BOOTSTRAPPING RARE OBJECT DETECTION IN HIGH-RESOLUTION SATELLITE IMAGERY

Akram Zaytar[*] **Caleb Robinson** **Gilles Q. Hacheme** **Girmaw A. Tadesse**

Rahul Dodhia **Juan M. Lavista Ferres** **Lacey F. Hughey** **Jared A. Stabach**

Irene Amoke

ABSTRACT

Rare object detection is a fundamental task in applied geospatial machine learning, however is often challenging due to large amounts of high-resolution satellite or aerial imagery and few or no labeled positive samples to start with. This paper addresses the problem of bootstrapping such a rare object detection task assuming there is no labeled data and no spatial prior over the area of interest. We propose novel offline and online cluster-based approaches for sampling patches that are significantly more efficient, in terms of exposing positive samples to a human annotator, than random sampling. We apply our methods for identifying bomas, or small enclosures for herd animals, in the Serengeti Mara region of Kenya and Tanzania. We demonstrate a significant enhancement in detection efficiency, achieving a positive sampling rate increase from 2% (random) to 33%. This advancement enables effective machine learning mapping even with minimal labeling budgets, exemplified by an F_1 score on the boma detection task of 0.36 with a budget of 300 total patches.

1 INTRODUCTION

Rare object detection over remotely sensed (satellite, aerial, drone) imagery is a common task in geospatial machine learning with applications ranging from identifying damaged structures in post-disaster imagery [7; 10], renewable energy infrastructure mapping at country and global scales [6], to ecological studies like counting large mammals over vast landscapes [16]. The nature of such tasks is that they are often not accompanied by relevant labeled datasets – if such datasets existed over the area of interest (AOI), then there would not be a need to solve the task in the first place. As such, a first step in any rare object detection task is often manually finding *few* instances of the rare object, i.e., positive class examples, that can then be used to *bootstrap* a few-shot model-based approach [3; 18].

It is possible to incorporate known spatial priors into the *bootstrapping* process in some object detection tasks. For example, to find examples of damaged structures in post-disaster satellite imagery, it is possible to search over known existing structures instead of over the entire AOI, dramatically reducing the search space. However, in other problem instances, such as finding large mammals in high-resolution satellite/aerial imagery, the expected spatial distribution of the object of interest is less obvious.

Similarly, in some problem instances, it is possible to substitute related label datasets to bootstrap the modeling process. For example, OpenStreetMap (OSM) contains a large amount of data on solar photovoltaic plants and windmills [1] that can be joined with satellite imagery to train machine learning models that can then identify solar panels and windmills in imagery. However, such models will be limited by the coverage of existing labels and accompanying imagery. For example, the coverage of OSM data varies greatly by country, and imagery from different countries can vary greatly, therefore models trained under such conditions may easily fail to generalize over the entire

[*] Corresponding author: akramzaytar@microsoft.com

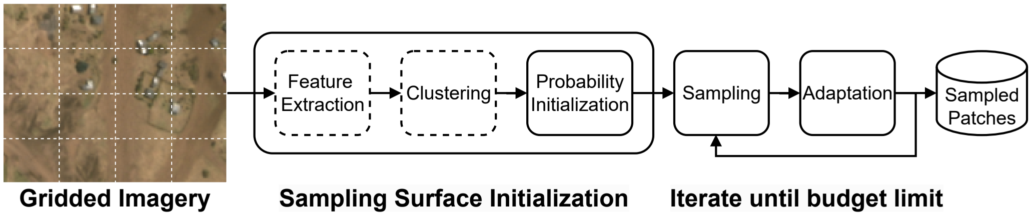
Feature Extraction
Clustering
Probability Initialization
Sampling
Adaptation
Sampled Patches
Gridded Imagery
Sampling Surface Initialization
Iterate until budget limit

Figure 1: **Bootstrapping rare object detection**. Given input imagery, we create a grid of image patches, initialize a sampling surface over the same grid, and sample iteratively from the surface looking for rare object instances until we hit a budget limit. Sampling strategies that surface these rare objects more frequently than random allow for quicker instantiating of model based methods for finding such objects.

AOI. Domain adaptation methods using synthetic imagery [6; 9], strong augmentation [19], and near-class object detectors [8] have all been proposed to alleviate this problem, but the fundamental problem of how to *start* modeling given a novel application and no labels remains an open problem in the state-of-the-art rare object detection literature.

In this paper, we formulate and address the basic problem of *bootstrapping a dataset of positive rare object samples under the assumptions of no initial labeled data and no spatial priors*. We propose novel offline and online clustering-based methods for selecting initial patches to annotate that only depends on imagery inputs, and a way to set the parameters of these methods without labels. Generally, our approach relies on the intuition that rare objects will, by definition, appear differently than their surroundings.

We apply our approach on the real-world problem of identifying Bomas from satellite imagery in the Serengeti Mara, a region of high ecological importance in Kenya and Tanzania. The distribution of Bomas is critical for various ecological and conservation efforts [15]. Our approach significantly increased the positive sampling rate from 2% to 33%, this sampling efficiency translates to the downstream task of Boma detection, achieving an F_1 score of 0.36 with just 300 initial labels.

The implications of this research extend beyond the specific case study, offering a scalable and efficient framework for rare object detection in various geospatial and remote sensing applications.

2 Problem Formulation

We assume that we are given a large unlabeled high-resolution satellite imagery scene, $\mathbf{X}$, that covers some AOI and a limited labeling budget, b (i.e., total area of imagery that can be labeled). We would like to detect some instances of a rare object class in $\mathbf{X}$ in order to *bootstrap* a modeling process. Specifically, we are looking for n^+ *positive* instances, i.e. examples of the object. While we look for the positives, we will annotate a number of *negative* instances, n^-, i.e. where the rare object is not present. Our objective is thus to minimize $\frac{n^-}{n^+}$ using b labelling iterations (or budget).

To achieve this, we split $\mathbf{X}$ into an $H \times W$ grid of non-overlapping image patches. For each grid cell, $X_{i,j}$, we assign a probability, $P_{i,j}$, thus initializing a *sampling surface*, $\mathcal{P}$ (i.e. a discrete probability distribution on a 2D grid). We discuss different strategies for initializing $\mathcal{P}$ in Section 3. We then use a sampling strategy to choose b patches for a labeler to annotate.

3 Methods

Our framework for bootstrapping rare object detection in $\mathbf{X}$ depends on two steps: *initializing the sampling surface*, and defining a *sampling strategy* to select b patches while, optionally, updating the sampling surface (as shown in Fig. 1).

Initializing the sampling surface: Naively, we can use equal weights to initialize the sampling surface as a `Uniform` baseline, i.e., $P_{i,j} = \frac{1}{H*W}, \forall i, j$. We further propose cluster-based approaches that extract feature vectors for each image patch $X_{i,j}$ using 3 different strategies: 1) RCF—uses random convolutional features [11] to extract color/texture feature vectors; 2) ColorStats—calculates the mean, standard deviation, minimum, and maximum values for each channel in $X_{i,j}$, providing a simple representation of the colors in a patch; 3) Pre-trained ResNet—employs a pre-trained ResNet-18 [5] to extract a feature representation. The clustering step assigns each grid cell into one of several clusters based on its feature representation. Here we utilize KMeans & Bisecting K-Means [14] with a hyperparameter for the number of clusters, K, and DBSCAN [2] with hyperparameters for the maximum distance between two samples to be neighbors, ϵ, and the number of neighbors of a point to be considered a core point, η. We describe an unsupervised method for choosing these hyperparameters based on the silhouette score [13] in Appendix Section 1.

We use the feature representations per patch to fit the clustering algorithm, resulting in K clusters, where each $X_{i,j}$ is assigned to one of the clusters. We then initialize $P_{i,j}$ based on the *inverse size of the cluster that $X_{i,j}$ is in*. Specifically, let $C_{i,j}$ be the size of the cluster that $X_{i,j}$ is in, then:

$$P_{i,j} \leftarrow \frac{1}{K} * \frac{1}{C_{i,j}}, \forall i, j \tag{1}$$

Sampling strategies: Once $\mathcal{P}$ is initialized, we can sample from it with methods from two broad categories: 1.) offline sampling where $P_{i,j}$ do not change based on incoming online annotations; and 2.) online sampling where we use the incoming annotations to change the probability surface. Specifically, we propose the Online and Proximity methods. In the Proximity sampling method, if we sample a positive, then we increase the probability of all neighboring patches within a set radius, r, by some weight, w, following the intuition that rare objects may be clustered in space. Similarly, in the Online method we increase the probability of patches that are in the same cluster as the sampled positive by w. In both cases, we sample from $\mathcal{P}$ without replacement and renormalize after reweighting based on observed positives.

With both online and offline methods, we sample until we exhaust our budget, after which we can use the set of found positives, negatives, and unlabeled patches to train a downstream machine learning model for object detection.

4 Case Study: Boma Mapping in the Serengeti Mara

We apply our methods in a case study for finding “bomas” in high-resolution satellite imagery captured over the Serengeti Mara region of Kenya and Tanzania. Bomas are temporary cattle enclosures used by local population that are relatively rare given the low population density of the region. Information on boma locations are crucial for identifying human-predator conflict hotspots and, as such, are used by organizations such as the Kenya Wildlife Trust.

We use 3 pansharpened WorldView-2 satellite scenes with 50cm/px spatial resolution. Combined, the images cover approximately $4,300\ km^2$ over three points in time (August 6th, 2022, January 2 & October 8 2020). We have polygon based labels for all bomas in these scenes from prior work which we use to run simulations.

Bootstrapping For our experiments we choose 3 *low-resource* labeling budgets of 300, 950, and 3000 image patches to evaluate different sampling strategies. For Proximity weighting, the radius was set at $r = 200\ m$, informed by previous knowledge of Boma settlement patterns. For both Proximity and Online, a value of $w = \max(P_{ij})$ (highest initial weight) was used. For clustering-based methods for initializing the sampling surface, we set hyperparameters (including which feature representation to use) based on an *unsupervised* method described in Appendix Section 1 that attempts to create clusters of rare objects. After setting the hyperparameter values, we compared the performance of various sampling approaches (i.e., Uniform, Proximity, and clustering methods in both *Offline* and *Online* scenarios using three labeling simulations per method. We report the number of positives samples found with each method in Table 1.

Downstream training Each bootstrapping method produces a set of positive and negative labels that were used to create training sets for the downstream task of semantic segmentation of bomas. Our aim is to assess the benefit provided by each method in the task of rare object detection. To this

Sampling Strategy	**300 Patches**			**950 Patches**			**3K Patches**		
	$\mathbf{n}^+(\%)$	**CE**	**RCE**	$\mathbf{n}^+(\%)$	**CE**	**RCE**	$\mathbf{n}^+(\%)$	**CE**	**RCE**
Random	1.6 ± 0.2	.00 ± .00	.06 ± .14	1.5 ± 0.3	.05 ± .07	.33 ± .19	1.8 ± 0.3	.62 ± .05	.65 ± .04
Proximity Weighting	14.0 ± 2.8	.00 ± .00	.15 ± .12	16.5 ± 6.4	.06 ± .07	.42 ± .19	14.0 ± 1.1	.55 ± .17	.76 ± .02
Static DBSCAN	3.3 ± 0.7	.00 ± .00	.00 ± .01	2.8 ± 0.2	.01 ± .01	.30 ± .16	4.0 ± 0.01	.60 ± .21	.70 ± .03
Adaptive DBSCAN	3.3 ± 0.4	.00 ± .00	.02 ± .02	5.2 ± 1.1	.04 ± .05	.31 ± .21	7.1 ± 0.2	.51 ± .23	.55 ± .24
Static KMeans	13.3 ± 3.3	.05 ± .08	.30 ± .23	28.9 ± 8.2	.13 ± .10	**.72 ± .04**	29.4 ± 4.2	.57 ± .21	.75 ± .04
Adaptive KMeans	**28.3 ± 8.6**	.03 ± .07	.18 ± .07	32.6 ± 7.7	.16 ± .13	.42 ± .17	28.1 ± 2.2	.46 ± .20	.65 ± .19
Static BKMeans	6.0 ± 0.8	.01 ± .01	.28 ± .18	6.1 ± 0.2	.19 ± .12	.71 ± .03	5.6 ± 0.2	**.73 ± .03**	**.77 ± .02**
Adaptive BKMeans	19.0 ± 1.0	**.06 ± .14**	**.36 ± .10**	**39.0 ± 8.6**	**.36 ± .25**	.70 ± .13	**33.6 ± 4.8**	.48 ± .23	.76 ± .01

Table 1: Results of various sampling strategies across different labeling budgets (300, 950, and 3,000 patches). We report the percentage of identified positive instances ($\mathbf{n}^+$) and downstream task performance of a U-Net model trained with the resulting samples for each sampling strategy. We report F_1 scores at an object level for models trained with Cross Entropy (CE) and Regularized Cross Entropy (RCE) losses. Note BKMeans refers to Bisecting KMeans.

end, we report object-level F_1 scores over unseen patches in Table 1. We maintained a consistent training setup across all methods, utilizing a U-Net [12] architecture with a ResNeXt50 $(32x4d)$ backbone [17], and training for up to 200 epochs. The AdamW optimizer was employed alongside data augmentation techniques, including flipping, rotation, and color jitter. Our training approach encompassed two loss configurations: 1) traditional cross-entropy loss (`CE`), and 2) regularized cross entropy (RCE) loss - a hybrid loss that combines cross-entropy for labeled pixels with an entropy minimization regularization term for unlabeled patches, formulated as $J(y, \hat{y}) = \rho(CE(y, \hat{y}) \odot Y_L) + (1 - \rho)(H(\hat{y}) \odot Y_{\bar{L}})$. Here, ρ represents the proportion of labeled pixels, H signifies entropy, Y_L is the binary mask for labeled pixels, and $Y_{\bar{L}}$ indicates the mask for unlabeled pixels. This is a semi-supervised training technique based on [4] that aims to minimize the class entropy of predictions over unlabeled pixels, which stabilizes training in low-label settings.

4.1 Results and Discussion

Our experiments demonstrate that all methods significantly surpass the uniform sampling baseline in identifying rare objects (see Table 1). Notably, `Online Bisecting KMeans` increased the sampling ratio from 2% (i.e., the object's density over the AOI) to 39% in the 950 patch scenario and 33% in the 3,000 patch scenario. In the 950 patch scenario, the `Online Bisecting KMeans` method found $26\times$ as many positives as uniform sampling—human annotators would need to process an additional 23,750 patches under the uniform sampling method to find an equivalent number of positives.

For the downstream task of detecting Bomas, the improved sampling strategies translated into notable gains in object detection performance, especially at lower labeling budgets. For instance, while uniform sampling fails with a budget of 300 patches, Online Bisecting KMeans achieves an F_1 score of 0.36. Overall, the cluster-based and online sampling methods achieve the best down-stream performance with a best result of 0.77 F_1 at the largest labeling budget. We find the advantage of these sampling techniques diminish as more labels become available, which is where other search methods can also take over.

Finally, we find that the inclusion of the entropy regularization term (RCE) consistently boosted performance across all experiments. While not the focus of this paper, this result warrants further exploration across other low-label modeling problems.

5 Conclusion

In this paper, we formalized the challenge of label bootstrapping for rare object detection in satellite imagery, establishing benchmark results for both heuristic-based methods, like proximity weighting, and clustering-based approaches. In our case study, the proposed methods significantly improved the positive sample identification rate from 2% to 33%, while enabling machine learning mapping with minimal labeling resources, achieving an F1 score of 0.36 with just 300 labeled patches. We hope this work paves the way for the exploration of bootstrapping strategies for geospatial ML and their connections to adjacent techniques in active learning and subset selection.

REFERENCES

[1] Sebastian Dunnett, Alessandro Sorichetta, Gail Taylor, and Felix Eigenbrod. Harmonised global datasets of wind and solar farm locations and power. *Scientific data*, 7(1):130, 2020.

[2] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pp. 226–231, 1996.

[3] Alex Goupilleau, Tugdual Ceillier, and Marie-Caroline Corbineau. Active learning for object detection in high-resolution satellite images. *arXiv preprint arXiv:2101.02480*, 2021.

[4] Yves Grandvalet and Yoshua Bengio. Entropy regularization.

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

[6] Wei Hu, Tyler Feldman, Eddy Lin, Jose Luis Moscoso, Yanchen J Ou, Natalie Tarn, Baoyan Ye, Wendy Zhang, Jordan Malof, and Kyle Bradbury. Synthetic imagery aided geographic domain adaptation for rare energy infrastructure detection in remotely sensed imagery. In *NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning*, 2021.

[7] Aparna R Joshi, Isha Tarte, Sreeja Suresh, and Shashidhar G Koolagudi. Damage identification and assessment using image processing on post-disaster satellite imagery. In *2017 IEEE Global Humanitarian Technology Conference (GHTC)*, pp. 1–7. IEEE, 2017.

[8] Lily Lee, Benjamin Smith, and T Chen. Fine-grain uncommon object detection from satellite images. In *2015 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pp. 1–6. IEEE, 2015.

[9] Eric Martinson, Bridget Furlong, and Andy Gillies. Training rare object detection in satellite imagery with synthetic gan images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2769–2776, 2021.

[10] Caleb Robinson, Simone Fobi Nsutezo, Anthony Ortiz, Tina Sederholm, Rahul Dodhia, Cameron Birge, Kasie Richards, Kris Pitcher, Paulo Duarte, and Juan M Lavista Ferres. Rapid building damage assessment workflow: An implementation for the 2023 rolling fork, mississippi tornado event. *arXiv preprint arXiv:2306.12589*, 2023.

[11] Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature communications*, 12(1):4392, 2021.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

[13] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[14] Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. 2000.

[15] Peter Tyrrell, Irene Amoke, Koen Betjes, Femke Broekhuis, Robert Buitenwerf, Sarah Carroll, Nathan Hahn, Daniel Haywood, Britt Klaassen, Mette Løvschal, et al. Landscape dynamics (landdx) an open-access spatial-temporal database for the kenya-tanzania borderlands. *Scientific Data*, 9(1):8, 2022.

[16] Zijing Wu, Ce Zhang, Xiaowei Gu, Isla Duporge, Lacey F Hughey, Jared A Stabach, Andrew K Skidmore, J Grant C Hopcraft, Stephen J Lee, Peter M Atkinson, et al. Deep learning enables satellite-based monitoring of large populations of terrestrial mammals across heterogeneous landscape. *Nature communications*, 14(1):3072, 2023.

[17] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

[18] Fahong Zhang, Yilei Shi, Zhitong Xiong, and Xiao Xiang Zhu. Few-shot object detection in remote sensing: Lifting the curse of incompletely annotated novel objects. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[19] Boya Zhao, Yuanfeng Wu, Xinran Guan, Lianru Gao, and Bing Zhang. An improved aggregated-mosaic method for the sparse object detection of remote sensing imagery. *Remote Sensing*, 13(13):2602, 2021.

SUPPLEMENTAL – BOOTSTRAPPING RARE OBJECT DETECTION IN HIGH-RESOLUTION SATELLITE IMAGERY

Akram Zaytar[*] **Caleb Robinson** **Gilles Quentin Hacheme** **Girmaw Abebe Tadesse**

Rahul Dodhia **Juan M. Lavista Ferres** **Lacey F. Hughey** **Jared A. Stabach**

Irene Amoke

1 UNSUPERVISED METHOD FOR CHOOSING HYPERPARAMETERS

Our proposed methods for initializing a sampling surface involve 1.) computing some feature representation per image patch (we test three methods); and 2.) clustering these representations in some manner (a method that will require some hyperparameter choice, e.g. K in KMeans). As we assume that we have no access to labeled data, or more generally, any prior information about the distribution of the object of interest, there is a large question of how to set the parameters of our method. Here we propose an *unsupervised* approach based on the silhouette score [1].

Rousseeuw [1] define a *silhouette coefficient* given a clustering of data and a data point x_i that is the ratio $s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$. Here a_i is the mean intra-cluster distance of x_i, i.e. the average distance to data points in its own cluster, and b_i is the mean nearest cluster distance, i.e. the average distance to all data points in the nearest cluster. The *silhouette score* is the average silhouette coefficient over all data points in a dataset and can range from -1 (indicating that all samples are in wrong clusters, i.e. closer to other clusters than to the cluster they are assigned) to 1 (indicating that samples are well clustered).

We find that the silhouette score from a choice of feature representation and clustering method hyperparameters is highly correlated with the number of samples required to find 100 positive samples. For example in Figure 1 we show the number of samples required to find 100 positive samples with the `Offline KMeans` for different feature representations and choices for k. Overall, the expected number of samples is minimized when the silhouette score is also minimized, and there is high correlation between the two for each feature representation (e.g. $R^2 = 0.93$ when using MOSAIKS based representations). Given this, we use a Bayesian hyperparameter search to minimize the absolute value of the silhouette score given all free parameters, and use the resulting values in each experiment.

2 LABELING SIMULATIONS FOR SAMPLING FUNCTIONS

After finding suitable hyperparameters for each clustering method, we simulate 3 annotation runs per method. Figure 2 plots a log-scale running budget versus the number of found positives for various selection algorithms.

The results show that adaptive clustering methods, which initialize weights based on cluster sizes and change the sampling surface when positive samples are found, perform better. We also note that some methods are perform better at low-regimes (i.e., bisecting KMeans), while others (i.e., proximity sampling) outperform later due to the initial use of uniform weights.

[*] Corresponding author: akramzaytar@microsoft.com

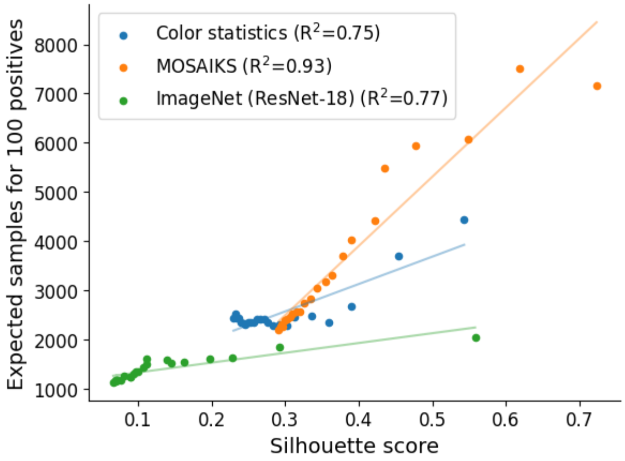

Figure 1: Silhouette scores of different clustering methods versus the expected number of samples required to find 100 positives with an `offline` sampling scheme.

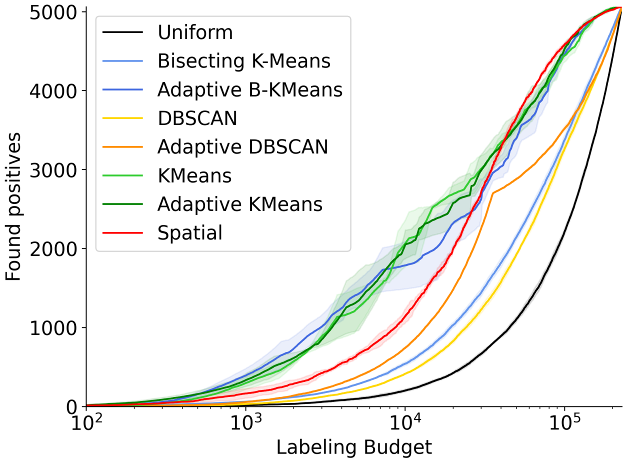

Figure 2: A log-scale comparison of labeling efficiency across sampling algorithms, highlighting the differential performance in early and extended budget scenarios.

REFERENCES

[1] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.