On-the-Fly OVD Adaptation with FLAME: Few-shot Localization via Active Marginal-Samples Exploration

Yehonathan Refael, Amit Aides, Aviad Barzilai,
George Leifman, Vered Silverman, Bolous Jaber, Tomer Shekel, Genady Beryozkin
Google Research

Abstract

Open-vocabulary object detection (OVD) models offer remarkable flexibility by detecting objects from arbitrary text queries. However, their zero-shot performance in specialized domains like Remote Sensing (RS) is often compromised by the inherent ambiguity of natural language, limiting critical downstream applications. For instance, an OVD model may struggle to distinguish between fine-grained classes such as "fishing boat" and "yacht" since their embeddings are similar and often inseparable. This can hamper specific user goals, such as monitoring illegal fishing, by producing irrelevant detections. To address this, we propose a cascaded approach that couples the broad generalization of a large pre-trained OVD model with a lightweight few-shot classifier. Our method first employs the zero-shot model to generate high-recall object proposals. These proposals are then refined for high precision by a compact classifier trained in real-time on only a handful of user-annotated examples—drastically reducing the high costs of RS imagery annotation. The core of our framework is FLAME, a one-step active learning strategy that selects the most informative samples for training. FLAME identifies, on the fly, uncertain marginal candidates near the decision boundary using density estimation, followed by clustering to ensure sample diversity. This efficient sampling technique achieves high accuracy without costly full-model fine-tuning and enables instant adaptation, within less than a minute, which is significantly faster than state-of-the-art alternatives. Our method consistently surpasses state-of-the-art performance on RS benchmarks, establishing a practical and resource-efficient framework for adapting foundation models to specific user needs.

1 Introduction

The recent advancements in large-scale vision-language models (VLMs) such as CLIP [20] have catalyzed a paradigm shift in computer vision, giving rise to Open-Vocabulary Object Detection (OVD) [28]. Unlike traditional detectors limited to predefined categories, OVD models can identify objects described by arbitrary natural language text, offering unprecedented flexibility. This is particularly transformative for remote sensing (RS), where cataloging every possible class is intractable. Early OVD methods adapted standard detectors by replacing the classifier head with text embeddings [10], leveraging the semantic richness of VLMs to generalize to unseen categories. However, the inherent ambiguity of text queries often leads to significant drops in precision, limiting the utility of pure zero-shot systems.

To overcome the limitations of pure zero-shot systems, one alternative is Few-Shot Object Detection (FSOD) [13], which adapts models to novel categories using only a handful of annotated examples. In RS, FSOD is critical due to the difficulty and cost of acquiring dense labels [1]. While effective, common FSOD strategies like meta-learning or fine-tuning [25] can be computationally intensive. To address this, Parameter-Efficient Fine-Tuning (PEFT) techniques such as LoRA [11] have emerged to alleviate these costs by reducing the number of trainable parameters.

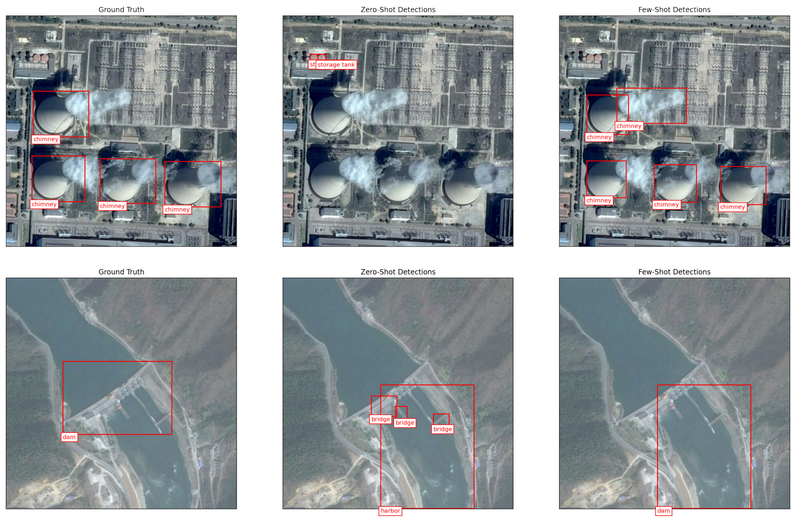

Figure 1: A visual demonstration of performance improvement from Zero-Shot to Few-Shot detection using DIOR dataset [29]. The Zero-Shot model (center) produces noisy and unreliable results, identifying the 'chimneys' but with low confidence and accompanied by several false positives. Our Few-Shot method (right) refines this output, successfully eliminating the false positives and accurately detecting all four chimneys shown in the Ground Truth (left). Similarly, the bottom row showcases the detection of a 'dam'. The Zero-Shot model struggles with false positives such as 'bridge' and 'harbor', which are corrected by the more precise Few-Shot approach.

These FSOD and PEFT strategies are primarily designed to create specialized detectors optimized for a new, specific set of target classes; for example, the methods in [2, 12, 14] are tailored for RS. However, these more efficient adaptation methods still involve a computationally demanding fine-tuning step. Even some recent prototype-based methods [2] require tuning for hundreds of epochs, a process that can take hours and necessitates an accelerator like a GPU (a phase our proposed method eliminates, as we demonstrate in Section 3).

A distinct paradigm explores a hybrid approach that merges OVD and FSOD, using few-shot supervision to enhance and expand an open-vocabulary detector's existing knowledge within a single, unified framework [5]. Several strategies explore this hybrid model: prompt-based methods [9, 30] learn continuous prompts from support sets to improve category alignment, while Transformer-based methods like OV-DETR [27] and OWL-ViT [18] show strong generalization.

The success of these hybrid approaches, which use only a handful of examples, hinges on the efficient selection of the most informative ones. This challenge is addressed by Active Learning (AL) [23], which queries an oracle for the most beneficial labels. Common AL strategies include uncertainty-based sampling [15], diversity-based sampling [22], or their combinations [6]. Building on this foundation, our work proposes a cascaded OVD–FSOD framework with a novel AL strategy specifically designed to resolve semantic ambiguity in RS imagery efficiently and effectively.

2 Method

Theory and Motivations. Our framework relies on the observation that a binary classifier, whether an SVM [24] or a positively homogeneous neural network [19], can be determined entirely by its margin (support) examples. Equivalently, if one removes all non-support training points and retrains, the resulting classifier is unchanged. Building on this, our few-shot procedure identifies a small set of near-boundary examples (the "few-shots"), asks the user to label them, and trains a lightweight model on the fly. Despite using only a handful of points, this model matches the classifier that would have been obtained from training on the full dataset, which may be too large or impractical for real-time training. The Lemma below formalizes this fact for the SVM. Similarly, we prove the same for soft margin (SVM) and for neural networks in Appendix B. The proofs of the Lemmas are also relegated to the Appendix B.

Lemma 2.1 (Support–determination for hard–margin SVM)**.** *Let $\{(x_i, y_i)\}_{i=1}^{n}$ be linearly separable with $y_i \in \{\pm 1\}$. Consider the hard–margin SVM*

$$\min_{w,b} \ \tfrac{1}{2}\|w\|^2 \quad \textit{s.t.} \quad y_i\,(w^\top x_i + b) \ \geq \ 1 \quad (i = 1, \dots, n). \tag{P}$$

Let $(w^\star, b^\star)$ be an optimal solution and define the support set $S \ := \ \{i \in [n] : y_i\,(w^{\star\top} x_i + b^\star) = 1\}$. Then,

1. *$(w^\star, b^\star)$ together with multipliers $\{\alpha_i^\star\}_{i \in S}$ forms a Karush-Kuhn-Tucker (KKT) [7] pair for the* reduced *problem that retains only constraints indexed by S:*

$$\min_{w,b} \ \tfrac{1}{2}\|w\|^2 \quad \textit{s.t.} \quad y_i\,(w^\top x_i + b) \ \geq \ 1 \quad (i \in S). \tag{P_S}$$

2. *Conversely, if $(\tilde{w}, \tilde{b})$ and multipliers $\{\mu_i\}_{i \in S}$ satisfy the KKT system of* (P_S)*, then extending the multipliers by $\tilde{\alpha}_i := \mu_i$ for $i \in S$ and $\tilde{\alpha}_i := 0$ for $i \notin S$ yields a KKT pair $(\tilde{w}, \tilde{b}, \tilde{\alpha})$ for the full problem* (P)*.*

Consequently, (P) *and* (P_S) *have the same optimal solutions. In particular, retraining the hard–margin SVM after removing all non–support points $[n] \setminus S$ leaves the classifier $x \mapsto \operatorname{sign}(w^\top x + b)$ unchanged.*

Remark 2.2 (Kernel SVM). The same argument holds verbatim for kernel SVMs by replacing x_i with $\varphi(x_i)$ in a feature space: at optimality $w^\star = \sum_{i \in S} \alpha_i^\star y_i \varphi(x_i)$, so only support vectors ($\alpha_i^\star > 0$) determine the classifier.

Marginal Samples Retrieval. We propose a one-stage active learning strategy that pinpoints the most informative samples for training a lightweight, class-specific binary classifier. This algorithm 1 allows a large-scale, zero-shot OVD model to be adapted to a new target class efficiently, in real-time, and with minimal human supervision. The method is illustrated in Figure 2.

Algorithm 1 FLAME: Few-shot Localization via Active Marginal-Samples Exploration

---

Require: Unlabeled pool of embeddings $X = \{x_i\}_{i=1}^{N} \subset \mathbb{R}^d$, text embedding $t \in \mathbb{R}^d$; number of target shots K; PCA dimension ℓ; Hyperparameters: Gaussian KDE bandwidth h, ratios $0 < r_l < r_u < 1$, imbalance threshold τ.
Ensure: Selected shots $\hat{X} := \{\hat{x}_k\}_{k=1}^{K}$

for $i = 1$ **to** N **do**
  Compute cosine similarities: $c_i \leftarrow \dfrac{x_i^\top t}{\|x_i\|\,\|t\|}$
  Augment examples: $\tilde{x}_i \leftarrow [x_i,\, c_i]$
end for
\# Marginal samples identification
Project $\{\tilde{x}_i\}$ to ℓ dimensions via PCA to get $S = \{s_i\}_{i=1}^{N}$
Fit Gaussian KDE $\hat{f}$ (bandwidth h) on S: $s^\star \leftarrow \arg\max_s \hat{f}(s)$
Find samples density boundaries s_L, s_U s.t. $\hat{f}(s_L) = r_l \hat{f}(s^\star)$, and $\hat{f}(s_U) = r_u \hat{f}(s^\star)$
\# Promote information diversity
Set $\mathcal{I}_{\text{marginal}} \leftarrow \{i \mid s_i \in [s_L, s_U]\}$, $\quad X_{\text{marginal}} \leftarrow \{x_i \mid i \in \mathcal{I}_{\text{marginal}}\}$
Run k-means clustering on X_{marginal} into K clusters $\{C_k\}_{k=1}^{K}$
Find examples closest to each center $\hat{X} \leftarrow \{\hat{x}_k\}_{k=1}^{K}$
\# User few shot labeling
User labels the few-shots $\hat{X}$ to obtain $D_{\text{labeled}} = \{(\hat{x}_k, y_k)\}_{k=1}^{K}$, $y_k \in \{0, 1\}$
\# Imbalance handling (Optional)
Compute imbalance ratio $\rho \leftarrow \dfrac{\max_{c \in \{0,1\}} |\{y_k = c\}|}{\min_{c \in \{0,1\}} |\{y_k = c\}|}$
if $\rho > \tau$ **then**
  $\hat{X} \leftarrow \text{SMOTE}(D_{\text{labeled}})$
end if
return $\hat{X}$

First, we identify uncertain candidates by augmenting image embeddings with their zero-shot similarity to the text query and applying density estimation in a projected (PCA) augmented-embedding-space. Samples at the distribution's margins are retained as they carry the most informative ambiguity. From this pool, we promote diversity by clustering and selecting one representative per cluster, yielding K candidate shots for annotation. The user then labels these few informative samples, forming an initial dataset. To mitigate imbalance, we apply Synthetic Minority Over-sampling Technique (SMOTE) [3] for extremely fast augmentation. This procedure contributes to a balanced, representative, and efficient training process.

Finally, using the (augmented) few-shots returned by Algorithm 1, we train a compact classifier, by default a Radial Basis Function (RBF) SVM [21], which is trained to find a non-linear separating hyperplane. Note that our efficient framework could support many lightweight alternatives such as: Two-Layer Multi-Layer Perceptron (MLP) under binary cross-entropy loss function, or encoder-classifier with Triplet Loss [8].

3 Experiments

To evaluate its performance, our few-shot method is benchmarked against a zero-shot baseline and leading state-of-the-art approaches, as summarized in Table 1. To that end, we leverage the following two RS datasets: (1) DOTA [26] (Dataset for Object Detection in Aerial Images): A large-scale RS dataset with multi-class, multi-oriented objects annotated in high-resolution aerial images for object detection. (2) DIOR [16] (Dataset for Object Detection in Optical RS Images): A diverse large-scale dataset of optical RS images containing numerous object categories across varying conditions and resolutions for robust detection.

Table 1: Comparison of few-shot object detection performance on the DOTA and DIOR datasets, based on 30-shot examples. The metric used is Average Precision (AP). Our proposed method achieves state-of-the-art results while demonstrating a significantly faster adaptation time.

Method	DOTA	DIOR
Zero-shot OWL-ViT-v2 (Baseline)	13.774%	14.982%
Zero-shot RS-OWL-ViT-v2	31.827%	29.387%
Jeune et. al [14]	37.1%	35.6%
SIoU [12]	45.88%	52.85%
Prototype-based FSOD with DINOv2 [2]	41.40%	26.46%
FLAME cascaded on RS-OWL-ViT-v2	**53.96%**	**53.21%**

We first evaluate the zero-shot performance of the baseline OWL-ViT-v2 model [17], which was pre-trained on the vast, generic multilingual WebLI dataset [4]. We then consider the RS-OWL-ViT-v2 model [1], a remote sensing variant of OWL-ViT-v2 fine-tuned on the RS-WebLI dataset [1], which consists of three million aerial and satellite images from the original WebLI dataset and on a collection of $67,000$ aerial images annotated for remote sensing object detection across 34 categories. Its improved zero-shot performance serves as the starting point for FLAME.

In addition to demonstrating the highest performance, our method adapts in approximately 1 minute for each label on a standard CPU. This is a considerable advantage over fine-tuning approaches that typically require GPU machines and several hours to complete. Further experiments can be found in Appendix C. Illustration of the algorithm is presented in Figure 1.

4 Discussion

This work presents a practical and resource-efficient paradigm for adapting large foundational models to specific user needs in real time. By confining the adaptation to a lightweight classifier trained on actively selected samples, our method avoids costly fine-tuning while effectively resolving semantic ambiguities. This on-the-fly approach paves the way for more dynamic, user-in-the-loop applications in RS, OVD, and other specialized domains.

References

[1] Aviad Barzilai, Yotam Gigi, Amr Helmy, Vered Silverman, Yehonathan Refael, Bolous Jaber, Tomer Shekel, George Leifman, and Genady Beryozkin. A recipe for improving remote sensing vlm zero shot generalization, 2025.

[2] Xavier Bou, Gabriele Facciolo, Rafael Grompone Von Gioi, Jean-Michel Morel, and Thibaud Ehret. Exploring robust features for few-shot object detection in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 430–439, 2024.

[3] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[4] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2023.

[5] Bowen Cheng, Bo Jiang, et al. Revisiting few-shot object detection with vision-language models. *arXiv preprint arXiv:2402.12345*, 2024.

[6] Jae-woo Choi, Jisoo Kim, and Chang-su Kim. Active learning for deep object detection via uncertainty and diversity. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1454–1458. IEEE, 2021.

[7] Tiziana Ciano and Massimiliano Ferrara. Karush-kuhn-tucker conditions and lagrangian approach for improving machine learning techniques: A survey and new developments. *Atti della Accademia Peloritana dei Pericolanti-Classe di Scienze Fisiche, Matematiche e Naturali*, 102(1):1, 2024.

[8] Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 459–474, 2018.

[9] Cheng Feng, Yu Zhong, Tao Zhang, and et al. Promptdet: Towards open-vocabulary detection using uncurated images. In *European Conference on Computer Vision*, pages 701–717. Springer, 2022.

[10] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[12] Pierre Le Jeune and Anissa Mokraoui. Rethinking intersection over union for small object detection in few-shot regime, 2023.

[13] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9539–9548, 2019.

[14] Pierre Le Jeune and Anissa Mokraoui. Improving few-shot object detection through a performance analysis on aerial and natural images. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 513–517. IEEE, 2022.

[15] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12, 1994.

[16] Ke Li, Guorui Wan, Gong Cheng, Lei Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020.

[17] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection, 2024.

[18] Matthias Minderer, Alexey Gritsenko, Austin Stone, and et al. Simple open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6373–6382, 2023.

[19] Andrey Polyakov. Homogeneous artificial neural network, 2023.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sameer Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[21] Bernhard Schölkopf, Christopher JC Burges, and Alexander J Smola. *Advances in kernel methods: support vector learning*. MIT press, 1999.

[22] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

[23] Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin-Madison, Department of Computer Sciences, 2009.

[24] Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[25] Xinlong Wang, Thomas E Huang, Trevor Darrell, Fisher Yu, and Joseph E Gonzalez. Frustratingly simple few-shot object detection. In *International conference on machine learning*, pages 9937–9946. PMLR, 2020.

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

[27] Yiyi Zang, Limin Li, Zechun Wang, Xiang Li, and Jian Sun. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, pages 106–122. Springer, 2022.

[28] Alireza Zareian, Mehrdad Zolfaghari, and Thomas Brox. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14575–14584, 2021.

[29] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

[30] Ming Zhang, Hao Fang, Wen Wang, and et al. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, pages 493–510. Springer, 2022.

A Illustrations

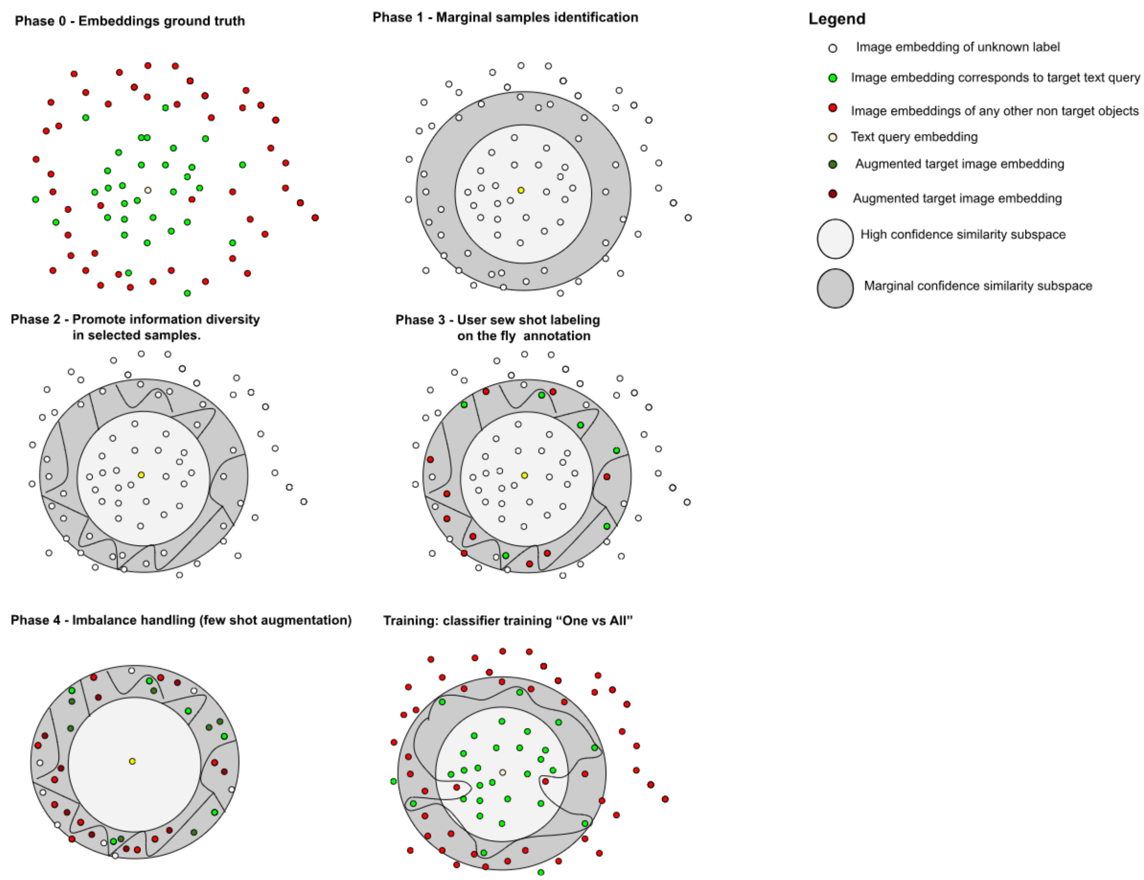

Figure 2: Overview of the proposed few-shot sampling method. The method is followed by the stages: (1) uncertainty-based filtering using density estimation to identify ambiguous candidates near the decision boundary, (2) clustering-based diversity sampling to ensure representative coverage, (3) interactive user annotation of the selected samples, (4) conditional data augmentation with SMOTE or SVM-SMOTE to balance classes, and (5) lightweight classifier training (e.g., SVM or MLP) on the augmented set. This cascaded process refines the zero-shot proposals from a large open-vocabulary detector into an accurate, real-time few-shot classifier without full-model fine-tuning.

B Proofs

Proof of Lemma 2.1. Introduce Lagrange multipliers $\alpha_i \geq 0$ for the constraints in (P). The Lagrangian is

$$\mathcal{L}(w, b, \alpha) = \tfrac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \big(y_i(w^\top x_i + b) - 1\big).$$

The KKT conditions for (P) are:

$$\begin{aligned}
&\text{(stationarity)} && \nabla_w \mathcal{L} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0, && (1)\\
& && \nabla_b \mathcal{L} = -\sum_{i=1}^{n} \alpha_i y_i = 0, && (2)\\
&\text{(primal feasibility)} && y_i(w^\top x_i + b) \geq 1 \quad (\forall i), && (3)\\
&\text{(dual feasibility)} && \alpha_i \geq 0 \quad (\forall i), && (4)\\
&\text{(complementary slackness)} && \alpha_i\big(y_i(w^\top x_i + b) - 1\big) = 0 \quad (\forall i). && (5)
\end{aligned}$$

Let $(w^\star, b^\star, \alpha^\star)$ be a KKT triple for (P). By complementary slackness (5), if $i \notin S$ then $y_i(w^{\star\top} x_i + b^\star) > 1$ and hence $\alpha_i^\star = 0$. Therefore (1)–(2) reduce to

$$w^\star = \sum_{i \in S} \alpha_i^\star y_i x_i, \qquad \sum_{i \in S} \alpha_i^\star y_i = 0,$$

and feasibility plus (5) hold with indices restricted to S. These are precisely the KKT conditions for the reduced problem (P_S). This proves item (1).

Suppose $(\tilde{w}, \tilde{b}, \{\mu_i\}_{i \in S})$ satisfies the KKT system for (P_S). Define $\tilde{\alpha}_i := \mu_i$ for $i \in S$ and $\tilde{\alpha}_i := 0$ for $i \notin S$. Dual feasibility is immediate. Complementary slackness holds for $i \in S$ by assumption and is trivial for $i \notin S$ because $\tilde{\alpha}_i = 0$. The stationarity equations (1)–(2) for (P) coincide with those of (P_S) since the extra terms vanish. Finally, primal feasibility on S holds by construction; constraints for $i \notin S$ are inactive at the optimum of (P) and their multipliers are zero in the constructed KKT triple. Thus $(\tilde{w}, \tilde{b}, \tilde{\alpha})$ satisfies all KKT conditions of the full problem (P).

Combining (1) and (2), the optimal solution set of (P) coincides with that of (P_S). Since the primal objective $\frac{1}{2}\|w\|^2$ is strictly convex in w, the optimal $w^\star$ is unique; hence the decision function $x \mapsto \operatorname{sign}(w^{\star\top} x + b^\star)$ is unchanged when the non–support training points are removed. □

Lemma B.1 (Support examples–determination for homogeneous networks)**.** *Let $\Phi(\theta; \cdot)$ be binary classifier L-homogeneous*[1] *in the weights parameters θ (e.g., ReLU, Leaky ReLU, sigmoid etc), and let the binary training set $\{(x_i, y_i)\}_{i=1}^n$ be linearly separable by $\Phi(\theta; \cdot)$. Consider gradient flow on logistic loss and assume it converges in* direction *to a KKT point $(\theta^\star, \lambda^\star)$ of the maximum–margin program*

$$\min_{\theta} \ \tfrac{1}{2}\|\theta\|^2 \quad \textit{s.t.} \quad y_i\, \Phi(\theta; x_i) \ \geq 1 \quad (i = 1, \ldots, n). \tag{6}$$

Let the (margin/support) set be $S \ := \ \{\, i \in [n] : y_i\, \Phi(\theta^\star; x_i) = 1 \,\}$. Then,

1. *$(\theta^\star, \{\lambda_i^\star\}_{i \in S})$ satisfies the KKT system of the* reduced *problem that keeps only constraints with indices in S:*
$$\min_{\theta} \ \tfrac{1}{2}\|\theta\|^2 \quad \textit{s.t.} \quad y_i\, \Phi(\theta; x_i) \ \geq 1 \quad (i \in S). \tag{7}$$

2. *Conversely, if $(\tilde{\theta}, \{\mu_i\}_{i \in S})$ is a KKT pair for* (7) *and we define $\tilde{\lambda}_i := \mu_i$ for $i \in S$ and $\tilde{\lambda}_i := 0$ for $i \notin S$, then $(\tilde{\theta}, \tilde{\lambda})$ is a KKT pair for the full problem* (6).

Consequently, the sets of KKT solutions of (6) *and* (7) *coincide. In particular, retraining after removing all non-support points $[n] \setminus S$ produces the same limiting classifier $x \mapsto \operatorname{sign}(\Phi(\theta; x))$.*

Proof of Lemma B.1. The KKT conditions for (6) at (θ, λ) are:

$$\begin{aligned}
\text{stationarity:} \quad & \nabla_\theta \left(\tfrac{1}{2}\|\theta\|^2 - \sum_{i=1}^{n} \lambda_i\, y_i\, \Phi(\theta; x_i) \right) = 0, && (8)\\
\text{primal feasibility:} \quad & y_i\, \Phi(\theta; x_i) \geq 1 \quad (\forall i), && (9)\\
\text{dual feasibility:} \quad & \lambda_i \geq 0 \quad (\forall i), && (10)\\
\text{complementary slackness:} \quad & \lambda_i \big(y_i\, \Phi(\theta; x_i) - 1 \big) = 0 \quad (\forall i). && (11)
\end{aligned}$$

[1] A network $\Phi(\theta; x)$ is called *homogeneous* of degree $c > 0$ if for all $b > 0$ and all θ, x, it holds that $\Phi(b, \theta; \mathbf{x}) = b^c\, \Phi(\theta; x)$.

At $(\theta^\star, \lambda^\star)$, complementary slackness gives $\lambda_i^\star = 0$ for every $i \notin S$. Hence (8) reduces to

$$\nabla_\theta \left(\tfrac{1}{2}\|\theta\|^2 - \sum_{i \in S} \lambda_i^\star \, y_i \, \Phi(\theta; x_i) \right) \Big|_{\theta=\theta^\star} = 0,$$

and together with feasibility and complementary slackness restricted to $i \in S$ this is exactly the KKT system for the reduced problem (7), proving (1).

Take any KKT pair $(\tilde{\theta}, \{\mu_i\}_{i \in S})$ for (7) and extend the multipliers by $\tilde{\lambda}_i := \mu_i$ for $i \in S$ and $\tilde{\lambda}_i := 0$ for $i \notin S$. Dual feasibility is immediate and complementary slackness holds because it holds on S and $\tilde{\lambda}_i = 0$ on S^c. The stationarity condition for (6) with $(\tilde{\theta}, \tilde{\lambda})$ coincides with the stationarity condition for (7) with $(\tilde{\theta}, \mu)$, since the sum over $i \notin S$ vanishes. Primal feasibility on S is inherited from (7); at optimality the constraints in S^c are inactive in (6) and do not affect the KKT system. Therefore $(\tilde{\theta}, \tilde{\lambda})$ satisfies the KKT conditions of (6). □

Lemma B.2 (Support-determination for soft-margin SVM)**.** *Let $\{(x_i, y_i)\}_{i=1}^n$ be possibly non-separable with $y_i \in \{\pm 1\}$. Consider a penalty parameter $C > 0$, then the soft-margin SVM is formulated by*

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C \sum_{i=1}^n \xi_i \quad \textit{s.t.} \quad y_i(w^\top x_i + b) \geq 1 - \xi_i, \quad (i = 1, \ldots, n) \qquad (P)$$

Let (w^*, b^*, ξ^*) be an optimal solution to the soft-margin problem (P) with corresponding dual multipliers $\{\alpha_i^*\}_{i=1}^n$ and $\{\beta_i^*\}_{i=1}^n$. Define the support set S as the set of indices with non-zero multipliers α_i^*, $S := \{i \in [n] \mid \alpha_i^* > 0\}$. Then,

1. *$(w^*, b^*, \{\xi_i^*\}_{i \in S})$ together with multipliers $\{\alpha_i^*, \beta_i^*\}_{i \in S}$ forms a Karush-Kuhn-Tucker (KKT) [7] pair for the reduced problem that retains only constraints indexed by S:*

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C \sum_{i=1}^n \xi_i \quad \textit{s.t.} \quad y_i(w^\top x_i + b) \geq 1 - \xi_i, \quad (i \in S) \qquad (P_S)$$

2. *Conversely, if $(\tilde{w}, \tilde{b}, \{\tilde{\xi}_i\}_{i \in S})$ and multipliers $\{\tilde{\alpha}_i, \tilde{\beta}_i\}_{i \in S}$ satisfy the KKT system for (P_S), then extending the solution by setting $\tilde{\alpha}_i = 0$, $\tilde{\xi}_i = 0$, and $\tilde{\beta}_i = C$ for all $i \notin S$ yields a full KKT pair $(\tilde{w}, \tilde{b}, \tilde{\xi}, \tilde{\alpha}, \tilde{\beta})$ for the full problem (P).*

Consequently, (P) and (P_S) have the same optimal solutions (w, b). Retraining the soft-margin SVM after removing all non-support points ($i \notin S$) leaves the classifier $x \mapsto \operatorname{sign}(w^\top x + b)$ unchanged.

Proof. **KKT system for** (P)**.** The Lagrangian is

$$\mathcal{L}(w, b, \xi; \alpha, \beta) = \tfrac{1}{2}\|w\|^2 + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i \big(y_i(w^\top x_i + b) - 1 + \xi_i \big) - \sum_{i=1}^n \beta_i \xi_i,$$

with $\alpha_i \geq 0$ (margin) and $\beta_i \geq 0$ (nonnegativity). The KKT conditions are

$$\begin{aligned}
&\nabla_w \mathcal{L} = 0 \;\Rightarrow\; w = \sum_{i=1}^n \alpha_i y_i x_i, \qquad \partial_b \mathcal{L} = 0 \;\Rightarrow\; \sum_{i=1}^n \alpha_i y_i = 0, \\
&\partial_{\xi_i} \mathcal{L} = 0 \;\Rightarrow\; \alpha_i + \beta_i = C \quad (\forall i), \\
&y_i(w^\top x_i + b) \geq 1 - \xi_i, \;\; \xi_i \geq 0, \quad \alpha_i \geq 0, \; \beta_i \geq 0, \\
&\alpha_i \big(1 - \xi_i - y_i(w^\top x_i + b) \big) = 0, \quad \beta_i \xi_i = 0 \quad (\forall i).
\end{aligned}$$

Let $(w^*, b^*, \xi^*; \alpha^*, \beta^*)$ satisfy these conditions, and set $S := \{i : \alpha_i^* > 0\}$.

(1) Restriction to S. Consider (P_S) with Lagrangian

$$\mathcal{L}_S(w, b, \xi_S; \alpha_S, \beta_S) = \tfrac{1}{2}\|w\|^2 + C\sum_{i\in S}\xi_i - \sum_{i\in S}\alpha_i\big(y_i(w^\top x_i + b) - 1 + \xi_i\big) - \sum_{i\in S}\beta_i\xi_i.$$

Since $\alpha_i^* = 0$ for $i \notin S$, the stationarity of (P) at the starred point reduces to

$$w^* = \sum_{i\in S}\alpha_i^* y_i x_i, \qquad \sum_{i\in S}\alpha_i^* y_i = 0, \qquad \alpha_i^* + \beta_i^* = C \ (i \in S),$$

which are exactly the stationarity equations of (P_S). Primal/dual feasibility and complementary slackness for $i \in S$ are inherited verbatim. Hence $(w^*, b^*, \{\xi_i^*\}_{i\in S}; \{\alpha_i^*, \beta_i^*\}_{i\in S})$ is a KKT pair for (P_S).

Moreover, for $i \notin S$, $\alpha_i^* = 0$ and $\alpha_i^* + \beta_i^* = C$ imply $\beta_i^* = C$. Then $\beta_i^*\xi_i^* = 0$ forces $\xi_i^* = 0$, and primal feasibility gives $y_i((w^*)^\top x_i + b^*) \geq 1$ for all $i \notin S$; thus the dropped constraints are strictly satisfied at (w^*, b^*).

(2) Extension from S to all indices. Let $(\tilde{w}, \tilde{b}, \{\tilde{\xi}_i\}_{i\in S}; \{\tilde{\alpha}_i, \tilde{\beta}_i\}_{i\in S})$ satisfy the KKT system of (P_S), and define for $i \notin S$:

$$\tilde{\alpha}_i := 0, \qquad \tilde{\beta}_i := C, \qquad \tilde{\xi}_i := 0.$$

Then

$$\tilde{w} = \sum_{i\in S}\tilde{\alpha}_i y_i x_i = \sum_{i=1}^{n}\tilde{\alpha}_i y_i x_i, \qquad \sum_{i=1}^{n}\tilde{\alpha}_i y_i = \sum_{i\in S}\tilde{\alpha}_i y_i = 0, \qquad \tilde{\alpha}_i + \tilde{\beta}_i = C \ \ (\forall i),$$

so full stationarity holds; dual feasibility is clear; complementary slackness holds on S by assumption and off S because $\tilde{\alpha}_i = 0$ and $\tilde{\beta}_i\tilde{\xi}_i = C \cdot 0 = 0$. If furthermore (as at any full optimum) the removed constraints satisfy $y_i(\tilde{w}^\top x_i + \tilde{b}) \geq 1$ for $i \notin S$, then primal feasibility holds for (P) and the extended tuple is a full KKT pair.

Equality of optima and unchanged classifier. Let v_P and v_S be the optimal values of (P) and (P_S). By (1), $(w^*, b^*, \{\xi_i^*\}_{i\in S})$ is feasible for (P_S) with value

$$\tfrac{1}{2}\|w^*\|^2 + C\sum_{i\in S}\xi_i^* \;=\; \tfrac{1}{2}\|w^*\|^2 + C\sum_{i=1}^{n}\xi_i^* \;=\; v_P,$$

so $v_S \leq v_P$. Conversely, any feasible $(w, b, \{\xi_i\}_{i\in S})$ of (P_S) can be augmented to a feasible point of (P) by setting, for $i \notin S$, $\xi_i^\uparrow := \max\{0,\, 1 - y_i(w^\top x_i + b)\}$, which yields objective $\geq \tfrac{1}{2}\|w\|^2 + C\sum_{i\in S}\xi_i$; hence $v_P \leq v_S$. Therefore $v_P = v_S$.

Because the primal objective is strictly convex in w, both problems share the same optimal w (unique), and one may choose a consistent optimal bias b (e.g., via any support with $0 < \alpha_i < C$). Hence removing non-support points and retraining on S leaves $x \mapsto \mathrm{sign}(w^\top x + b)$ unchanged. □

C Additional experiments

Following, Table 2 provides a detailed per-class breakdown of the Average Precision (AP) on both the DIOR and DOTA datasets, comparing our few-shot method against the zero-shot baseline using the Zero-shot OWL-ViT-v2 fine-tuned on RS-WebLI (appear in second line of Table 1). The missing values in the 'Few-Shot' columns indicate instances where the initial zero-shot retrieval step failed to find any relevant image embeddings, thereby preventing the few-shot selection process. The results highlight the substantial performance gains achieved through our approach across a wide range of object categories.

Table 2: Detailed per-class Average Precision (AP) comparison of our few-shot method against a zero-shot baseline (OWL-ViT-v2 fine-tuned on RS-WebLI) on the DIOR (left) and DOTA (right) datasets. The '–' symbol denotes a failure case for our method, occurring when the initial zero-shot step retrieved no relevant candidate images for a given class, thereby preventing the few-shot selection process. The results highlight the substantial AP gains achieved by our approach across a diverse range of object categories. Hyper-parameters were chosen via standard validation grid search.

DIOR Dataset

Class	**Few Shot**	**Zero Shot**
expressway service area	0.82	0.03
expressway toll station	0.99	0
airplane	0.99	0.84
airport	–	0
baseball field	0.93	0.62
basketball court	0.87	0.66
bridge	0.49	0.21
chimney	0.94	0.11
dam	0.71	0.04
golf field	0.72	0.01
ground track field	0.79	0.5
harbor	0.64	0.33
overpass	0.75	0.1
ship	0.93	0.72
stadium	0.86	0.57
storage tank	0.68	0.73
tennis court	0.57	0.8
train station	–	0.01
vehicle	0.79	0.25
windmill	1	0.67

DOTA Dataset

Class	**Few Shot**	**Zero Shot**
Baseball Diamond	0.88	0.32
Basketball Court	0.83	0.56
Bridge	0.28	0.09
Container Crane	0.95	0.03
Ground Track Field	0.68	0.4
Harbor	0.82	0.36
Helicopter	0.73	0.39
Large Vehicle	0.87	0.32
Plane	0.54	0.78
Roundabout	0.91	0.24
Ship	0.82	0.71
Small Vehicle	0.77	0.28
Soccer Ball Field	0.77	0.48
Storage Tank	0.55	0.79
Swimming Pool	0.58	0.71
Tennis Court	–	0.77