (2020); Guo et al. (2022); Feldman (2020) typically require having the full dataset labeled or a model trained on a labeled subset.

Data and Coreset Selection Beyond Image Classification. So far, most approaches for data or coreset selection have been evaluated on image classification. However, the labeling costs for dense prediction tasks are higher which motivates specific approaches for these tasks. To the best of our knowledge, Yao et al. (2023), Zhou et al. (2024) and Lee et al. (2024) are the only works considering coreset selection for tasks beyond image classification. However, they require fully labeled datasets for their selection. USL Wang et al. (2022) considers the combination of selected data with annotations and unlabeled data for semi-supervised learning. Similarly, ReCo Shin et al. (2022) considers the selection of reference images from an unlabeled dataset but specifically targets co-segmentation. Li et al. (2023b) and Xie et al. (2023) included evaluations of their methods on dense prediction tasks but perform the selection using image-level representation which we observe to be inferior in the setting with class imbalance.

Pre-Training with Autolabels. Training small models for downstream tasks with autolabels from a foundation model can be viewed as training under weak or noisy supervision Zhang et al. (2023); Kim et al. (2024). This approach has been explored for specific domains such as remote sensing segmentation Zhang et al. (2024) or tasks such as local feature learning Wu et al. (2024). While these works focus purely on training with autolabels on entire datasets which are assumed to be pre-selected, we use autolabels for initial pre-training to enhance the data selection problem that subsequently considers which images to label.

3 Method: Object-Focused Data Selection

3.1 Strategy

The goal of Object-Focused Data Selection (OFDS) is to select a representative set of images to be labeled for dense prediction tasks given a large pool of unlabeled images, a fixed annotation budget and the target classes. OFDS leverages feature representations of individual objects with the aim of sampling a balanced and semantically diverse subset. Therefore, it is important to consider both inter- and intra-class diversity. To address this, we introduce a four stage selection process.

1. **Object Proposals and Feature Extraction:** For dense prediction tasks, single images can contain objects from multiple classes. Thus, frequent classes or the background can dominate image-

level features. To mitigate this, we adopt foundation models to extract features on object-level to guide the selection process.
2. **Class-level Clustering by Semantic Similarity:** To gain additional knowledge about intra-class semantic similarity, we cluster the object features within each class.
3. **Object Selection Through Adaptive Clustering:** To ensure intra-class semantic diversity in the selected subset, we choose representative objects from the identified clusters. To ensure inter-class balance, we adaptively set the number of clusters by evenly distributing the overall annotation budget between the classes.
4. **Exhaustive Image Annotation:** For images containing selected objects, all objects from the target categories are considered for human labeling. This ensures correct background information which is required by most common training setups for dense prediction tasks and provides additional information value.

The complete steps for OFDS are summarized by Algorithm 1. The selection strategy operates without randomness, which is beneficial in practical situations where only one selection round is possible and reliability is essential.

Algorithm 1 OFDS: Object-Focused Dataset Selection

Input: Set of unlabeled images $\{\mathbf{I}_1, \ldots, \mathbf{I}_N\}$,
Annotation budgets by number of units: B,
Estimated number of annotation units per image: N_O
Classes to label, sorted by ascending number of object proposals per class: $\{C_1, ..., C_M\}$
Output: Subset of images selected for labeling

Generate a set of object features and corresponding labels $\{(\mathbf{O}_j^{\mathbf{I}_i}, l_j^{\mathbf{I}_i})\}_{j=1}^{K_i}$ for every image $\mathbf{I}_i$ using the object proposer
Initialize the subset $\mathcal{S} = \{\}$
for $l \in \{1, \ldots, M\}$ **do**
    Select the object features predicted as class C_l by the object proposer: $\mathcal{D}_l = \{\mathbf{O}_{l_j}^{\mathbf{I}_i} | l_j = C_l\}$
    Determine the number of images to add for the current class: $N_{C_l} = \frac{B - N(\mathcal{S})}{(M-l+1)N_O}$ where $N(\mathcal{S})$ is the number of units annotated in $\mathcal{S}$
    Perform k-means clustering on $\mathcal{D}_l$ with adaptive k to feature N_{C_l} clusters without images from $\mathcal{S}$
    For every cluster without images from $\mathcal{S}$ select the object $\mathbf{O}_{j^*}^{\mathbf{I}_i^*}$ which is closest to the cluster mean
    Annotate the images with cluster medoids and update $\mathcal{S}$: $\mathcal{S} = \mathcal{S} \cup \{\mathbf{I}_i^*\}$
end for
Return $\mathcal{S}$

3.2 Detailed Steps

Object Proposals and Feature Extraction. In the first step of Algorithm 1, a foundation model - referred to as the *object proposer* - is used to detect which objects are present in an image. For this purpose, we use Grounding DINO Liu et al. (2024); Ren et al. (2024) as pre-trained open-world object detector. Given an image and a set of class names, it returns a set of object detections consisting of bounding boxes, labels and confidences. The bounding boxes are used as queries to SAM 2 Ravi et al. (2024) for generating object features. These models were chosen as they are state-of-the-art open-world detection and segmentation models performing well across various benchmarks Liu et al. (2024); Ren et al. (2024); Ravi et al. (2024). A critical aspect of using the object proposer is calibrating the confidence threshold that determines which object proposals to consider. For data selection, the important aspect is to only obtain reliable, high-quality predictions for class objects instead of noisy predictions for all potential objects. This requires controlling the precision of the object proposer. Thus, the confidence threshold is set on a reference dataset such that the false positive rate of object proposals is 5% which is a commonly used value Hendrycks & Gimpel (2017). Based on the object proposals, we construct the object features to provide semantic information for clustering similar objects. We leverage object pointers from the SAM2 memory bank, which contain high-level information of objects and are stored as 256-dimensional vectors.

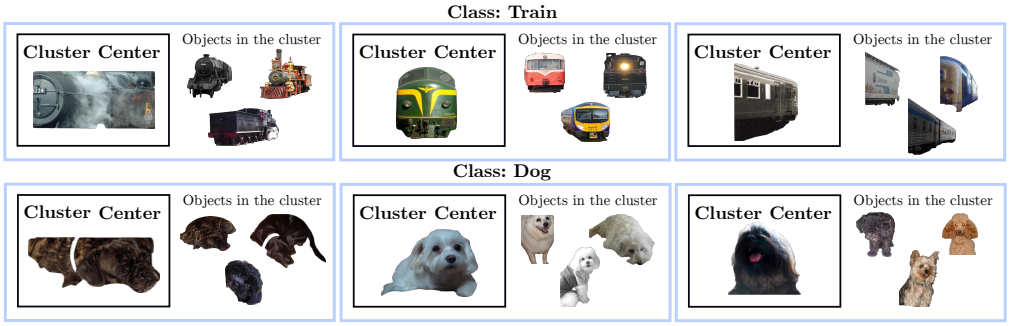

Figure 2: **Illustration of Clusters for Two Classes from PASCAL VOC.** Every object corresponds to one object feature generated by the object proposer as described in Section 3.2. The clustering was performed with $N_{C_l} = 30$ clusters per class. The clusters provide information about intra-class semantic similarity. By cluster center we refer to the datapoint closest to the cluster mean.

Class-level Clustering and Semantic Similarity. For clustering the object features, we perform k-means clustering on the set of object features from each target class individually. The use of k-means clustering is motivated by its effectiveness in the context of data selection for image-level classification Sorscher et al. (2022); Abbas et al. (2024) and the covering of the feature space that can be obtained from the cluster centers Pollard (1981). An illustration with example clusters is given in Figure 2. The resulting clusters provide information about intra-class similarity in addition to the inter-class information given through the object proposals.

Object Selection Through Adaptive Clustering. We select the objects closest to the cluster centers to get representative samples and choose at most one object per cluster to ensure semantic diversity. Like this, we construct a density-based covering for the semantic feature space of the individual classes. The number of clusters is chosen adaptively to accommodate for the overall annotation budget as well as already annotated images that were selected for previous classes. Initially, the annotation budget is evenly distributed between all classes and the class with the fewest object proposals is considered first. After every class, the leftover budget is evenly redistributed among the remaining classes. Importantly, this budget is measured in annotation units rather than images. Given a budget in units to annotate per class, the number of objects N_{C_l} to select is obtained by dividing this budget through an estimate for the number of units per image. We increase the number of clusters until there are N_{C_l} clusters without object features from already selected images. Only from these clusters, we select objects. This ensures that only one object per cluster is selected, providing a diverse semantic covering of the target classes and avoids the explicit selection of near duplicates. Steps 4 to 7 of Algorithm 1 summarize the object selection process.

Exhaustive Image Annotation. On images with selected objects, objects from all target classes are being labeled by the human annotators. Although this may initially seem contrary to the object-focused approach, exhaustive labeling yields information about the background of the selected images. This background information is required by most common training frameworks for dense prediction tasks, either as a separate class Zhao et al. (2017) or to ensure correct negative samples Xu et al. (2019). In particular, the negative samples sampled from the background are class-agnostic, which provide greater information value than class-specific negative samples that could show objects from other target classes. While there exist potential solutions for training models for dense downstream tasks with partial labeling Jain et al. (2022); Cour et al. (2011), we aim to ensure compatibility with standard setups.

4 Experiments

In this section, we outline the setup used to conduct our experiments and subsequently discuss the results in four parts. First, we discuss the model performance on downstream tasks when training purely with autolabels. Second, we conduct an extensive comparison of OFDS against existing data selection baselines across six distinct settings. Third, we highlight the advantage of pre-training on the full dataset with autolabels and fine-tuning with human annotations on selected subsets. Finally, we demonstrate that OFDS can be used to improve active learning by selecting the initial data to be labeled.

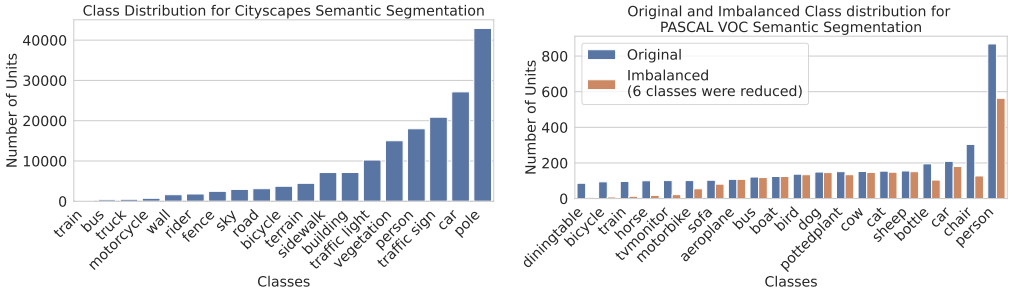

Figure 3: **Class Distributions for PASCAL VOC and Cityscapes for Semantic Segmentation.** As the distribution for PASCAL VOC is relatively balanced, we additionally construct a more realistic setting for data selection with class imbalance by pruning the six classes with the fewest objects. The class distribution for Cityscapes already features rare classes, such that we only consider the entire dataset.

4.1 Experimental Setup

Tasks and Datasets As downstream tasks, we consider object detection and semantic segmentation. For object detection we use the joint training set from the PASCAL VOC Everingham et al. (2010) 2007 and 2012 splits, evaluated on the validation set from 2012 as well as the Cityscapes Cordts et al. (2016) dataset with the classes featuring instance-level annotations. Similarly, for segmentation we consider the PASCAL VOC 2012 dataset and Cityscapes. The PASCAL VOC datasets feature a relatively balanced object distribution as shown in Figure 3. This can be attributed to the fact that the data set has already been selected and labeled by humans. However, unlabeled real-world datasets typically follow class imbalanced distributions Johnson & Khoshgoftaar (2019). Thus, we construct two additional settings with rare classes. We reduce the six smallest classes of the PASCAL VOC datasets by 99%, 95%, 85%, 80%, 75% and 50%. We refer to this setting as *class imbalanced*. The original and class imbalanced object distributions are shown in Figure 3. The class distribution of Cityscapes naturally contains rare classes such that we directly use the full dataset. Results on two additional datasets from specific domains can be found in Section C.

Models and Training Setup For our main experiments on object detection we use a Faster RCNN Ren et al. (2015) with ResNet-18 backbone He et al. (2011) and for semantic segmentation a Segmenter Strudel et al. (2021) with ViT-T backbone Dosovitskiy et al. (2021). Ablations with a Deformable DETR Zhu et al. (2021) for object detection and a PSP Net Zhao et al. (2017) for semantic segmentation can be found in the appendix. The backbones were pre-trained on ImageNet. In Sections 4.2 and 4.3 we train the decoder parts from scratch with the obtained human labels to evaluate the influence of the data selection. In Section 4.4 we initialize the model with the checkpoint obtained after pre-training with autolabels to improve the downstream performance. For every setting consisting of dataset and tasks we train for the same number of steps on all subsets. We use augmentations consistent with Xie et al. (2023). The complete hyperparameters can be found in the Section O. For the experiments with active learning we follow experimental setups from Yang et al. (2024); Mittal et al. (2023). Details can be found in Section F.

Baselines In this section we provide a comprehensive overview over the baselines for data selection.
FreeSel: Xie et al. (2023) introduced FreeSel as a method for data selection based on a single pass of the unlabeled dataset through DINO. The selection is based on local semantic features of images. Contrary to our approach, the number of features per image is a fixed hyperparameter independent of the number of objects on a specific image.
UP-DP: UP-DP Li et al. (2023b) performs data selection based on unsupervised prompt learning using vision-language models, in particular BLIP-2 Li et al. (2023a). Contrary to our method, UP-DP requires training. We compare to the UP-DP selection based on probabilities predicted by cluster-level head which exhibited the best performance in the original publication.
Prototypes: The current state-of-the art approach for unsupervised data selection for image classification was presented by Sorscher et al. (2022) for ImageNet Deng et al. (2009) and has since been scaled to webdatasets Abbas et al. (2024). The method consists of two steps. First, extract image features from an pre-trained model and perform k-means clustering using these features where k equals the number of classes. Subsequently, the datapoints closest to the cluster centers are selected. We compare to this protoype selection using image features from DINO-T Caron et al. (2021).
Coreset: By coreset selection we refer to the k-centers algorithm, introduced to the coreset problem

by Sener & Savarese (2018). The selection requires image features and starts with a random image as initial subset. The images with the greatest distance to the subset are then incrementally added to the subset. We compare to coreset selection using features from DINO-T and the $\mathcal{L}_2$ in the feature space to determine the newly added points.
HaCON: Chen et al. (2024) developed HaCON to address the cold start problem for active learning. Yet, the method can be applied to general data selection. HaCON clusters the features of a self-supervised model and selects the samples which are located at the cluster boundaries.
Random Selection The simplest baseline is the selection of a random subset from the unlabeled images.

4.2 Can Open-World Foundation Models Eliminate the Need for Dense Human Annotations?

To analyze the model performance on downstream tasks when training with autolabels, we use Grounding DINO-T and Grounding SAM2-T to annotate PASCAL VOC and Cityscapes. Abltations with larger foundation models can be found in Section B. These models are also used as object proposer in OFDS. Unlike for OFDS, the goal when generating autolabels is not to detect or segment objects with high precision but to balance precision and recall. Therefore, to generate autolabels we calibrate Grounding DINO by selecting the threshold that yields the highest F1 Score on reference datasets. These reference datasets are constructed by selecting subsets of MSCOCO Lin et al. (2014) consisting of relevant classes for PASCAL VOC or the streetscenes found in Cityscapes. Details on the two calibration approaches for autolabeling and OFDS can be found in Section J. The performance of downstream models trained using the resulting autolabels can be found in Figures 4 and 5 (horizontal lines). For PASCAL VOC, we observe that the performances of models trained with autolabels are comparable to a 30% random subset with human labels for object detection and a 20% subset for semantic segmentation. The performance of the model trained with autolabels achieves a mAP of 70.7 which outperforms the reported mAP of 55.7 achieved by the Grounding DINO-T model used to generate the autolabels. This can be attributed to the fact that Grounding DINO is a model for open-set object detection while the downstream models are trained for closed-set detection. For Cityscapes the models trained with autolabels perform substantially worse than even a 5% random subset with human labels. In summary, these findings indicate that for simpler datasets and very limited annotation budgets, autolabels can outperform training on human-annotated subsets. However, with increasing annotation budgets or more complex datasets, human annotations are indispensable.

4.3 Data Selection for Dense Prediction Tasks

In this section, we compare OFDS to six baselines in both class-imbalanced and balanced settings. We note that the imbalanced settings are more representative for real-world data selection. We train the decoder from scratch to investigate the influence of data selection alone and avoid confounding influences from pre-trained network weights.

Full PASCAL VOC. Figure 4 displays the results for the full PASCAL VOC dataset. As the class distribution is relatively balanced (see Figure 3), we observe that random selection serves as a strong baseline, with no other selection method achieving substantial improvements over it. FreeSel, HaCON and OFDS perform on par with random selection, while the prototype-based approach consistently yields the lowest performance. UP-DP and Coreset underperform compared to random selection. We attribute this to the fact that these methods rely on image-level representations and were developed for image classification while object detection and segmentation are multi-label tasks.

PASCAL VOC with Class Imbalance. The results for object detection and semantic segmentation on the PASCAL VOC datasets with class imbalance are shown in Figure 4. We observe that none of the existing baselines consistently outperform random selection. In contrast, our method outperforms all baselines, including random selection for both object detection and semantic segmentation. Notably, the difference to the baselines is largest when assessing the performance on the six rare classes. While there exist post-hoc approaches to adjust training setups to the presence of rare classes Dong et al. (2023); Tan et al. (2021); Wang et al. (2020; 2021a), OFDS targets the class imbalance problem already at the level of data selection.

Cityscapes. The results for object detection and semantic segmentation on Cityscapes are shown in Figure 5. We observe that OFDS, FreeSel and Coreset outperform random selection. As the dataset naturally has a more imbalanced class distribution than PASCAL VOC (see Figure 3), OFDS

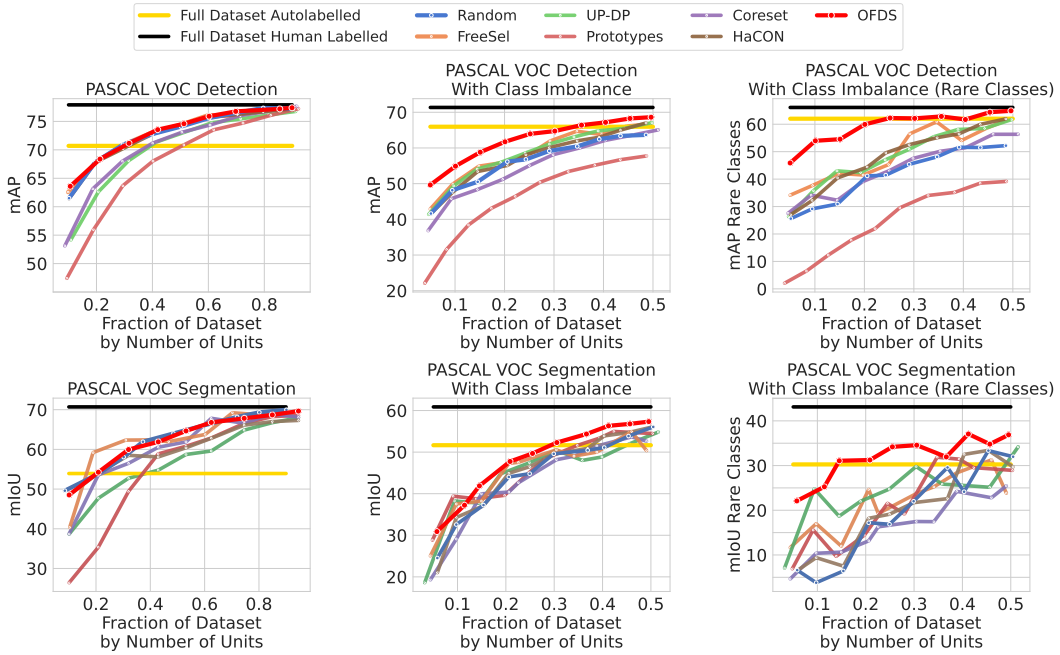

Figure 4: **Data Selection on PASCAL VOC.** The results correspond to a FasterRCNN with ResNet-18 backbone and a Segmenter with ViT-T backbone with the decoder part of the models is trained from scratch. The yellow and black line show the performance of the models being trained on the full dataset with either autolabels or human labels. The remaining points correspond to training on subsets with human annotations. We observe that OFDS consistently performs best amongst the on the methods full dataset and outperforms all baselines on the class imbalanced datasets.

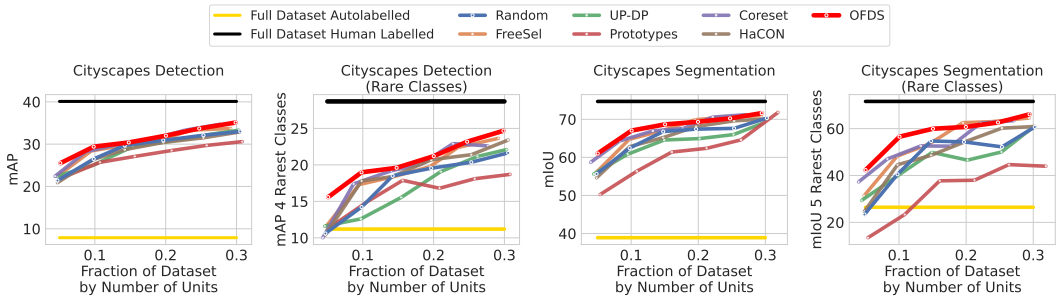

Figure 5: **Data Selection on Cityscapes.** The results are obtained with the same models as for Figure 4. When training on subsets with human annotations, OFDS consistently performs best, in particular when evaluating on rare classes.

achieves the highest overall performances. This improvement is especially notable when evaluating only on the rarest classes, highlighting OFDS's ability to select effective instances from rare classes.

Comparison Across All Settings. Conclusively, we highlight that OFDS outperforms or performs on par with the best baselines across *all* experimental settings and models. OFDS performs best in both balanced and imbalanced class scenarios. This is crucial for practical applications where the presence of class imbalance may not be known in advance. Random selection remains a strong baseline in the class-balanced setting but is substantially surpassed by OFDS in the class imbalanced setting. FreeSel and Coreset outperform random selection on Cityscapes but fail to consistently perform better than random selection on PASCAL VOC. The selection based on prototypes yields the worst performance across all settings. This underlines the finding by Sorscher et al. (2022) that dataset selection with the prototypical approach amplifies class imbalance.

4.4 Combining Autolabels With Data Selection

In Section 4.2, we observe that training on the full dataset with autolabels on PASCAL VOC can be superior to training with human-labeled subset under very constrained annotation budgets.

As the autolabeling process can in any case be carried out at little cost (see Section L), we assess whether the performance under constrained annotation budgets can be improved by incorporating autolabels in addition to human-labeled subsets. Therefore, we first pre-train the models using autolabels on the entire dataset and then fine-tune on human-annotated subsets. In this setup, the purpose of the methods for data selection is to determine the subset used for fine-tuning. We compare random selection, FreeSel and OFDS as the best performing methods from Section 4.3. The results on the PASCAL VOC datasets with class imbalance and Cityscapes are shown in Figures 6 and 23. We observe that fine-tuning with human-labeled images improves the performance over training purely with autolabels, even for the smallest annotation budget. The improvements on PASCAL VOC are larger in comparison to Cityscapes. This is a result of the stronger performance achieved by training with autolabels on PASCAL VOC. For both datasets, selecting the data for fine-tuning through OFDS leads to the best performance. This highlights that data selection can be effectively combined with autolabeling to obtain a holistic training setup for limited annotation budgets.

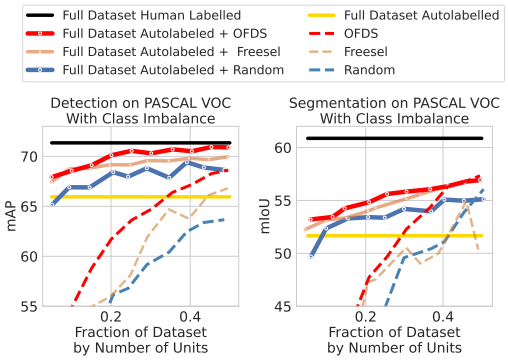

Figure 6: **Combining Autolabels With Data Selection.** The solid lines correspond to fine-tuning from the checkpoint pre-trained with autolabels. For the dashed lines the models were trained as in Section 4.3. We find that selecting the subsets for fine-tuning with OFDS consistently leads to the best performances. Results for the cityscapes datasets are shown in Figure 23.

4.5 Initial Data Selection For Active Learning

As discussed in Section 2, the cold start problem refers to selecting an initial pool of labeled images for active learning which is typically done randomly. In Figure 7, we replace this random initial dataset by datasets selected through OFDS or HaCON, which is the only baseline specifically devised for the cold start problem. Exemplarily, we use the state-of-the-art framework PPAL Yang et al. (2024) for object detection and EquAL Golestaneh & Kitani (2020) for semantic segmentation which performed best in a benchmarking study Mittal et al. (2023) and apply both methods on PASCAL VOC with class imbalance. Further experiments that validate our observation with additional active learning methods and the cityscapes dataset can be found in Section F. We highlight that selecting the initial data through OFDS improves the performance during the entire active learning process and yields better results than a random selection and HaCON.

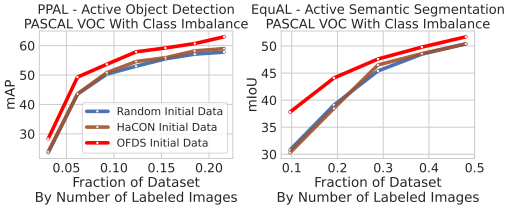

Figure 7: **Initial Data Selection Through OFDS Improves Active Learning.** We train a ResNet-50 RetinaNet Lin et al. (2017)for detection using PPAL and Wide-ResNet38 DeepLabv3+ Chen et al. (2018) for semantic segmentation using EquAL. The only difference between the different lines is the initial data selection. Further experiments and details on the cold start problem of active learning can be found in Section F.

5 Conclusion

In this work, we discuss how foundation models can be leveraged to effectively utilize a fixed annotation budget for training compact models for dense prediction tasks. We find that only for simple datasets and under very constrained annotation budget, training purely with autolabels yields competitive results. For more complex datasets, human annotations remain indispensable. Next, we address the question of which images to select for annotation with OFDS. Our method demonstrates an advantage by consistently improving the performance in comparison to existing baselines for data selection, particularly in class imbalanced settings. This is due to the fact that unlike prior approaches, OFDS guides the data selection at the level of objects rather than images and constructs a semantic covering of all target classes. Finally, we demonstrate that pre-training with autolabels on the full dataset before fine-tuning on human labeled OFDS-selected subset further enhances the final performance of the downstream models.

REFERENCES

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.

Amro Abbas, Evgenia Rusak, Kushal Tirumala, Wieland Brendel, Kamalika Chaudhuri, and Ari S. Morcos. Effective pruning of web-scale datasets based on complexity of concept clusters. In *ICLR*, 2024.

Cathaoir Agnew, Anthony Scanlan, Patrick Denny, Eoin M. Grua, Pepijn van de Ven, and Ciarán Eising. Annotation quality versus quantity for object detection and instance segmentation. *IEEE Access*, 12:140958–140977, 2024. doi: 10.1109/ACCESS.2024.3467008.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

Liangyu Chen, Yutong Bai, Siyu Huang, Yongyi Lu, Bihan Wen, Alan Yuille, and Zongwei Zhou. Making your first choice: To address cold start problem in medical active learning. In *Medical Imaging with Deep Learning*, 2024.

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 *CVPR*, 2016.

Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *JMLR*, 12(42):1501–1536, 2011.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Na Dong, Yongqiang Zhang, Mingli Ding, and Gim Hee Lee. Boosting long-tailed object detection via step-wise learning on smooth-tail data. In *ICCV*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Zeyad Ali Sami Emam, Hong-Min Chu, Ping-Yeh Chiang, Wojciech Czaja, Richard Leapman, Micah Goldblum, and Tom Goldstein. Active learning at the imagenet scale. *arXiv preprint arXiv:2111.12880*, 2021.

Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

Dan Feldman. Introduction to core-sets: an updated survey. *arXiv preprint arXiv:2011.09384*, 2020.

S. Alireza Golestaneh and Kris Kitani. Importance of self-consistency in active learning for semantic segmentation. *BMVC*, 2020.

Chengcheng Guo, Bo Zhao, and Yanbing Bai. Deepcore: A comprehensive library for coreset selection in deep learning. In *Database and Expert Systems Applications*, 2022.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2011.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

Alex Holub, Pietro Perona, and Michael C. Burl. Entropy-based active learning for object recognition. In *CVPR workshops*, 2008.

Sehyun Hwang, Sohyun Lee, Hoyoung Kim, Minhyeon Oh, Jungseul Ok, and Suha Kwak. Active learning for semantic segmentation with multi-class label query. In *NeurIPS*, 2023.

Achin Jain, Kibok Lee, Gurumurthy Swaminathan, Hao Yang, Bernt Schiele, Avinash Ravichandran, and Onkar Dabeer. Completr: Reducing the cost of annotations for object detection in dense scenes with vision transformers. *arXiv preprint arXiv:2209.05654*, 2022.

Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6:1–54, 2019.

Rui Ju and Weiming Cai. Fracture detection in pediatric wrist trauma x-ray images using yolov8 algorithm. *Scientific Reports*, 13, 2023.

Hoyoung Kim, Sehyun Hwang, Suha Kwak, and Jungseul Ok. Active label correction for semantic segmentation with foundation models. In *ICML*, 2024.

Hojun Lee, Suyoung Kim, Junhoo Lee, Jaeyoung Yoo, and Nojun Kwak. Coreset selection for object detection. In *CVPR Workshops*, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023a.

Xin Li, Sima Behpour, Thang Doan, Wenbin He, Liang Gou, and Liu Ren. UP-DP: Unsupervised prompt learning for data pre-selection with vision-language models. In *NeurIPS*, 2023b.

Hubert Lin, Paul Upchurch, and Kavita Bala. Block annotation: Better image annotation with sub-image decomposition. In *ICCV*, 2019. doi: 10.1109/ICCV.2019.00539.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *ICML*, 2020.

Sudhanshu Mittal, Joshua Niemeijer, Jörg P. Schäfer, and Thomas Brox. Best practices in active learning for semantic segmentation. In *GCPR*, 2023.

Eszter Nagy, Michael Janisch, Franko Hrvzic, Erich Sorantin, and Sebastian Tschauner. A pediatric wrist trauma x-ray dataset (grazpedwri-dx) for machine learning. *Scientific Data*, 9, 2022.

Vishwesh Nath, Dong Yang, Holger R. Roth, and Daguang Xu. Warm start active learning with proxy labels and selection via semi-supervised fine-tuning. In *MICCAI*, 2022.

David Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, 9(1):135 – 140, 1981.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Lukas Rauch, Matthias Aßenmacher, Denis Huseljic, Moritz Wirth, Bernd Bischl, and Bernhard Sick. Activeglae: A benchmark for deep active learning with transformers. In *ECML PKDD 2023*, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Comput. Surv.*, 54(9), 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *NeurIPS*, 2015.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Nermin Samet, Oriane Siméoni, Gilles Puy, Georgy Ponimatkin, Renaud Marlet, and Vincent Lepetit. You never get a second chance to make a good first impression: Seeding active learning for 3d semantic segmentation. In *ICCV*, 2023.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 1948.

Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. In *NeurIPS*, 2022.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In *NeurIPS*, 2022.

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. *arXiv preprint arXiv:2105.05633*, 2021.

Jingru Tan, Xin Lu, Gang Zhang, Changqing Yin, and Quanquan Li. Equalization loss v2: A new gradient balance approach for long-tailed object detection. In *CVPR*, June 2021.

Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In *CVPR*, 2021a.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *NeurIPS Track on Datasets and Benchmarks*, 2021b.

Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *ECCV*, 2020.

Xudong Wang, Long Lian, and Stella X Yu. Unsupervised selective labeling for more effective semi-supervised learning. In *ECCV*, 2022.

Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. On the de-duplication of laion-2b. *arXiv preprint arXiv:2303.12733*, 2023.

Jingqian Wu, Rongtao Xu, Zach Wood-Doughty, Changwei Wang, Shibiao Xu, and Edmund Y. Lam. Segment anything model is a good teacher for local feature learning. *arXiv preprint arXiv:2309.16992*, 2024.

Yichen Xie, Mingyu Ding, Masayoshi Tomizuka, and Wei Zhan. Towards free data selection with general-purpose models. In *NeurIPS*, 2023.

Mengmeng Xu, Yancheng Bai, and Bernard Ghanem. Missing labels in object detection. In *CVPR Workshops*, 2019.

Chenhongyi Yang, Lichao Huang, and Elliot J. Crowley. Plug and Play Active Learning for Object Detection. In *CVPR*, 2024.

Yue Yao, Tom Gedeon, and Liang Zheng. Large-scale training data search for object re-identification. In *CVPR*, June 2023.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*, 2023.

Song Zhang, Qingzhong Wang, Junyi Liu, and Haoyi Xiong. Alps: An auto-labeling and pre-training scheme for remote sensing segmentation with segment anything model. *arXiv preprint arXiv:2406.10855*, 2024.

Yixin Zhang, Shen Zhao, Hanxue Gu, and Maciej A. Mazurowski. How to efficiently annotate images for best-performing deep learning based segmentation models: An empirical study with weak and noisy annotations and segment anything model. *arXiv preprint arXiv:2312.10600*, 2023.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

Changyuan Zhou, Yumin Guo, Qinxue Lv, and Ji Yuan. Optimizing object detection via metric-driven training data selection. In *CVPR Workshops*, 2024.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021.

APPENDIX

We start with an overview of the content of the Appendix:

- In Section A, we perform ablations with different models for the downstream tasks.
- In Section B, we evaluate OFDS with object proposals from larger foundation models.
- Results on two additional datasets from specialized domains are shown in Section C.
- To demonstrate the advantage of using object-level instead of image-level features for the data selection, we compare to an additional image-focused baseline in Section D.
- We provide further insights into the effect of clustering the object features for OFDS in Section E.
- In Section F, we provide additional experiments on improving the cold start problem for active learning with OFDS.
- In Section G, we report class balance scores for the selected subsets to validate the effectiveness of OFDS in selecting subsets with improved class balance.
- Since the random baseline and FreeSel are based on a probabilistic selection process, we repeat the data selection on PASCAL VOC with class imbalance and assess the extent of the resulting fluctuations in Section H.
- In Section I, we provide further details on the class distributions including the rare classes used for evaluation and the subsets used to calibrate the object proposer.
- In Section J, we discuss the calibration of the foundation model for generating autolabels and object proposals.
- Further details on the implementation of OFDS and coreset selection are discussed in Section K.
- In Section L, we provide a discussion on the computational cost of OFDS and generating autolabels.
- The complete results for combining autolabels with data selection are shown in Section M.
- In Section O, we provide the hyperparameter configurations used for all training runs.
- Potential limitations of our approach are discussed in Section N.

A MODEL ABLATIONS FOR DOWNSTREAM TASKS

To highlight that the performance advantages of selecting data through OFDS are independent of the model chosen for the downstream tasks, we perform ablations using a Deformable DETR Zhu et al. (2021) for object detection and a PSP Net Zhao et al. (2017) for semantic segmentation. Both models are based on ResNet-18 backbones which were pre-trained on ImageNet. As in Section 4.3, the decoders are trained from scratch to assess only the influence of data selection without any confounding effects from autolabels. The results are shown in Figure 8. We observe that selecting the data through OFDS results in the best performances which confirms our previous findings.

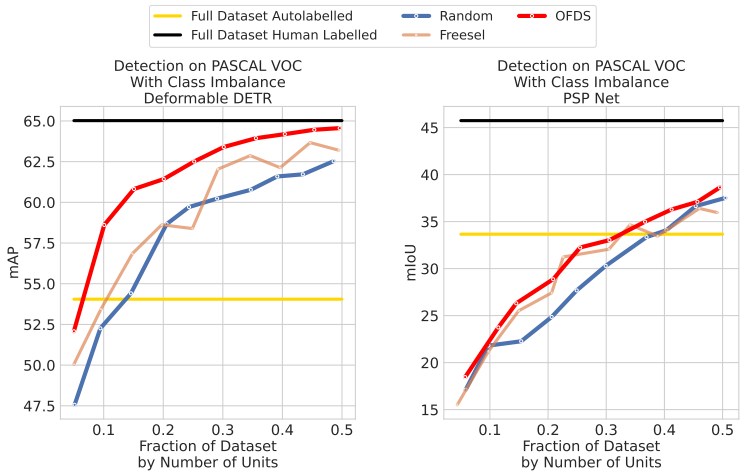

Figure 8: **Ablations with Different Models for the Downstream Tasks.** OFDS also leads to the best results for training a Deformable DETR for object detection and a PSP Net for semantic segmentation.

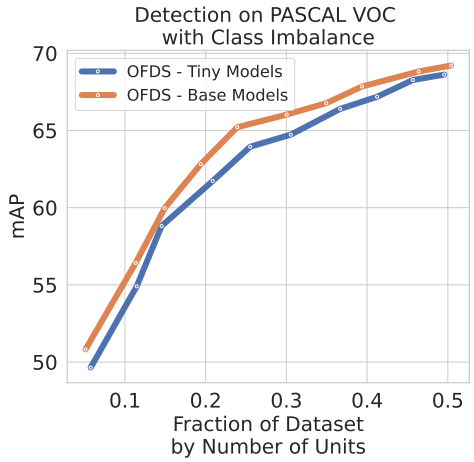

Figure 9: **Stronger foundation models improve OFDS.** We compare object feature extraction for OFDS using Grounding SAM with tiny and base models. The results are obtained through training downstream models on the selected subsets with the same setup as for Figure 4. Using the base instead of the tiny models yields small but consistent improvements.

B Ablations on the Choice of Foundation Models

In Section 4, we utilize the Grounding DINO-T and SAM2-T backbones to generate object-level representations. This choice is motivated by the fact that these models achieve the fastest inference time among the Grounding SAM variants. In this section we ablate on this choice by using OFDS with features from the Grounding DINO and SAM2 base models. We highlight that the base models not only feature an architecture with more parameters than the tiny models but were also trained on larger datasets. The results for object detection on the PASCAL VOC dataset with class imbalance are shown in Figure 9. We observe that using features from the base yields small but consistent improvements. As these performance gains remain relatively modest and we use the tiny models for our main experiments as these provide a more favorable trade-off between performance and inference cost of the foundation models.

C Additional Datasets

In this section, we evaluate OFDS in more specialized domains. We consider the LoveDA Wang et al. (2021b) dataset for semantic segmentation of satellite images and the GRAZPEDRWRI-DX Nagy et al. (2022) dataset for fracture detection on radiology images. Both of these dataset features imbalanced class distributions. Due to the lack of calibration datasets for these domain, we use the best confidence threshold for Grounding DINO determined by Kim et al. (2024). For some classes in GRAZPEDRWRI-DX we do not obtain any object proposals. With the annotation budget allocated to these classes in Algorithm 1, we perform a random selection and perform the clustering-based approach with the remaining budget on the classes with object proposals. For training the downstream model use the same experimental setup as for the cityscapes dataset in Section 4. The results are shown in Tables 1 and 1. We observe that on LoveDA, OFDS yields a consistent improvement over all baselines. However, the margin is smaller than on PASCAL VOC with class imbalance or cityscapes. On GRAZPEDRWRI-DX, OFDS performs only as good as a random selection but not worse like the coreset approach for example. This indicates that the more specialized the target domain of the dataset is, the smaller the improvement of OFDS over random selection becomes. However, we highlight that OFDS never performs worse than random selection unlike coreset for example.

D Image-Focused vs. Object-Focused Features

In order to motivate the use of object-level features in OFDS, we compare to an additional baseline which uses image-level features. Therefore, we perform CLIP retrieval on the unlabeled training datasets with a CLIP ViT-B/32 model Radford et al. (2021). We evenly distribute the annotation budget between all classes and retrieve the images with the highest text-to-image similarity using the prompts "a photo of a {classname}". The major drawback of such an approach is that image-level features can be dominated by large or frequent objects and can be confounded by objects outside of

Table 1: **Object detection on GRAZPEDWRI-DX** Reported results are the mAP on the test set of FasterRCNN models with ResNet-18 backbone trained on the selected subsets. The train and test splits is taken from Ju & Cai (2023). The subset size is determined by number of annotation units. Random selection was performed three times and averaged. OFDS performs on par with random selection, while Coreset and FreeSel exhibit lower performance.

Subset Size	5%	10%	15%	20%
OFDS	36.39	**40.87**	**43.00**	43.48
Random	**36.60**	40.03	42.29	**43.66**
FreeSel	34.55	37.96	40.88	42.55
Coreset	32.04	36.65	39.07	41.07

Table 2: **Semantic Segmentation on LoveDA** Reported results are the mIoU on the test set of Segmenter models with ViT-T backbone trained on the selected subsets. The subset size is determined by number of annotation units. As for Table 1, random selection was performed three times and averaged. OFDS consistently outperforms all baselines.

Subset Size	5%	10%	15%	20%	25%	30%
OFDS	**44.10**	**45.39**	**46.12**	**46.51**	**47.16**	**47.35**
Random	43.20	44.14	44.61	45.16	46.06	46.37
FreeSel	41.13	44.51	45.48	45.77	46.24	47.01
Coreset	40.60	43.87	44.63	45.63	45.94	46.29

the target classes. In Figure 11, we highlight the three images with the highest similarity to the class "train" in the cityscapes dataset. None of the images actually contains a train (even though there are several images with clearly visible trains in the dataset) but only streets with streetcar tracks. In Figure 10, we compare the model performances when training on subsets selected through CLIP retrieval in comparison to OFDS on the cityscapes dataset. We observe a clear improvement of OFDS over the image-focused baseline for both object detection and semantic segmentation.

E Influence of Clustering Object Features

In this section, we discuss the impact of clustering object features in OFDS in greater detail. As outlined in Section 3.2, the purpose of clustering the features and selecting individual objects close to the cluster centers is to obtain a density-based covering of the semantic feature space of the individual classes and ensure intra-class diversity. In Section E.1, we compare the clustering-based object selection in OFDS to an ablated variant that uses random selection per class without clustering. In Section E.2, we illustrate further examples of semantic object groups identified through the clustering.

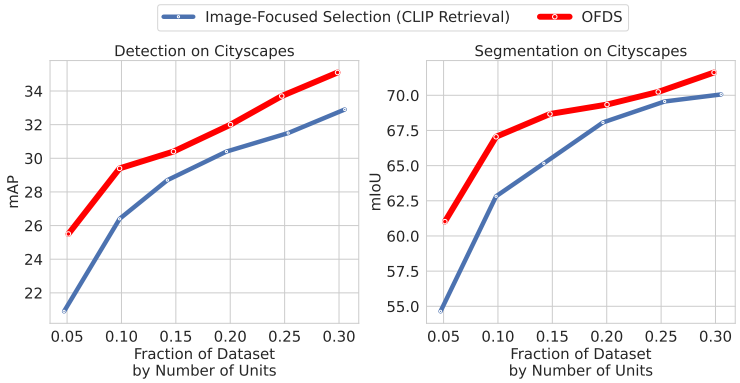

Figure 10: **OFDS Outperforms Image-Focused Baseline.** The image-focused baseline is based on CLIP retrieval with an evenly split budget between all classes. The model and training hyperparameters are the same as in Figure 5.

E.1 Influence on the Performance

To assess the influence of the clustering of object features on the performance, we construct an ablation of OFDS where the objects from every class are chosen randomly from the object proposals instead of the clustering-based approach. More precisely, in Step 7 of Algorithm 1 we annotate images with N_{C_l} randomly chosen objects from the current class. When using already annotated datasets like PASCAL VOC, the difference between these two variants of OFDS is diminished by the fact that the images and class objects in the datasets were selected and annotated by humans such that no or few very similar objects are contained. However, in datasets of e.g. webcrawled images such as LAION, up to 30% of images have been found to be exact or near duplicates Abbas et al. (2023); Webster et al. (2023). As no dense human annotations are available for such datasets, we construct a comparable setting by adding 20% random duplicates to the PASCAL VOC datasets with class imbalance. The results for training downstream models for object detection and semantic segmentation are shown in Figure 12 both with and without duplicates. We observe that the clustering steps leads to a small but consistent improvement in performance. Furthermore, the difference increases when training on the datasets with duplicates. Hence, we conclude that using OFDS with the object selection based on clustering of object features is favorable.

E.2 Additional Cluster Illustrations

In Section 3.2, we motivate the use of unsupervised clustering to construct a density-based covering of the class semantics. In particular, the clustering is used to group semantically similar objects. Figure 13 further illustrates this aspect with object clusters for two additional classes.

F Further Details on the Cold Start Problem for Active Learning

F.1 Experimental Setup

For the experiments on active object detection using PPAL, we use a RetinaNet Lin et al. (2017) with ResNet50 backbone and the same training hyperparameters as in the original publications. We start with a subset consisting of 2.5% of the overall images and add 400 images (roughly 5%) in every active learning round. For semantic segmentation we use the training setup from the benchmarking study by Mittal et al. (2023). The segmentation model is based on the DeepLabv3+ Lin et al. (2017) architecture with a Wide-ResNet38 Zagoruyko & Komodakis (2016) backbone. We start with a subset consisting of 10% of the overall images and add 10% in every active learning round. Note that Mittal et al. (2023) use weaker augmentations than we used for our main experiments in Section 4. For the active learning experiments, the budget is counted by the number of labeled images due to the conception of the used frameworks.

F.2 Combining Autolabels with OFDS for the Cold Start Problem

In this section, we incorporate autolabels into the cold start problem for active learning. For methods like PPAL Yang et al. (2024) or Sener & Savarese (2018), the cold start for active learning cannot be started just with autolabels. This is due to the fact that these active learning methods base the selection of new datapoints to label on an already existing set of labeled images and still require an initial pool of labeled images. Thus, we start active learning methods using a model checkpoint trained on the entire dataset with autolabels together with an initial labeled subset. In Figure 14 we perform active learning with PPAL for object detection on PASCAL VOC with class imbalance and initial datasets selected through OFDS, HaCON and random drawing. We observe that pre-training

Figure 11: **Images with Highest CLIP Text-to-Image Similarity for Class "Train" in Cityscapes.** We use a ViT-B/32 CLIP model to perform retrieval on cityscapes. The images contain streets with streetcar tracks but no actual trains. This highlights the downside of an image-focused approach where the image features can be dominated by objects that are not actually from the target class.

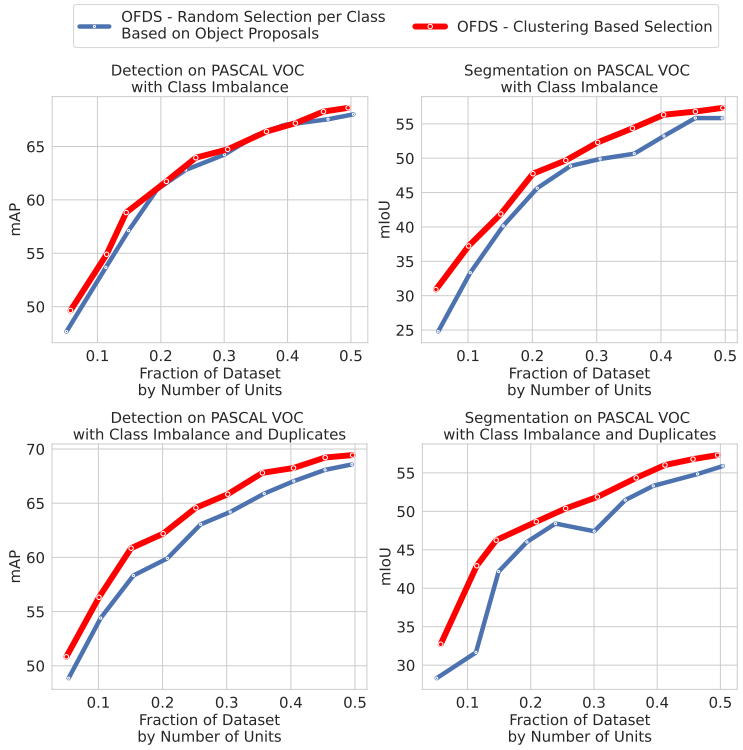

Figure 12: **Influence of Clustering Object Features in OFDS.** We perform an ablation of OFDS where in Step 7 of Algorithm 1 the objects are chosen randomly per class from the object proposals instead of using the clustering-based selection. We compare these two variants of OFDS on the PASCAL VOC dataset with class imbalance. As an additional setting, we add 20% random duplicate images to the dataset. This is motivated by the fact that for example unlabeled web crawled datasets are known to feature a substantial amount of duplicates Webster et al. (2023). In such case, it is particularly important to ensure that the objects in the selected subset are semantically diverse. We observe that the clustering-based object selection yields small but consistent performance improvements over the random selection of object proposals per class, in particular when duplicates are present. The results are obtained using the same model and training setup as in Section 4.3.

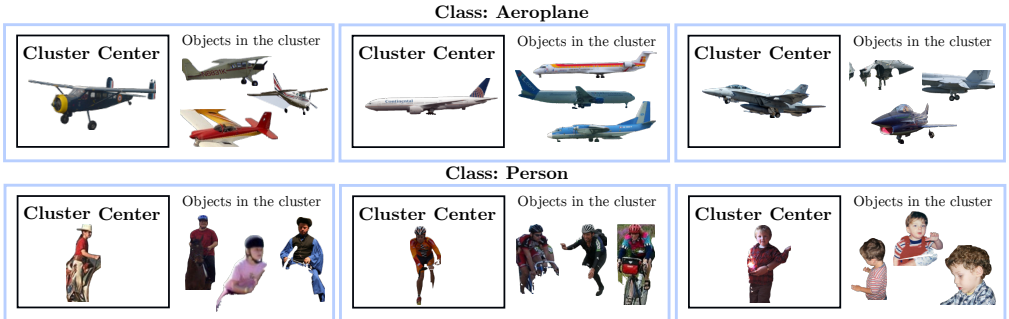

Figure 13: **Illustration of Clusters for Two Additional Classes from PASCAL VOC.** As in Figure 2 the clustering was performed with $N_{Cl} = 30$ clusters per class.

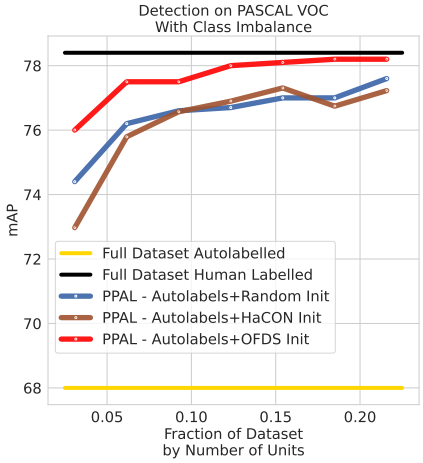

Figure 14: **Leveraging Autolabels Together with OFDS further improves the Cold Start Problem.** The models are trained with the same setup as for Figure 7 but the model is first pre-trained with autolabels before fine-tuning it on the initial labeled data and subsequently performing active learning. Importantly, it is not possible to start active learning with PPAL only with autolabels as the selection of additional datapoints requires a labeled pool to compare to. The only difference between the three lines is the initial dataset. We observe that selecting the initial labeled dataset with OFDS yields the best results.

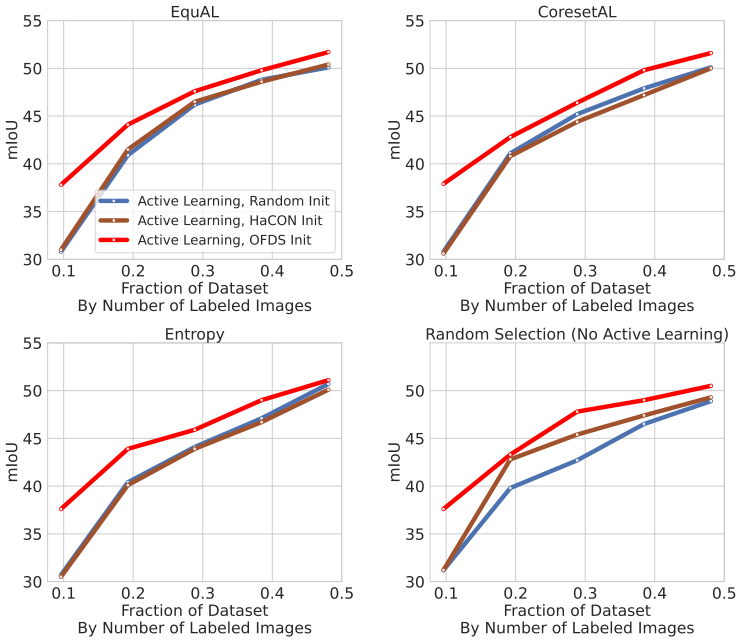

Figure 15: **Different Initial Labeled Datasets for Active Semantic Segmentation on PASCAL VOC with Class Imbalance.** We train a DeepLavV3+ model with Wide-ResNet38 backbone on PASCAL VOC with class imbalance using the setup from Mittal et al. (2023). For every plot, the models were trained with the same active learning frameworks only with different initial datasets. Random selection refers to randomly selecting the additional images in every round. For all frameworks apart from random selection, OFDS improves the performance of the model during the active learning process.

with autolabels substantially improves the performance over training the model from scratch in the initial round. Furthermore, selecting the initial dataset through OFDS improves the performance of the entire training in comparison to random selection.

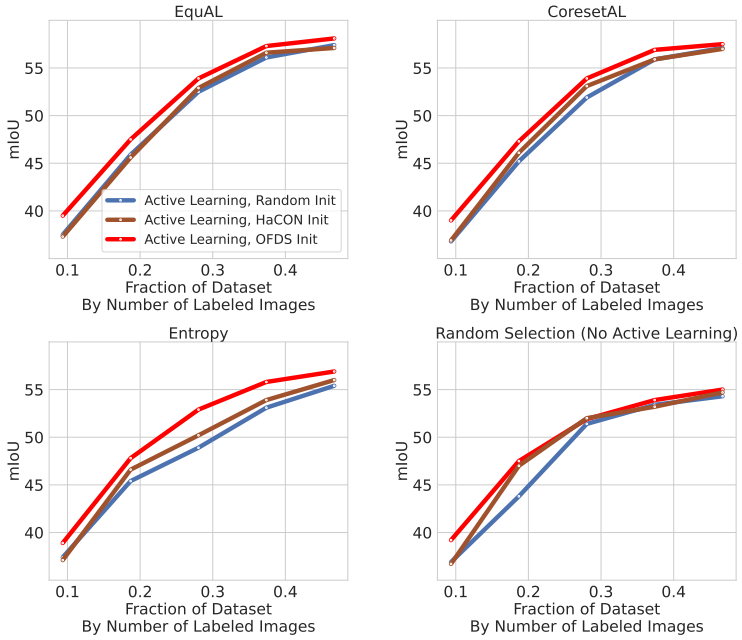

Figure 16: **Different Initial Labeled Datasets for Active Semantic Segmentation on Cityscapes.** We use the same setup as in Figure 15 to train models for semantic segmentation on cityscapes. We again observe that for all frameworks apart from random selection, selecting the initial dataset through OFDS improves active learning.

F.3 Further Active Learning Methods

In this section, we provide further results for cold start problem of different active learning methods. Due to the vast amount of methods for active learning, we restrict ourselves to the tasks of semantic segmentation and assess the cold start problem on two datasets using four baselines from a benchmarking study Mittal et al. (2023). This includes the EquAL framework Golestaneh & Kitani (2020), which achieved the most consistent results in the benchmarking study Mittal et al. (2023), as well as entropy-based active learning Holub et al. (2008); Shannon (1948), coreset Sener & Savarese (2018) and the random baseline. Entropy and coreset are two of the most commonly used frameworks for active learning and remain strong baselines even in more recent works Yang et al. (2024). Coreset for active learning differs from coreset for data selection from Section 3.2. Coreset for active learning, which we refer to as CoresetAL, utilizes the features of the downstream model being trained while coreset for data selection uses features from a pre-trained self-supervised model, in our case DINO. In Figures 15 and 16, we report the results for semantic segmentation on PASCAL VOC with class imbalance. We observe that for all frameworks apart from random selection, OFDS improves the performance of the model during the entire active learning process. This validates our observation from Section 4.

F.4 Direct Comparison of OFDS and Active Learning

In Figure 17, we compare training a model with PPAL active learning for object detection and EquAL for semantic segmentation to a model trained from scratch on subsets selected with OFDS. We observe that on PASCAL VOC with class imbalance, training a model from scratch on data selected through OFDS yields results which are on par with the tested active learning methods. Importantly, we do not claim that training models on data selected through OFDS is generally on par with active learning as there is a vast amount of literature on active learning methods, many of which are highly optimized for specific tasks Hwang et al. (2023), model architectures Rauch et al. (2023) or even datasets Emam et al. (2021) to reach the best performance. Adequately comparing to this line of methods requires extensive experiments which go beyond the scope of this work. Instead, we highlight once more that data selection through OFDS is qualitatively different from active learning as it does not require an initial dataset and is both model and task agnostic.

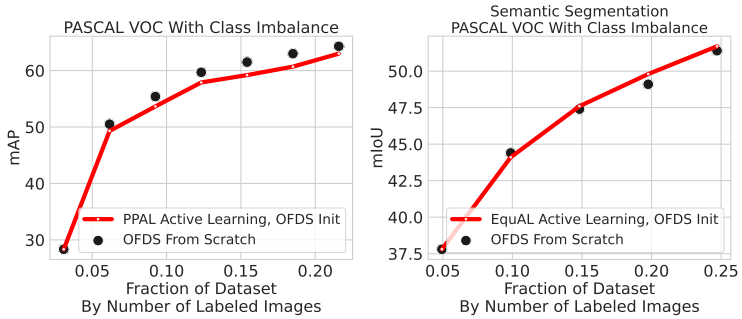

Figure 17: **Direct Comparison of OFDS and Active Learning.** We compare the models trained with the active learning frameworks from Figure 7, where the dataset sizes are iteratively enlarged, to the same models trained from scratch on datasets selected through OFDS. We observe that training on the subsets selected through OFDS without any active learning achieves results that are on par to the active learning methods.

G Class Balance Scores

To quantify how balanced the class distributions of the selected subsets are, we compute the class balance scores introduced by Sorscher et al. (2022). The class balance score $b \in [0, 1]$ is defined as the average balance between any two pairs of classes. It is determined by taking the expectation of drawing two random classes and computing the fraction between the number of objects in the smaller class in comparison to the larger class. A balance score of 1 corresponds to evenly balanced classes and higher scores are generally better. In Figure 18, we show the balance scores for subsets selected from PASCAL VOC with class imbalance and Cityscapes using the six baselines and OFDS. The subsets selected by OFDS consistently feature higher class balance scores compared to all baselines as well as the full datasets. This demonstrates the effectiveness of OFDS in selecting subsets with improved class balance. Since some of the classes in the segmentation split of PASCAL VOC with class imbalance contain only very few objects, these classes are not represented in the small subsets selected through random drawing. As a result, the balance scores for random selection on smaller subsets are lower than the score of the full dataset.

H Repeated Dataset Selection with FreeSel and Random

Since both the random baseline and FreeSel rely on a probabilistic selection process, we repeat the data selection multiple times to assess the extent of the resulting fluctuations. Due to the high computational cost of repeating all experiments, we focus on object detection and semantic segmentation on PASCAL VOC with class imbalance and perform the selection three times for every subset size. The results are depicted in Figure 19. The performance improvement achieved by OFDS over the baselines clearly surpasses the fluctuations caused by the randomness in FreeSel or random selection. Importantly, OFDS features a deterministic selection process, which is advantageous for practical applications where the data selection can only be performed once.

I Class Distributions

In this section, we provide additional details on the class distributions as well as the subsets selected for the evaluation on rare classes.

I.1 Class Distributions for Object Detection

Figure 20 displays the class distributions for the object detection datasets. It complements Figure 3 which shows the class distributions for the segmentation splits.

I.2 Rare Classes

For the evaluation on rare classes of PASCAL VOC with class imbalance, we consider the six classes that were pruned from the full dataset. On the Cityscapes dataset, we focus on the smallest four classes for object detection, representing half of the total classes. For semantic segmentation, we evaluate the performance on the five smallest classes, accounting for one fourth of the total classes.

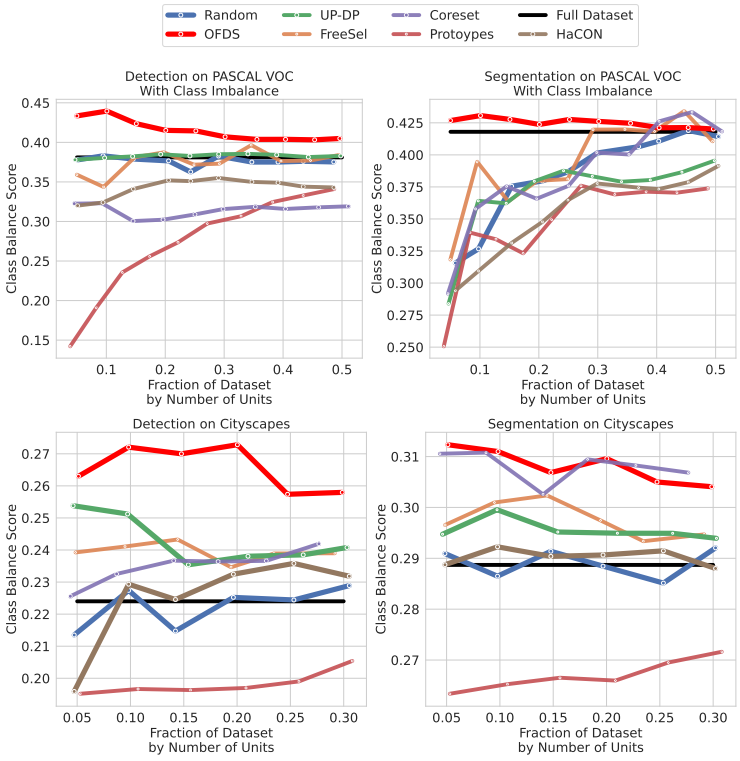

Figure 18: **OFDS Selects More Class-Balanced Subsets.** We compute the class balance score introduced by Sorscher et al. (2022) and find that the subsets selected by OFDS consistently feature higher class balance scores than the baselines. This indicates more evenly distributed class distributions.

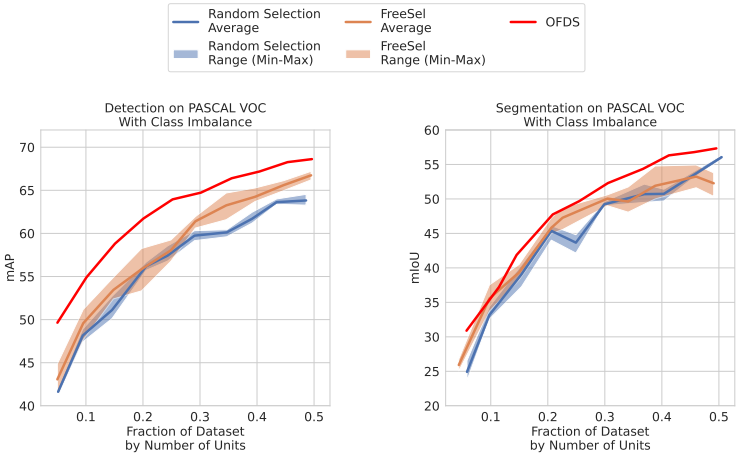

Figure 19: **Repeated Data Selection For Baselines With Randomness.** Since the random baseline and FreeSel are based on probabilistic selection, we repeat the data selection process for object detection and semantic segmentation on PASCAL VOC with class imbalance three times. The solid lines represent the mean performance and the shaded areas indicate the range between the minimum and maximum performance for each subset size. The improvement achieved by OFDS over the baselines consistently exceeds the fluctuations resulting from the randomness in their selection processes. OFDS is included for reference but features a fully deterministic selection.

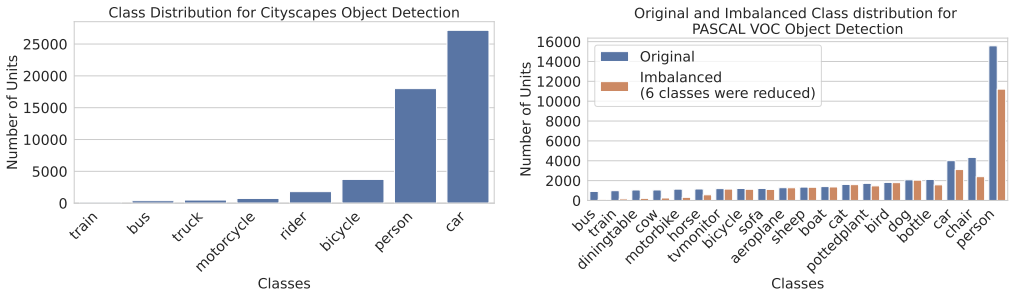

Figure 20: **Class distributions for the Object Detection Datasets.**

In both cases on Cityscapes, the rare classes jointly contain less than 3.5% of the overall number of objects.

- **Rare Classes for PASCAL VOC Object Detection:** bus, train, diningtable, cow, motorbike, horse
- **Rare Classes for PASCAL VOC Semantic Segmentation:** diningtable, bicycle, train, horse, tvmonitor, motorbike
- **Rare Classes for Cityscapes Object Detection:** train, bus, truck, motorcycle
- **Rare Classes for Cityscapes Semantic Segmentation:** train, bus, truck, motorcycle, wall

J Calibrating the Foundation Model for Generating Autolabels and Object Proposals

J.1 Calibration Data

As discussed in Section 4, we utilize a subset of the MSCOCO Lin et al. (2014) validation split to calibrate the object proposer. Therefore, we select only images containing objects from classes related to the target classes. Since MSCOCO does not feature the same classes as the target datasets, we manually identify and select these related classes:

- **PASCAL VOC:** airplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, diningtable, dog, horse, motorcycle, person, potted plant, sheep, couch, train, tvmonitor
- **Cityscapes:** car, bus, truck, motorcycle, bicycle, traffic light

Importantly, these classes are only used to calibrate the confidence threshold. For generating the object proposals and the autolabels, we use the actual target classes.

J.2 Threshold Calibration of the Object Proposer

In Figure 21, we visualize the threshold calibration of the foundation model for autolabeling and for generating the object proposer. The threshold bounds the confidence with respect to the bounding boxes and the final class predictions are obtained by taking an argmax over the text scores. When generating autolabels, we use the threshold which achieves the highest F1 score on the calibration data. For generating object proposals, we use a threshold that yields 5% false positives on the calibration data to control the precision of the object proposals.

K Implementation Details

In this section, we provide further details on the implementation of OFDS and coreset.

OFDS In step 5 of Algorithm 1, the number of objects to select is set as $N_{C_l} = \frac{B-N(\mathcal{S})}{(M-l+1)N_O}$. Here, N_{C_l} is determined from the leftover budget of annotation units $B - N(\mathcal{S})$ which is updated after every class. This budget is equally distributed between the $M - l + 1$ remaining classes at every iteration of step 5. Therefore, we divide the leftover budget by the number of remaining classes to obtain the annotation budget per class. To obtain the number of objects to annotate from this annotation budget, we further divide by N_O which is expected number of annotations per selected object.
Since we select objects only from clusters that do not contain any annotated objects from previous steps, the number of selected objects does not necessarily correspond to the number of clusters. Instead, we initialize the number of clusters by N_{C_l} and gradually increase it until we find N_{C_l}

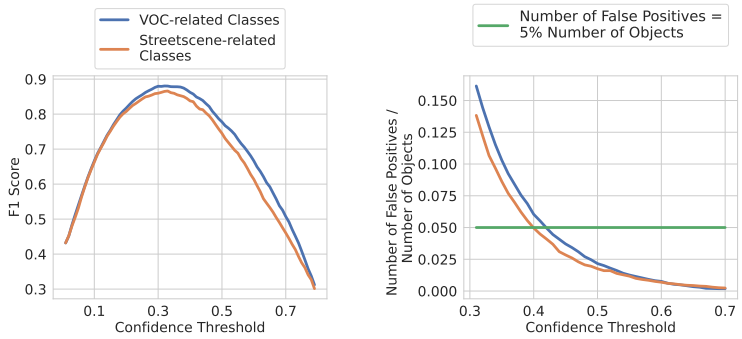

Figure 21: **Calibration of the Object Proposer** using the FPR to control the precision for OFDS and the F1 score for autolabeling. The calibration data is as described in Section J.1

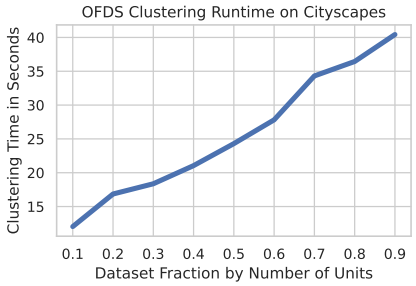

Figure 22: **Compute Time for the Data Selection in OFDS Without the Time to Generate the Object Proposals** The clustering and selection was performed on a Xeon Gold 6150 CPU.

clusters without any annotated objects. In practice, we achieve this by iteratively multiplying the number of clusters with 1.05 until enough clusters are present.
Furthermore, in step 1 of Algorithm 1 we only consider object proposals with bounding boxes smaller than 0.05% of the overall image area to filter out noisy proposals.

Coreset When using the k-centers algorithm to select subsets that are relatively large compared to the full dataset, the complexity becomes prohibitive due to the quadratic cost of computing the distances between all points in the selected subset and the non-selected subset. To overcome this problem, we use a batched version that considers batches of size 512 when selecting new points for the subset. Thereby, the complexity becomes independent of the size of the unlabeled image pool.

L Computational Cost of Autolabeling and OFDS

We use a modified pipeline of Grounding SAM 2 to compute the object proposals using a NVIDIA v100 GPU. We selected the DINO-T and SAM2-T models to achieve the highest throughput. In this setup, the throughput is 3 images per second on average. The clustering and selection process of OFDS takes only in the order of seconds on a Xeon Gold 6150 CPU. The times for the clustering steps on the cityscapes dataset are shown in Figure 22. We emphasize that generating autolabels or object proposals and the selection process are thereby substantially less computationally expensive than the model trainings carried out in Section 4. Each training took between 6 hours for object detection on PASCAL VOC to 21 hours for semantic segmentation on PASCAL VOC using the same compute hardware as for the data selection.

M Combining Dataset Selection with Autolabels

M.1 Combining Autolabels with OFDS on Cityscapes

To complement Figure 6, we report the results of the models pre-trained with autolabels and then fine-tuned on selected subsets with human labels on the cityscapes dataset in Figure 23. As for PASCAL VOC, we observe that fine-tuning with human-labeled images improves the performance over training purely with autolabels. The improvements achieved through pre-training on autolabels

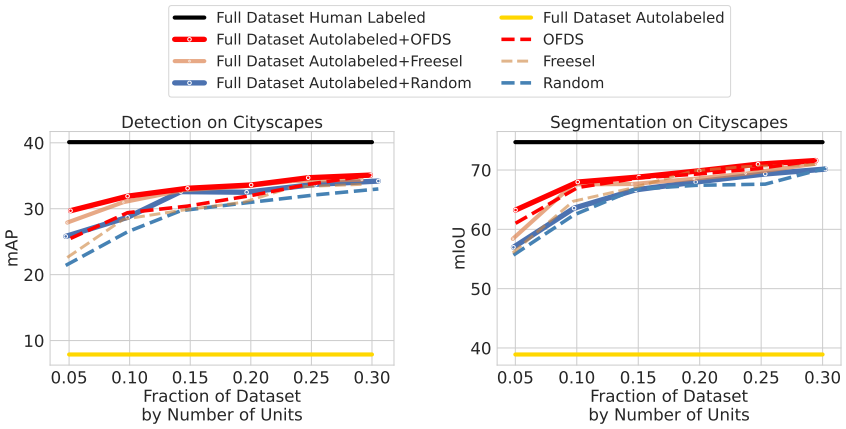

Figure 23: **Combining Autolabels With Data Selection on Cityscapes.** The training setup is the same as for Figure 6 but with hyperparameters adjusted for the cityscapes dataset as stated in Section O. The results validate our observation that fine-tuning with human-labeled improves the performance over training purely with autolabels and selecting the subset for labeling with OFDS yields the best performances.

are smaller in comparison on PASCAL VOC as a results of the weaker performance of the models trained with autolabels.

M.2 Complete Results on PASCAL VOC with Class Imbalance and Cityscapes

The complete results including all six baselines for fine-tuning on selected datasets with human labels are shown in Figures 24 and 25. As in Section 4.4, the pre-trained checkpoints are obtained through training with autolabels on the full datasets. The results confirm that selecting the data for fine-tuning through OFDS consistently leads to the best performances and yields improvements over all baselines.

N Limitations

OFDS depends on features generated by the object proposer and thereby inherits its biases and limitations, which may reduce its effectiveness in specialized target domains. Furthermore, achieving a balanced class distribution in some cases can be challenging even with OFDS due to class co-occurrences on image level.

O Training Hyperparameters

The hyperparameter configurations for all models trained in this work can be found in Table 3 for object detection and in Table 4 for semantic segmentation respectively. Due to the smaller resolution and subset size for the PASCAL VOC segmentation split, we trained with higher weight decay and learning rate in comparison to Cityscapes. Loss functions including loss weights are taken as in the original works that presented the model architectures.

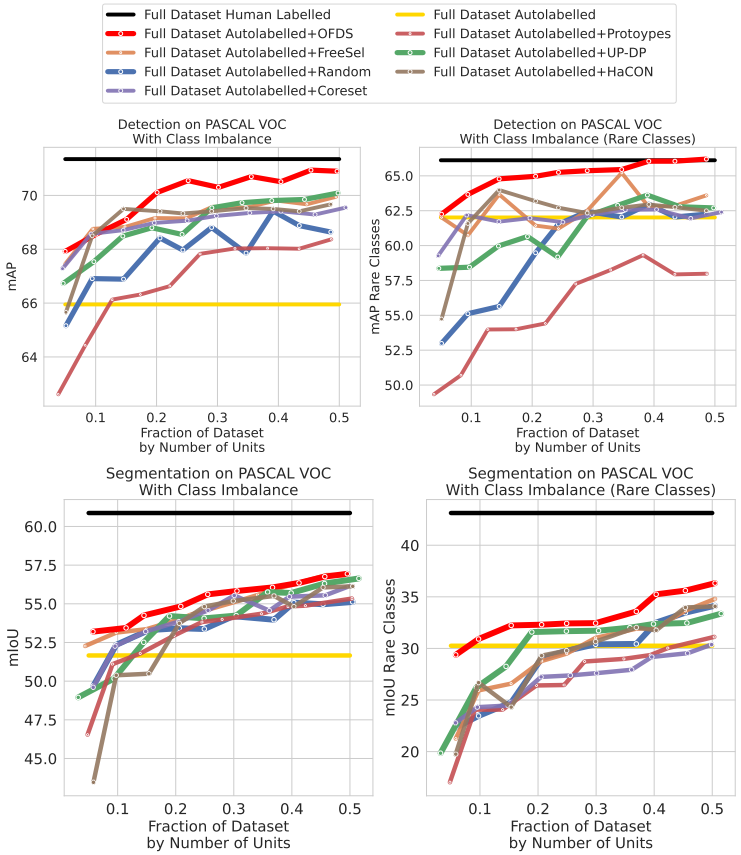

Figure 24: **Complete Results for Combining Autolabels with Data Selection on PASCAL VOC with Class Imbalance.** The results correspond to a FasterRCNN with ResNet-18 backbone and a Segmenter with ViT-T backbone. The models were first pre-trained on the full dataset with autolabels and then fine-tuned on selected subsets with human labels. These subsets were selected by the six baselines or OFDS given the fixed annotation budgets indicated on the x-axis.

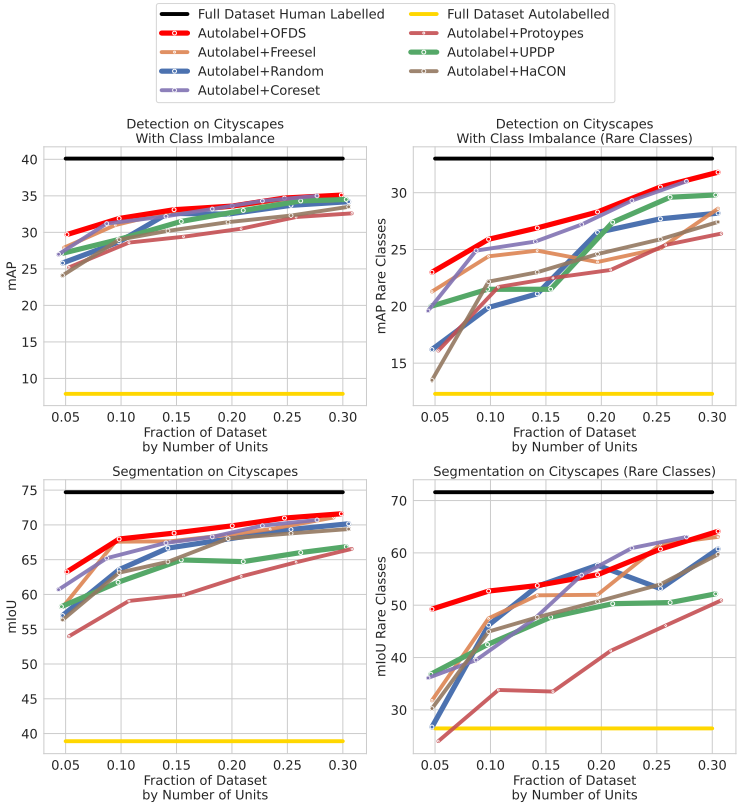

Figure 25: **Complete Results for Combining Autolabels with Data Selection on Cityscapes.** The results were obtained using the same setup as for Figures 6 and 24.

Hyperparameters	Faster-RCNN	Deformable DETR
Backbone	ResNet-18	ResNet-18
Optimizer	AdamW Loshchilov & Hutter (2019)	AdamW
Optimizer Parameters	ϵ =10e-8, β ∈(0.9, 0.999)	ϵ =10e-8, β ∈(0.9, 0.999)
Base lr	1e-4	1e-4(VOC) backbone scaled by factor 0.1
Weight decay	5e-4, 3e-1 (VOC Fine-Tuning)	1e-4 (VOC)
Optimizer Steps	80k, 60k (Fine-Tuning)	80k
Batchsize	8	8
Lr schedule	Cosine Annealing	Cosine Annealing
Warmup steps	1k	1k
Warmup configuration	Linear Warmup, Factor 0.1	Linear Warmup, Factor 0.1
Augmentations	PhotoMetric Distortion, Random Crop, Random Flip	PhotoMetric Distortion, Random Crop, Random Flip

Table 3: **Hyperparameters for Object Detection.** We provide the hyperparameter configurations for all object detection models trained in this work. The specified values correspond to the setup described in Section 4.3. Unless explicitly stated otherwise, the same configurations are used for fine-tuning in Section 4.4.

Hyperparameters	Segmenter	PSPNet
Backbone	ViT-T	ResNet-18
Optimizer	AdamW	SGD
Optimizer Parameters	ϵ =10e-8, β ∈(0.9, 0.999)	momentum 0.9
Base lr	1e-5 (VOC), 1e-4 (Cityscapes) 1e-6 (VOC Fine-Tuning)	1e-2 (VOC)
Weight decay	1e-2 (VOC), 5e-4 (Cityscapes) 1e-1 (VOC Fine-Tuning)	5e-4 (VOC)
Optimizer Steps	80k, 60k (Fine-Tuning)	80k
Batchsize	4	4
Lr schedule	Cosine Annealing	Cosine Annealing
Warmup steps	1k	0
Warmup configuration	Linear Warmup, Factor 0.1	
Augmentations	PhotoMetric Distortion, Random Crop, Random Flip	PhotoMetric Distortion, Random Crop, Random Flip

Table 4: **Hyperparameters for Semantic Segmentation.** For all models trained for semantic segmentation in this work, we list the hyperparameters configurations. The stated values refer to the setup for Section 4.3. When not explicitly stated otherwise the configurations used for fine-tuning in Section 4.4 are the same.