

# Active Learning Meets Foundation Models: Fast Remote Sensing Data Annotation for Object Detection

Marvin Burges\*

TU Wien

Vienna, Vienna, Austria

mburges@cvl.tuwien.ac.at

Philipe Ambrozio Dias & Carson Woody & Sarah Walters & Dalton Lunga

Oak Ridge National Laboratory

Oak Ridge, Tennessee, USA

{ambroziodiap, woodycb, waltersse, lungadd}@ornl.gov

\*This work was conducted during a fellowship at Oak Ridge National Laboratory.

## Abstract

*Object detection in remote sensing demands extensive, high-quality annotations—a process that is both labor-intensive and time-consuming. In this work, we introduce a real-time active learning and semi-automated labeling framework that leverages foundation models to streamline dataset annotation for object detection in remote sensing imagery. For example, by integrating a Segment Anything Model (SAM), our approach generates mask-based bounding boxes that serve as the basis for dual sampling: (a) uncertainty estimation to pinpoint challenging samples, and (b) diversity assessment to ensure broad data coverage. Furthermore, our Dynamic Box Switching Module (DBS) addresses the well-known cold start problem for object detection models by replacing its suboptimal initial predictions with SAM-derived masks, thereby enhancing early-stage localization accuracy. Extensive evaluations on multiple remote sensing datasets, along with a real-world user study, demonstrate that our framework not only reduces annotation effort but also significantly boosts detection performance compared to traditional active learning sampling methods. The code for training and the user interface is available under [https://github.com/mburges-cvl/ICCV\\_AL4FM](https://github.com/mburges-cvl/ICCV_AL4FM).*

## 1. Introduction

Object Detection (OD) is a fundamental computer vision task that relies on vast amounts of annotated training data [22, 42, 45]. However, the manual annotation

process, which must accurately capture both object classes and their corresponding coordinates, is labor-intensive and time-consuming. Active Learning (AL) has emerged as a promising strategy to alleviate this burden by selectively choosing the most informative samples for labeling. This approach is predicated on the idea that by intelligently selecting data points—rather than annotating a randomly sampled dataset—it is possible to train models that achieve higher accuracy and yet this process requires substantially fewer labeled instances.

Various AL strategies have been proposed. Uncertainty-based methods [13, 16, 48, 49] prioritize samples for which the model exhibits low confidence, thereby targeting cases that are most likely to improve model performance upon annotation. In contrast, diversity-based methods [1, 5, 33, 37] aim to cover a broad spectrum of the data distribution by selecting samples that are distinct from the already labeled set. More recently, combined approaches that integrate both uncertainty and diversity metrics have been developed, aiming to combine their complementary strengths [45].

Despite these advances, applying AL to OD presents unique challenges. Unlike classification tasks, OD involves the dual objectives of accurate localization and classification. This duality necessitates the development of AL methods that can effectively assess and balance uncertainties related to both the positioning of objects and their categories. Moreover, an additional challenge for AL is the cold start problem, where the initial lack of annotated data can lead to suboptimal model performance, thus complicating the early stages of sample selection [40].

In this paper, we present a novel interactive AL framework for OD that leverages recent advances in foundation

models (FMs). Our approach utilizes the bounding boxes derived from the predicted masks of the Segment Anything Model (SAM) [17, 19] for uncertainty estimation, while the features extracted from individual masks are employed to assess sample diversity. Notably, our method addresses the cold start problem by initially using the bounding boxes provided by SAM instead of those predicted by the object detector. This is particularly beneficial because, although the classification performance of a detector—especially one built on a FM backbone—improves rapidly, its localization performance is initially suboptimal. The main contributions of this work are: (i) the integration of FMs into an interactive AL framework for OD, (ii) a dual sampling strategy that combines uncertainty estimation from SAM-derived boxes with diversity estimation from mask features, (iii) an effective solution to the cold start challenge that enhances early-stage localization performance, (iv) a real-world user study, that highlights the advantages of our methodology, and (v) an interactive user interface that can be adapted for annotation of novel and rare objects in imagery.

## 2. Related Work

AL for OD follows a three-step iterative process: (1) deploying a detector trained on labeled images to make inferences on unlabeled data, (2) strategically selecting and annotating images within a predefined budget, and (3) retraining the model and evaluating performance. This approach optimizes annotation resources by prioritizing informative images. In the paragraphs below, we review the different strategies introduced in the related literature for sample selection, as well as recent works exploring SAM for annotation assistance.

**Uncertainty-Based Active Learning:** Uncertainty-based methods leverage model prediction confidence to select informative samples. Yoo and Kweon [47] augment networks with loss prediction modules to identify potential misclassifications. MI-AOD [49] models the gap between instance- and image-level uncertainty through adversarial estimation. ORCR [15] measures classification and localization inconsistencies without additional network overhead, while HUALE [26] employs a two-stage framework with novel uncertainty metrics.

**Diversity-Based Active Learning:** Diversity-based approaches aim to select a representative subset of the data that covers the full feature space. Sener and Savarese [32] reformulate AL as a core-set selection problem, showing that minimizing the “core-set loss” is equivalent to solving a k-Center problem and achieving performance close to full dataset training. In parallel, Wang et al. [38] extend traditional multiple-instance AL by integrating diversity criteria — via kernel k-means clustering and fuzzy rough set theory — with informativeness measures, thereby ensuring that selected samples capture a broader range of features.

**Dual Objective Active Learning:** Dual objective approaches integrate both uncertainty and diversity to maximize the value of queried samples. PPAL [44] calibrates instance uncertainty with difficulty coefficients and enforces diversity through category-conditioned similarity matching. USDM [46] combines uncertainty sampling with graph-based random walks, while Agarwal *et al.* [2] incorporate contextual diversity using KL-divergence. DivProto[42] employs Entropy-based Non-Maximum Suppression to filter redundant predictions and refines diversity by decomposing it into intra-class and inter-class components.

**Segment Anything-based Annotations Tools:** Various projects have integrated SAM into OD workflows with mixed success. Label Studio [14] employs SAM but requires user-supplied inputs like points or prompts [29]. Napari-SAM [27], demands manual guidance and cannot independently generate precise annotations. Autodistill [31] combines SAM with Grounding DINO [23] for text-prompted annotations but struggles with RS data accuracy. Similarly, YOLO+SAM [3] has users draw initial boxes that SAM refines into masks before fine-tuning, creating an iterative but still manually-dependent process.

**Cold Start Active Learning:** In cold start scenarios, where initial labeled examples are scarce, the model struggles to accurately represent the full data distribution, complicating the selection of informative samples. To address the challenges of initial bias and limited annotations, cold start AL methods focus on robust sample selection at the onset of learning. ALWOD [39] integrates weakly and semi-supervised OD by employing an auxiliary image generator to warm-start the detector and an acquisition function that fuses student–teacher disagreement with image uncertainty, thereby reducing annotation requirements. Complementing this approach is Chen *et al.* [8], whose study tackles early-stage inefficiencies by leveraging self-supervised contrastive learning and K-means clustering to force label diversity and selection of representative samples, outperforming both traditional AL strategies and random sampling.

**Limitations of Existing Approaches:** Existing SAM-based methods fail to fully leverage RS imagery characteristics such as low occlusion in orthorectified, cloud-free images. While using points and bounding boxes speeds up mask prediction, annotators must still assign labels for each object. Additionally, no methods address the common scenario in RS where many images may be empty of target classes—an issue exacerbated in cold start situations where empty images contribute little to training. Our approach harnesses a pretrained FM (e.g. SAM) for dense mask extraction to automatically align masks with classes (predicted by the OD model), allowing users to simply confirm or adjust results, and employs dual-source uncertainty estimation where only images with objects confirmed by both the detector and the *FM* are proposed for annotation.

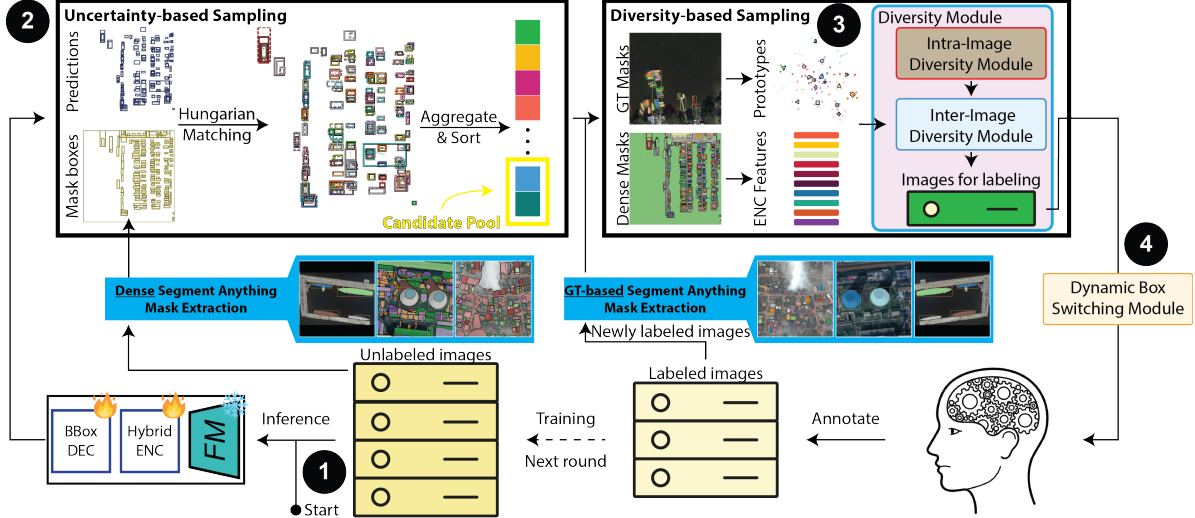


Figure 1. Our framework iteratively selects images by combining  $FM$ -based uncertainty and diversity sampling. We extract masks from all images using a  $FM$  (SAM [18] in our case), train an object detector on a small labeled set, and predict on unlabeled images. Uncertainty is estimated by matching detector boxes with  $FM$  masks. High-uncertainty images undergo diversity sampling to ensure both intra- and inter-image diversity. Finally, high-confidence predictions are selected with  $FM$ -derived boxes to accelerate annotation.

### 3. Methodology

Figure 1 illustrates our AL framework. Let  $\mathcal{I}_l^{(0)}$  denote a small initial set of labeled images, used to train the OD model  $OD^{(0)}$  at step 0. Let  $\mathcal{I}_u$  denote the set of unlabeled images to sample from at each step. To select the images  $I_{sel}$  to annotate in the next step  $s$ , we perform the following procedures. **1** For each image in  $\mathcal{I}_u$  we: i) use a pretrained segment anything-type of model  $SA$  (e.g., SAM) to densely extract a set of masks  $M_{SA}$ , and collect the bounding boxes  $B_{SA}$  associated with these masks – it suffices to run this process just once and store  $B_{SA}$  and  $M_{SA}$  in memory for each image; ii) perform OD inference using the latest model  $OD^{(s-1)}$ , collecting a set of boxes  $B_{OD}$ .

**2** We employ a novel *Dual-Source Uncertainty Estimation* mechanism, detailed in Sec 3.1, this strategy matches bounding-boxes across  $B_{SA}$  and  $B_{OD}$ , exploiting their matching cost (i.e., similarity) as a quality measure  $\sigma_{match}$ . This quality measure is then combined with the detectors’ prediction confidence  $\sigma_{OD}$  by means of a harmonic mean, yielding a combined uncertainty score  $\mu$  for each detected object. This uncertainty is averaged across all detected objects in an image. For a  $\theta$  budget expansion rate, we select the  $n \times \theta$  most uncertain images to compose a set of candidate images  $I_{cond}$ .

**3** We then run  $I_{cond}$  through a novel *Mask-Guided Diversity Estimation* scheme to guide diversity sampling, which is detailed in Sec 3.2. Inspired by DivProto [42], our strategy constructs class prototypes based on features extracted by a pretrained model and select samples such that both intra-class and inter-class diversity are promoted. In

contrast to DivProto, we rely on stronger features extracted by a modern FM, and exploit the masks  $M_{SA}$  pre-extracted in procedure 1 to mask the feature map and hence obtain higher-quality, object-specific feature representations for each object. The  $n$  images from  $I_{cond}$  with highest overall diversity score are selected to compose the  $I_{sel}$  set for annotation.

**4** Once the oracle annotates  $I_{sel}$ ,  $OD$  is fine-tuned on  $\mathcal{I}_l^{(s+1)} = \mathcal{I}_l^{(s)} \cup I_{sel}$ , and the AL process repeats from procedures 1-4 until the total budget or  $\mathcal{I}_u$  is exhausted.

**Semi-automated labeling:** For an OD model leveraging an FM-like backbone, at early stages the classification performance tends to be already good, but the localization of objects is less reliable. To address such cold start behavior, we introduce the Dynamic Box Switching Module (DBS), which replaces the bounding boxes predicted by  $OD$  with SAM mask bounding boxes for annotation. Details about the DBS module are provided in Sec 3.3, and its benefits are validated in a user study (see Sec 5).

#### 3.1. Dual-Source Uncertainty Estimation

For the DSUE, we first perform Hungarian matching between boxes in  $B^{SA}$  and  $B^{OD}$ , ignoring class labels. From the Hungarian matching process, a cost matrix  $C$  corresponding to the minimal matching cost between paired boxes in  $B^{SA}$  and  $B^{OD}$  is obtained for each image. In this one-to-one process, unpaired entries in  $B^{SA}$  and  $B^{OD}$  are removed, and a set of  $\hat{B}^{Match}$  paired boxes is obtained. We transform each cost into a matching quality measure,  $\sigma_{match}$ , using an exponential decay function ( $\sigma_{match} =$

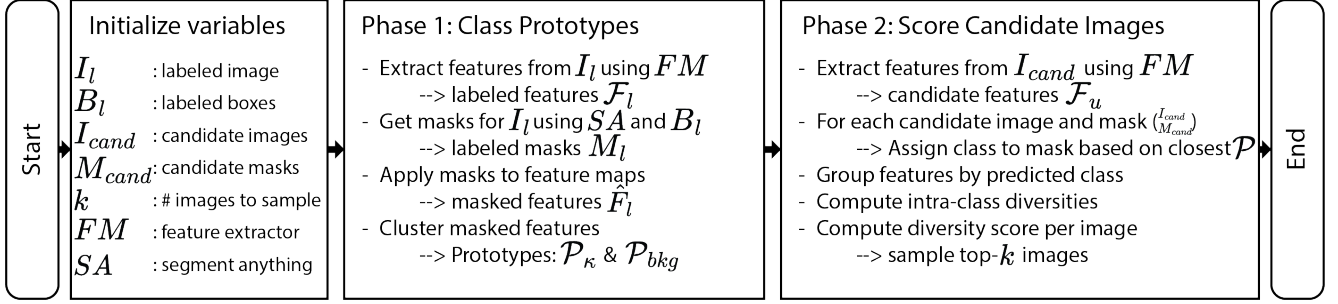


Figure 2. A flowchart highlighting the mask-guided diversity estimation. In phase 1, the class prototypes are constructed by first extracting the features ( $\mathcal{F}_l$ ) and masks ( $M_l \leftarrow SA(I_l, B_l)$ ), then extracting the mask based features and finally clustering these masked-based features per class. Phase 2 then utilizes these prototypes to cluster the features extracted from  $I_{cand}$ , based on the precomputed masks ( $M_{cand}$ ). Based on the clustered features we then compute the intra-class diversities from which we finally obtain the diversity score for the image.

$exp(\mathcal{H})$ ). This quality measure is then combined with the detectors prediction confidence  $\sigma_{OD}$  by means of a harmonic mean, yielding a combined uncertainty score  $\mu$  for each pair as per:

$$\mu = 1 - \frac{2\sigma_{OD}\sigma_{match}}{\sigma_{OD} + \sigma_{match} + \epsilon}, \quad (1)$$

with  $\epsilon = 10^{-6}$  to ensure numerical stability. Finally, the per-image uncertainty is obtained by averaging across all predicted uncertainties. We then sample the  $k$  highest uncertainty images.

### 3.2. Mask-guided Diversity Estimation

For diversity sampling, we take inspiration in DivProto [42] but under a novel approach where we exploit pre-computed masks to refine the feature-based characterization of each detection. The strategy is summarized in Fig. 2 and detailed below.

**Phase 1: Prototypes construction:** At each step  $s$ , for each labeled image in  $\mathcal{I}_l^{(s)}$  we extract features  $\mathcal{F}_l$  using a pretrained  $FM_{OD}$ . In our case, we exploit the same  $FM_{OD}$  used as backbone for our  $OD$  model. Using the labeled bounding boxes within an image as prompts, we deploy  $FM_{SA}$  on  $\mathcal{I}_l^{(s)}$  to extract a set  $M_l$  of one mask per object. We then mask the feature map  $\mathcal{F}_l$  with  $M_l$  to obtain a collection of features  $\hat{\mathcal{F}}_l$  for each labeled object.

We then build a set of  $\mathcal{P}_\kappa$  prototypes for each object category  $\kappa$  by clustering these features using the k-means++ algorithm with cosine similarity as the distance metric. Following DivProto [42], we opt for  $p = 5$  prototypes per category to capture intra-class diversity. We also construct a set of background prototypes  $\mathcal{P}_{bkg}$  by randomly sampling from the subset of the precomputed masks  $M_{SA}$  across all images that do not overlap with any ground truth bounding box.

**Phase 2: Intra-class and inter-class diversity estimation:** Similarly and in parallel to the process of prototype

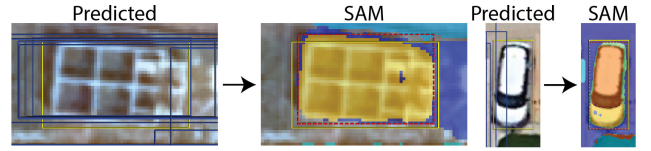


Figure 3. DBS module: Switching the predicted mask to the SAM masks to boost IoU, while keeping the box classification. Blue: Model predictions, Dotted Red: SAM mask bounding box, Yellow: Ground truth.

construction, we extract features  $\mathcal{F}_u$  (again using  $FM_{OD}$ ) from the unlabeled images by using the masks associated with the bounding boxes  $\hat{B}$  identified through the Hungarian matching. Each candidate mask is matched to a class based on cosine similarity between its features and the class prototypes. To identify the most diverse candidate samples to be annotated, we compare margins between the closest and second closest prototypes. We group detections by their predicted class, and quantify intra-class diversity as the maximum pairwise distance within each group. Then, we define the image’s overall diversity score as the mean of intra-class diversities within it. Finally, images with the highest diversity scores are selected, ensuring that our sampler prioritizes unlabeled samples that are both uncertain but also exhibit a broad range of distinct features relative to the existing labeled set.

### 3.3. Dynamic Box Switching Module

To mitigate the cold-start problem and accelerate the AL process, the implementation of semi-supervised labeling for objects classified with high certainty by the uncertainty sampler presents a potential solution. While this approach demonstrates efficacy in classification frameworks [21], where a single object per image is assumed, it exhibits limitations in OD contexts, where certainty of individual objects does not necessarily correlate with overall image certainty. Nevertheless, pre-labeling these high-confidence ob-

jects can expedite the annotation workflow. This approach introduces a significant challenge: in cold-start scenarios, initial localization exhibits considerable noise, despite rapid classification accuracy gains attributable to pretrained backbone architectures. To address this limitation, we leverage precomputed bounding boxes from SAM, substituting predicted bounding boxes with the SAM-generated alternatives with the highest overlap, thereby reducing localization noise while preserving the OD model’s classification capabilities. Examples are highlighted in Fig. 3.

## 4. Experimental Setup & Results

**Backbones & OD Model:** For the evaluation of the AL model, we employ the RTDETRv2 model [25] in combination with remote sensing FMs as backbones. RTDETRv2 was selected for quicker retraining on new data within the AL cycles ( $\approx 114s$  to finetune on 10 samples for 30 epochs in our user study, A100-PCI-E-40GBs) and based on experiments indicating that a frozen FM with a hybrid encoder neck performs equivalently to a non-frozen backbone on the DIOR [50], FAIR1M [34], and HRSC2016 [35] datasets. The backbone models used in this study are listed in Table 1, with an ablation of their performances described in Sec 4.2. Following [25] we extract three feature maps from each encoder.

**Datasets:** We evaluate our framework across five datasets: *DIOR*, which encompasses a diverse range of object categories; *DOTAv2*, which provides annotations for extremely small objects ( $\leq 10px$ ); *FAIR1M*, a fine-grained dataset; *HRSC2016*, dedicated to ship detection; and our proprietary *WaffleHomes* dataset. While DIOR and HRSC2016 come with predefined train/val/test splits, the DOTAv2 test split lacks labels, so we adopted the splits proposed by Lee *et al.* [20]. We split the FAIR1M dataset from [7] into randomly sampled 40/20/20% splits.

Our WaffleHome dataset consists of 363 images ( $1024 \times 1024px$ ) capturing “waffle homes”, which are defined as roofless buildings under construction, with visible interiors from overhead that yield grid-like patterns reminiscent of waffles. The localization of waffle homes can provide insights into typical floor plans and verification of building construction codes [41]. We select the annotated images

within a region of interest (see Figure 9) and then randomly divide the data into 40/20/40% splits.

**Model settings:** we largely adhere to the default hyperparameters and training settings from RTDETRV2 [25]. However, we observed that training for 36 epochs rather than the standard 72 is sufficient for our limited data experiments, with longer training being prone to overfitting. By default, we employ a ResNet50 backbone pretrained on FMOW, which provides an optimal balance between accuracy and speed. All experiments were conducted on a single Tesla V100-SXM3-32GB. For the uncertainty-based sampling we utilize a budget expansion ratio ( $\theta = 6$ ). For our diversity sampler, we use 5 prototypes per class and average the features for each mask.

### 4.1. Experimental Results

Figure 4 presents a quantitative comparison between our approach, which we refer to as AL4FM-OD for shortness, and established state-of-the-art AL methodologies. The DBS module was excluded from this analysis as it is primarily designed for annotation acceleration. Instead, detailed results for the DBS module are presented in the subsequent paragraphs.

Each method underwent five iterations with results averaged. The entropy-based sampling baseline quantifies image uncertainty by summing classification entropy of all detected objects, following [44]. This approach underperforms when trained with minimal data, scoring below random sampling across all datasets. DivProto exhibits inconsistent performance: it is effective on Wafflehome and DOTAv2, ineffective on HRSC2016, and demonstrates median performance on DIOR and FAIR1M, indicating cold start limitations. CoreSet shows median performance on most datasets except FAIR1M where it excels. While PPAL outperforms our method on DOTAv2, possibly due to the presence of small objects as discussed below, it achieves only median results on HRSC2016 and FAIR1M, and underperforms on the Wafflehome dataset due to selecting numerous empty images.

Our AL4FM-OD method demonstrates consistent performance improvement across most datasets, achieving significant gains: on DIOR from  $4.89mAP$  (PPAL) to  $10.68mAP$ , on HRSC2016 from  $47.79mAP$  (RANDOM) to  $50.52mAP$ , and on Waffle Homes from  $36.04mAP$  (DivProto) to  $42.05mAP$ . For datasets with objects more challenging to segment, AL4FM-OD performs comparably to the best methods during initial (cold-start) AL iterations ( $<100$  samples), with minimal performance differences in later stages:  $9.24mAP$  (PPAL) versus  $7.93mAP$  on DOTAv2 and  $12.12mAP$  (CoreSet) versus  $12.07mAP$  on FAIR1M. In summary, while previous methods exhibit dataset-specific inconsistency, our approach maintains robust performance stability across diverse datasets.

Table 1. Backbone models and pretraining datasets.

Model	Pretrained on Dataset
SWIN Transformer	SATLAS [6]
ResNet50 [4], ScaleMAE [30]	FMOW [9]
DOFA [43]	SATLAS, FiveBillionPixels [36] HySpecNet11k [12]
OREOLE [11]	MillionAID [24]
DinoV2 (ViT) [28]	ImageNet1K [10]

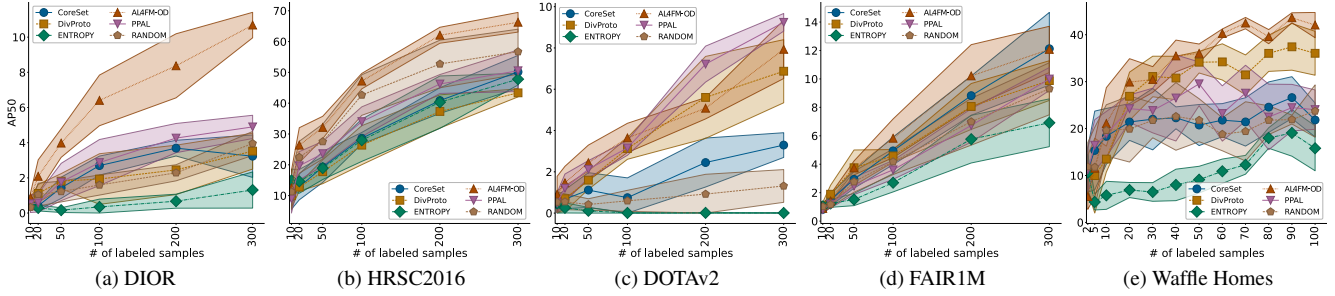


Figure 4. Comparison between the our method and the SOTA AL methods for OD on five remote sensing datasets.

**Dynamic Box Switching Results:** Figure 5 illustrates the performance of our DBS module. For evaluation, we employed the Recall@100 metric with a more stringent IoU threshold of 75%, reflecting the practical scenario where users would only accept bounding boxes that tightly encompass objects. The results demonstrate that during the initial AL phases ( $< 50$  annotated images), our model generates substantially more accurate bounding boxes when replacing predicted boxes with SAM-derived alternatives. Specifically, we observed a  $4\times$  improvement for the WaffleHome dataset during the first annotation round and a  $2\times$  improvement for the HRSC2016 dataset. However, we note DBS benefits diminish after approximately 50 annotated images. For the WaffleHome dataset, performance differences become negligible, while for the HRSC2016 the DBS module plateaus at  $\approx 58\%$  R@100, whereas the standard OD model continues improving to  $\approx 85\%$ . Overall, it is clear that with the DBS module and a pretrained backbone, our model is able to predict very accurate bounding boxes with very little data.

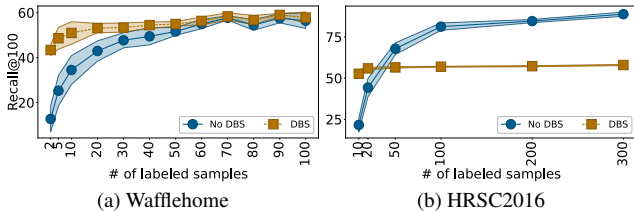


Figure 5. Comparing model predictions with/without DBS bounding box replacement. RTDETRv2, ResNet50 (FMOW pretrained).

## 4.2. Ablations Studies

To show the effectiveness of our proposed Dual-Source Uncertainty Estimation (DSUE) scheme we first compare it to random and entropy-based sampling in Table 2. We find that DSUE achieves better performance compared to an entropy-based uncertainty-sampling counterpart, showcasing its effectiveness. We also compare the proposed Mask-guided Diversity Estimation (MDE) with three alternative diversity samplers: CoreSet, DivProto, and the

Table 2. Ablation studies of the Dual-Source Uncertainty Estimation (DSUE) using RTDETRv2, with the ResNet50 backbone, trained on the DIOR dataset. Category Conditioned Matching Similarity (CCMS) refers to the diversity sampler of PPAL.

Stage 1	Stage 2	mAP on # of labeled images				
		10	50	100	200	300
Random		$0.3\pm 0.2$	$1.2\pm 0.6$	$1.6\pm 0.6$	$2.3\pm 0.5$	$4.0\pm 0.7$
Entropy	None	$0.5\pm 0.3$	$0.2\pm 0.1$	$0.4\pm 0.4$	$0.7\pm 0.4$	$1.3\pm 1.1$
	DSUE	$0.9\pm 0.4$	$2.4\pm 0.9$	$3.8\pm 0.5$	$4.3\pm 0.9$	$5.4\pm 0.8$
DSUE	CoreSet	$0.8\pm 0.4$	$2.6\pm 0.7$	$4.3\pm 1.1$	$4.6\pm 0.3$	$4.8\pm 0.6$
	DivProto	$0.9\pm 0.4$	$1.1\pm 0.5$	$0.7\pm 0.5$	$0.9\pm 0.5$	$0.9\pm 0.4$
DSUE	CCMS	$0.8\pm 0.4$	$1.7\pm 1.2$	$3.7\pm 0.4$	$4.6\pm 0.8$	$5.2\pm 0.8$
	MDE	$0.9\pm 0.3$	$4.1\pm 0.8$	$6.4\pm 1.4$	$8.4\pm 1.8$	$10.7\pm 0.7$

Category Conditioned Matching Similarity (CCMS) proposed by PPAL. We confirm the findings of the authors of PPAL [44] that global similarity-based diversity sampling methods (e.g., DivProto) perform worse than random sampling in multi-instance images; however, we find that mask-based similarity measurements are even more suitable to measure the similarity in multi-instance images.

**Generalization to different backbones.** In Table 3, we do an ablation study of different (frozen) backbones and present the results. We find that our approach can generalize well to different backbones. Only ScaleMAE underperforms, likely due to the smaller dataset (fMOW) and lack of stronger features due to its MAE-only pretraining. Surprisingly, the (frozen) DinoV2, pretrained on ImageNet1K, performs significantly better on the remote sensing DIOR dataset than its counterparts pretrained on remotely sensed data. This highlights the effectiveness of DINOv2’s self-supervised pretraining regime.

**Expansion Ratio.** Table 4 presents the ablation for the expansion ratio  $\theta$ , using RTDETRv2 with the ResNet50 backbone pretrained on FMOW. One can see that increasing  $\theta$  leads to better results up until  $\theta = 6$ , after which the time requirement becomes a limitation ( $\theta = 2$ : 27s,  $\theta = 4$ : 30s,  $\theta = 6$ : 45s,  $\theta = 7$ : 87s) for the real-time applicability.

**SAM segmentation abilities.** Effective performance of our AL framework fundamentally relies on the Segment

Table 3. Comparison of different backbones in combination with RTDETRv2 on DIOR. DOFA, ScaleMAE and DinoV2 results averaged over 3 runs.

Backbone	mAP on # of labeled images				
	10	50	100	200	300
ResNet50	0.9±0.3	4.1±0.8	6.4±1.4	8.4±1.8	10.7±0.7
ScaleMAE	0.1±0.1	1.2±0.1	2.2±0.7	3.6±1.1	4.3±0.9
SwinBase	0.5±0.5	1.9±2.4	2.8±3.2	5.3±4.4	10.9±3.1
DOFA	0.3±0.1	1.8±0.3	2.6±0.4	6.5±0.3	8.9±0.9
OREOLE	1.1±0.4	2.1±0.1	5.7±0.3	10.8±1.2	12.0±0.7
DinoV2	0.8±0.6	2.5±0.5	5.4±0.3	10.2±0.7	15.4±0.8

Table 4. Ablation study using DIOR + RTDETRv2 (ResNet50) on how the budget expansion ratio  $\theta$ , which determines the size of the candidate pool for the first stage.

Expansion ratio ( $\theta$ )	mAP on # of labeled images				
	10	50	100	200	300
1	0.9±0.4	2.4±0.9	3.8±0.5	4.3±0.9	5.4±0.8
2	0.8±0.3	2.8±0.6	3.9±0.7	5.1±0.3	6.5±0.8
3	0.7±0.4	3.2±0.7	3.5±0.9	5.7±1.5	6.0±1.2
4	0.7±0.4	2.4±0.7	4.2±1.3	3.9±1.7	7.7±1.4
5	0.7±0.3	3.2±0.7	4.3±0.7	6.3±0.7	8.9±1.4
6	0.9±0.3	4.1±0.8	6.4±1.4	8.4±1.8	10.7±0.7



Figure 6. Example of SAM limitations on DIOR dataset. The terminals lack coherent structure, thus SAM is not able to segment. Dotted red: SAM bounding boxes; Yellow: ground truth.

Anything-type of model (e.g., SAM) accurately segmenting target objects. To quantify SAM’s segmentation abilities for different objects as captured in remote sensing imagery, we compute recall rates at different IoU thresholds for the 5 datasets considered across this work, with different initialization grid sizes. As summarized in Figure 8, recall rates for grid-size 64 are  $\geq 50\%$  at  $IoU = 0.5$  for most datasets, which represents a solid performance for SAM’s exploitation within our AL framework.

**Limitations.** In more details, we observed that SAM excels for objects that exhibit clear, well-defined boundaries. This is typically the case when objects of interest are sufficiently large relative to the ground sampling distance. Examples include *aircraft*, *storage tanks* composing the DIOR, DOTAv2 and FAIR1M datasets, our waffle homes dataset, and *ships* composing the HRSC2016 dataset. In high-resolution imagery (GSD  $\approx 50\text{cm}$ ), these structures tend to yield precise masks due to their high contrast and distinctive structural

features. In turn, SAM underperforms for segmentation of smaller objects like *windmills*, *small chimneys*, or *small vehicles* present in the DOTAv2 and FAIR1M datasets, as they may fall between the dense extraction grid points set. Additional SAM failure cases include structures defined by more contextual features, such as *bridges*, *golf courses*, or *train stations* present in DIOR, DOTAv2 and FAIR1M, as they present ambiguous boundaries. In those few scenarios, such segmentation limitations result in suboptimal mask extraction that can cascade through the pipeline, undermining both uncertainty computation and final annotation quality.

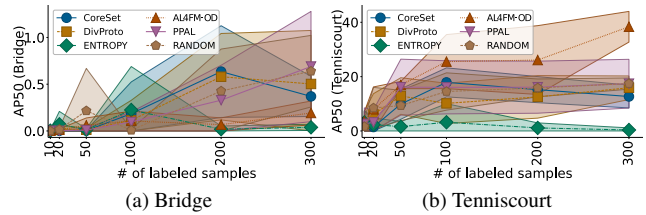


Figure 7. AP of bridge and tennis court classes from DIOR.

## 5. User Study

**Setup:** This study employed eight human geography experts on population density estimation. Each 1-hour session paired two experts annotating the same dataset independently, using a PyQT front-end with REST API server communication that allows bounding box creation. The backend system performed all AL procedures described in Sec 3. The study used 145 training images from the waffle home dataset. We compared our proposed method against a baseline random annotation approach. Five randomly selected pre-annotated images were used to pretrain the model, with identical initial weights and images across experiments. A between-subjects design assigned participants to either manual annotation or AL conditions. Both groups had options to draw boxes (two clicks per box), as well as accepting/rejecting model predictions. All participants had comparable image annotation experience and familiarity with the dataset. Four sessions were conducted, each with two participants (one per condition). Sessions included a 10-minute introduction followed by 50 minutes of annotation. Participants annotated 10 images before the AL system predicted the next 10 images.

**Objective Performance:** Our AL approach improved annotation efficiency compared to random selection. First, users annotated more images (11.75 vs. 10.67) and produced more total annotations (235 vs. 209), while reducing manually drawn bounding boxes (185 vs. 196). This improvement stems from our DBS method’s superior bounding box prediction accuracy, whereas the random approach using model predictions directly had poor localization due to cold start. Mann–Whitney U test has been conducted and

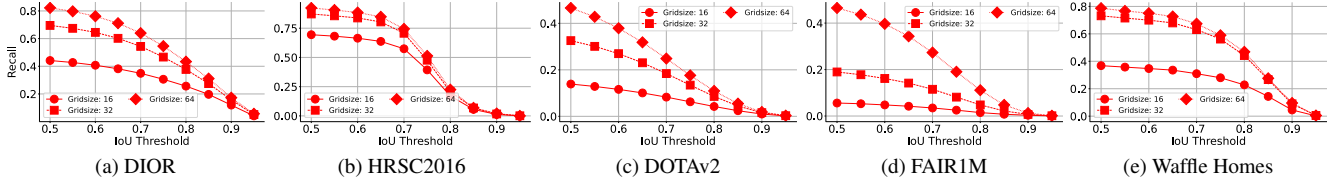


Figure 8. Recall vs. different IoU thresholds of SAM on 5 different remote sensing datasets.

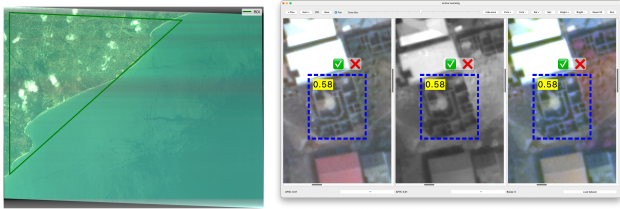


Figure 9. (Left) Region of interest. (Right) AL Interface. *Dotted bounding box*: model prediction with confidence score (yellow). Accept/reject buttons above. Users can zoom, pan, and annotate in synchronized views: RGB, NIR, *Costal Blue*, *Yellow*, and *Red Edge* channels.

Table 5. User Study Interaction Results: Our AL method with DMS vs. random annotation without DMS.

	Random	Ours
Avg. # of image annotated	10.67	<b>11.75</b>
Avg. # of accepted boxes	13.33	<b>51.75</b>
Avg. # of drawn boxes	196	<b>185</b>
Avg. total annotations	209	<b>235</b>

Table 6. User Study Results: Average Precision with an IoU of 0.5 and 0.75 and the Average Recall 100. The same model and hyperparameters were used for all datasets.

	AP50	AP75	AR100
Random	0.28±0.03	0.11±0.05	0.35±0.04
Ours	<b>0.34±0.04</b>	<b>0.17±0.05</b>	<b>0.43±0.03</b>

we find that the difference in the on average accepted boxes is statistically significant ( $U = 0$ ,  $p = 0.0294$ ). Moreover, we also analyze the final recall and precision of RT-DETR using the ResNet50 backbone trained on each of the annotations. Results in Table 6 show that within the same timeframe the users with our method were able to achieve a roughly 14% higher AP50, a 3% increase in AP75 and a 8% increase in AR100.

**Limitations of the user study:** Each annotator was allotted only one hour, which precluded the possibility of a within-subject design, as sessions shorter than 30 minutes would not yield sufficiently meaningful annotations. In addition, a larger sample size is required to further study the generalizability of the proposed approach.

## 6. Conclusion

We introduced a real-time AL framework that leverages modern FMs to accelerate dataset annotation for remote sensing OD. Our framework exploits a Segment Anything model for the extraction of candidate masks and boxes, and a pretrained FM for powering our object detector as well as extracting features that guide the sampling selection process. Our contributions include three novel components. Our dual-source uncertainty sampling strategy and mask-guided diversity estimation scheme effectively enable uncertainty and diversity-based sample selection, while our Dynamic Mask Switching Module effectively addresses the cold start problem to enable effective semi-automated annotation. Extensive experiments and a user study confirm that our approach reduces annotation effort while improving detection performance. Future work will target enhancing SAM’s segmentation for small or ambiguous objects and performing more extensive user studies.

## Acknowledgments

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>). This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725. This research was supported in part by an appointment to the Oak Ridge National Laboratory GRO Program, sponsored by the U.S. Department of Energy and administered by the Oak Ridge Institute for Science and Education.

## References

- [1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *Computer Vision – ECCV 2020*, pages 137–153, Cham, 2020. Springer International Publishing. 1
- [2] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVI*, pages 137–153. Springer, 2020. 2
- [3] amine0110. Yolo-sam for medical imaging. <https://github.com/amine0110/yolo-sam>, 2025. Accessed: 2025-03-05. 2
- [4] Kumar Ayush, Burak Uzket, Chenlin Meng, Kumar Tanmay, Marshall Burke, David B. Lobell, and Stefano Ermon. Geography-aware self-supervised learning. *CoRR*, abs/2011.09980, 2020. 5
- [5] Leah Bar, Boaz Lerner, Nir Darshan, and Rami Ben-Ari. Active learning via classifier impact and greedy selection for interactive image retrieval. *CoRR*, abs/2412.02310, 2024. 1
- [6] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Atlaspretrain: A large-scale dataset for remote sensing image understanding. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 16726–16736. IEEE, 2023. 5
- [7] blanchon. Fair1m. <https://huggingface.co/datasets/blanchon/FAIR1M>, 2021. Accessed: 2025-03-05. 5
- [8] Liangyu Chen, Yutong Bai, Siyu Huang, Yongyi Lu, Bihan Wen, Alan L. Yuille, and Zongwei Zhou. Making your first choice: To address cold start problem in vision active learning. *CoRR*, abs/2210.02442, 2022. 2
- [9] Gordon A. Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6172–6180. Computer Vision Foundation / IEEE Computer Society, 2018. 5
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, 20-25 June 2009, Miami, Florida, USA, pages 248–255. IEEE Computer Society, 2009. 5
- [11] Philippe A. Dias, Aristeidis Tsaris, Jordan Bowman, Abhishek Potnis, Jacob Arndt, Hsiuhan Lexie Yang, and Dalton D. Lunga. Oreole-fm: successes and challenges toward billion-parameter foundation models for high-resolution satellite imagery. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems, SIGSPATIAL 2024, Atlanta, GA, USA, 29 October 2024 - 1 November 2024*, pages 597–600. ACM, 2024. 5
- [12] Martin Hermann Paul Fuchs and Begüm Demir. Hyspecnet-11k: a large-scale hyperspectral dataset for benchmarking learning-based hyperspectral image compression methods. In *IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2023, Pasadena, CA, USA, July 16-21, 2023*, pages 1779–1782. IEEE, 2023. 5
- [13] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, page 1183–1192. JMLR.org, 2017. 1
- [14] HumanSignal. Label studio. <https://labelstudio.io/>, 2025. Accessed: 2025-03-05. 2
- [15] Ming Jing, Zhilong Ou, Hongxing Wang, Jiaxin Li, and Ziyi Zhao. Object recognition consistency in regression for active detection. *Mach. Vis. Appl.*, 35(5):121, 2024. 2
- [16] Ajay J. Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379, 2009. 1
- [17] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. (arXiv:2306.01567), 2023. arXiv:2306.01567 [cs]. 2
- [18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3992–4003. IEEE, 2023. 3
- [19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. (arXiv:2304.02643), 2023. arXiv:2304.02643 [cs]. 2
- [20] Chunggi Lee, Seonwook Park, Heon Song, Jeongun Ryu, Sanghoon Kim, Haejoon Kim, Sérgio Pereira, and Donggeun Yoo. Interactive multi-class tiny-object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14116–14125. IEEE, 2022. 5
- [21] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 2013. 4
- [22] Hojun Lee, Suyoung Kim, Junhoo Lee, Jaeyoung Yoo, and Nojun Kwak. Coreset selection for object detection. 1
- [23] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, pages 38–55, Cham, 2025. Springer Nature Switzerland. 2
- [24] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 14:4205–4230, 2021. 5

- [25] Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. Rt-detr2: Improved baseline with bag-of-freebies for real-time detection transformer. *CoRR*, abs/2407.17140, 2024. 5
- [26] Tai Nguyen, Khoa Nguyen, Thanh Nguyen, Tri Nguyen, Anh Nguyen, and Karrman Kim. Hierarchical uncertainty aggregation and emphasis loss for active learning in object detection. In *IEEE International Conference on Big Data, Big-Data 2023, Sorrento, Italy, December 15-18, 2023*, pages 5311–5320. IEEE, 2023. 2
- [27] Applied Computer Vision Lab (Helmholtz Imaging) & Division of Medical Image Computing (DKFZ). Segment anything model (sam) in napari. <https://github.com/MIC-DKFZ/napari-sam>, 2025. Accessed: 2025-03-05. 2
- [28] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafranec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. 5
- [29] Lucas Prado Osco, Qiusheng Wu, Eduardo Lopes de Lemos, Wesley Nunes Gonçalves, Ana Paula Marques Ramos, Jonathan Li, and José Marcato Junior. The segment anything model (sam) for remote sensing applications: From zero to one shot, 2023. 2
- [30] Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 4065–4076. IEEE, 2023. 5
- [31] Roboflow. autodistill. <https://docs.autodistill.com/>, 2025. Accessed: 2025-03-05. 2
- [32] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 2
- [33] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. (arXiv:1708.00489), 2018. arXiv:1708.00489 [stat]. 1
- [34] Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, Martin Weinmann, Stefan Hinz, Cheng Wang, and Kun Fu. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *CoRR*, abs/2103.05569, 2021. 5
- [35] Gang Tang, Shibo Liu, Iwao Fujino, Christophe Claramunt, Yide Wang, and Shaoyang Men. H-YOLO: A single-shot ship detection approach based on region of interest preselected network. *Remote. Sens.*, 12(24):4192, 2020. 5
- [36] Xin-Yi Tong, Gui-Song Xia, and Xiao Xiang Zhu. Enabling country-scale land cover mapping with meter-resolution satellite imagery. *CoRR*, abs/2209.00727, 2022. 5
- [37] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Trans. Cir. and Sys. for Video Technol.*, 27(12):2591–2600, 2017. 1
- [38] Ran Wang, Xi-Zhao Wang, Sam Kwong, and Chen Xu. Incorporating diversity and informativeness in multiple-instance active learning. *IEEE Trans. Fuzzy Syst.*, 25(6):1460–1475, 2017. 2
- [39] Yuting Wang, Velibor Ilic, Jiatong Li, Branislav Kisačanin, and Vladimir Pavlovic. ALWOD: active learning for weakly-supervised object detection. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 6436–6446. IEEE, 2023. 2
- [40] Yuting Wang, Velibor Ilic, Jiatong Li, Branislav Kisačanin, and Vladimir Pavlovic. Alwod: Active learning for weakly-supervised object detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 6436–6446, Paris, France, 2023. IEEE. 1
- [41] Carson Woody and Tyler Frazier. Waffle homes: Utilizing aerial imagery of unfinished buildings to determine average room size (short paper). In *12th International Conference on Geographic Information Science, GIScience 2023, September 12-15, 2023, Leeds, UK*, pages 85:1–85:6. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023. 5
- [42] Jiayi Wu, Jiayin Chen, and Di Huang. Entropy-based active learning for object detection with progressive diversity constraint. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 9387–9396, New Orleans, LA, USA, 2022. IEEE. 1, 2, 3, 4
- [43] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J. Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for earth observation, 2024. 5
- [44] Chenhongyi Yang, Lichao Huang, and Elliot J. Crowley. Plug and play active learning for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 17784–17793. IEEE, 2024. 2, 5, 6
- [45] Chenhongyi Yang, Lichao Huang, and Elliot J. Crowley. Plug and play active learning for object detection. (arXiv:2211.11612), 2024. arXiv:2211.11612 [cs]. 1
- [46] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G. Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *Int. J. Comput. Vis.*, 113(2):113–127, 2015. 2
- [47] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 93–102. Computer Vision Foundation / IEEE, 2019. 2
- [48] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 93–102, Long Beach, CA, USA, 2019. IEEE. 1

- [49] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5330–5339. Computer Vision Foundation / IEEE, 2021. [1](#), [2](#)
- [50] Yang Zhan, Zhitong Xiong, and Yuan Yuan. RSVG: exploring data and models for visual grounding on remote sensing data. *IEEE Trans. Geosci. Remote. Sens.*, 61:1–13, 2023. [5](#)