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.

Introduction

The recent advancements in large-scale vision-language models (VLMs) such as CLIP (Radford et al. 2021) have catalyzed a paradigm shift in computer vision, giving rise to Open-Vocabulary Object Detection (OVD) (Zareian, Zolfaghari, and Brox 2021). 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 (Gu et al. 2021), 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) (Kang et al. 2019), 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(Barzilai et al. 2025). While effective, common FSOD strategies like meta-learning or fine-tuning (Wang et al. 2020) can be computationally intensive. To address this, Parameter-Efficient Fine-Tuning (PEFT) techniques such as LoRA (Hu et al. 2021) have emerged to alleviate these costs by reducing the number of trainable parameters.

These FSOD and PEFT strategies are primarily designed to create specialized detectors optimized for a new, specific set of target classes, for example (Bou et al. 2024; Jeune and Mokraoui 2023; Le Jeune and Mokraoui 2022) are tailored for RS. However, these more efficient adaptation methods still involve a computationally demanding fine-tuning step. Even some recent prototype-based methods (Bou et al. 2024) 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 later in this study).

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 (Cheng, Jiang et al. 2024). Several strategies explore this hybrid model: prompt-based methods (Feng et al. 2022; Zhang et al. 2022) learn continuous prompts from support sets to improve category alignment, while Transformer-based methods like OV-DETR (Zang et al. 2022) and OWL-ViT (Minderer et al. 2023) 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) (Settles 2009), which queries an oracle for the most beneficial labels. Common AL strategies include uncertainty-based sampling (Lewis and Gale 1994), diversity-based sampling (Sener and Savarese 2018), or their combina-

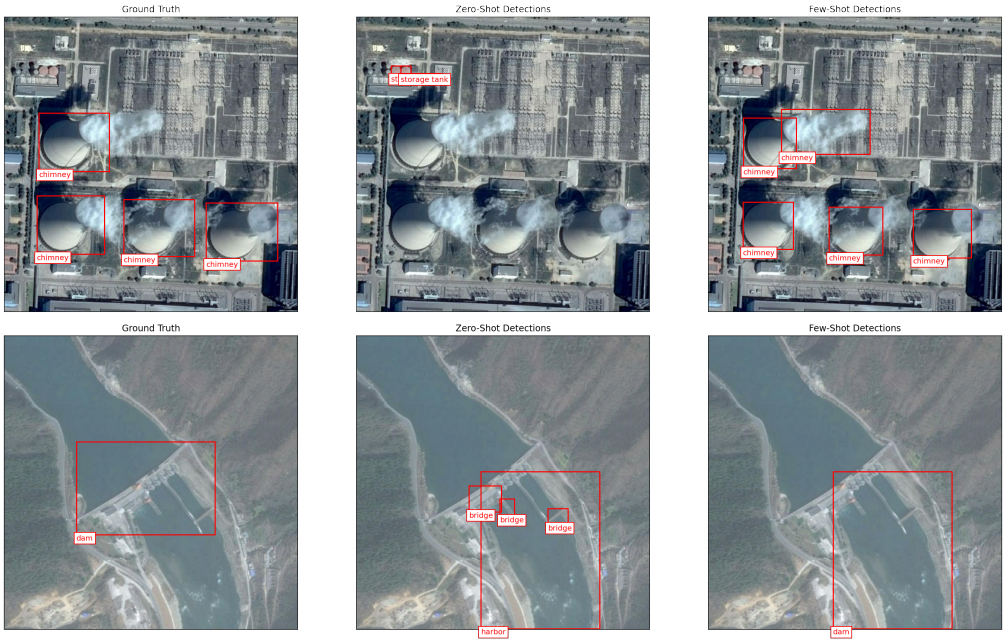

Figure 1: A visual demonstration of performance improvement from Zero-Shot to Few-Shot detection using DIOR dataset (Zhan, Xiong, and Yuan 2023). The Zero-Shot model (center) produces noisy and unreliable results, identifying the 'chimneys' but with low confidence and accompanied by several false positives. 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).

tions (Choi, Kim, and Kim 2021). 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.

Method

Theory and Motivations. Our framework lays on the observation that a binary classifier, whether an SVM (Vapnik 1995) or a positively homogeneous neural network (Polyakov 2023), 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 Lemmas below formalizes this fact for the SVM, soft margin SVM and for neural networks.

Lemma 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*

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

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

1. $(w^\star, b^\star)$ *together with multipliers* $\{\alpha_i^\star\}_{i\in S}$ *forms a Karush-kuhn-tucker (KKT) (Ciano and Ferrara 2024) pair for the* reduced *problem that retains only constraints indexed by* S*:*

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

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* $x \mapsto \mathrm{sign}(w^\top x + b)$ *unchanged.*

Remark 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.

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

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

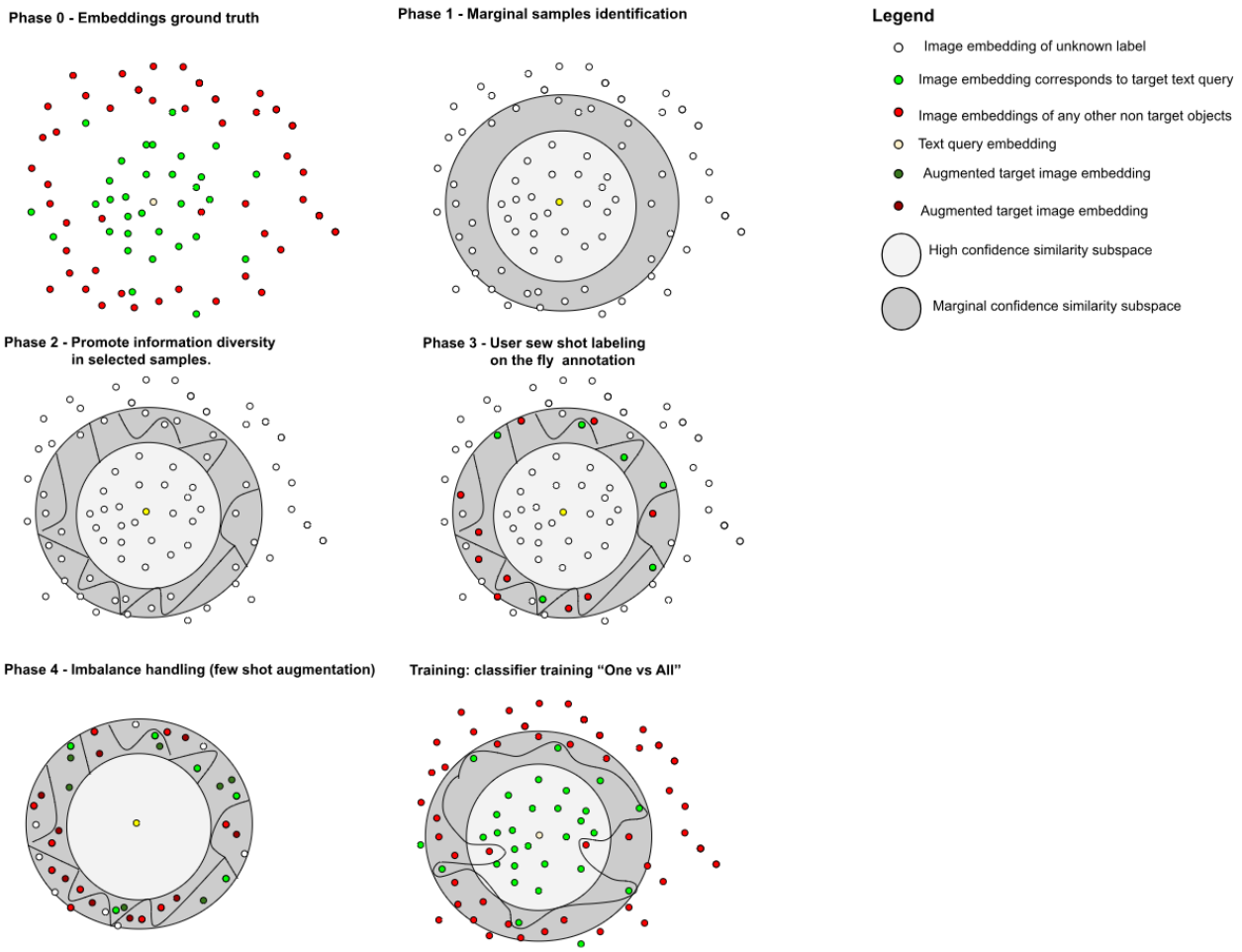

Figure 2: Overview of the proposed few-shot sampling method. The method is followed by the stages: (1) uncertainty-based filtering using density estimation to identify ambiguous candidates near the decision boundary, (2) clustering-based diversity sampling to ensure representative coverage, (3) interactive 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.

and the KKT conditions are

$$\begin{aligned}
&\text{(stationarity)} && w = \sum_{i=1}^{n} \alpha_i y_i x_i, \qquad \sum_{i=1}^{n} \alpha_i y_i = 0,\\
&\text{(primal feas.)} && y_i(w^\top x_i + b) \geq 1 \quad (\forall i),\\
&\text{(dual feas.)} && \alpha_i \geq 0 \quad (\forall i),\\
&\text{(comp. slackness)} && \alpha_i\big(y_i(w^\top x_i + b) - 1\big) = 0 \quad (\forall i).
\end{aligned}$$

(1) Full ⇒ reduced. Let $(w^\star, b^\star, \alpha^\star)$ be any KKT triple for (P), and set $S = \{i : y_i(w^{\star\top} x_i + b^\star) = 1\}$. By complementary slackness, $\alpha_i^\star = 0$ for every $i \notin S$. Hence stationarity reduces 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 together with feasibility and slackness on S these are exactly the KKT conditions of the reduced problem (P_S). Thus $\big(w^\star, b^\star, (\alpha_i^\star)_{i \in S}\big)$ is KKT for (P_S).

(2) Reduced ⇒ full. Conversely, let $(\tilde{w}, \tilde{b}, (\mu_i)_{i \in S})$ satisfy the KKT system for (P_S), and define $\tilde{\alpha}_i := \mu_i$ for $i \in S$ and $\tilde{\alpha}_i := 0$ for $i \notin S$. Then stationarity, dual feasibility, and complementary slackness for (P) hold immediately. To check the remaining primal feasibility on $[n] \setminus S$, compare duals: the dual of (P_S) is the dual of (P) restricted to indices S. Since an optimal dual solution of (P) has $\alpha_i^\star = 0$ for $i \notin S$, the restricted dual attains the same optimal value; by strong duality, (P) and (P_S) share the same optimal objective value. Because the primal objective is strictly convex in w, any optimal reduced solution must satisfy $\tilde{w} = w^\star$, and the equalities on S then fix $\tilde{b} = b^\star$. Hence $y_i(\tilde{w}^\top x_i + \tilde{b}) \geq 1$ for all $i \in [n]$, i.e., primal feasibility for the full problem. Thus $(\tilde{w}, \tilde{b}, \tilde{\alpha})$ is KKT for (P).

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 centers $\hat{X} \leftarrow \{\hat{x}_k\}_{k=1}^K$

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

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 $\hat{X}$

Parts (1)–(2) imply that (P) and (P_S) have the same optimal solutions. In particular, removing non–support points leaves the classifier $x \mapsto \operatorname{sign}(w^\top x + b)$ unchanged. □

Our claim for the non-separable embeddings case, which is the soft marginal SVM, is stated in the following lemma.

Lemma 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*

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

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) (Ciano and Ferrara 2024) pair for the reduced problem that retains only constraints indexed by S:*

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

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. Let $(w^*, b^*, \xi^*; \alpha^*, \beta^*)$ be a KKT pair of (P), where the Lagrangian is $\mathcal{L} = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_i \alpha_i\big(y_i(w^\top x_i + b) - 1 + \xi_i\big) - \sum_i \beta_i \xi_i$ with $\alpha_i, \beta_i \geq 0$ and the implicit constraints $\xi_i \geq 0$. The KKT conditions read: (i) $w = \sum_i \alpha_i y_i x_i$, $\sum_i \alpha_i y_i = 0$, and $\alpha_i + \beta_i = C$; (ii) $y_i(w^\top x_i + b) \geq 1 - \xi_i$, $\xi_i \geq 0$; (iii) $\alpha_i(1 - \xi_i - y_i(w^\top x_i + b)) = 0$, $\beta_i \xi_i = 0$. Define $S := \{i : \alpha_i^* > 0\}$. Since $\alpha_i^* = 0$ for $i \notin S$, the stationarity equations at the starred point reduce to $w^* = \sum_{i\in S} \alpha_i^* y_i x_i$ and $\sum_{i\in S} \alpha_i^* y_i = 0$, while $\alpha_i^* + \beta_i^* = C$ holds for $i \in S$. Together with primal/dual feasibility and complementary slackness restricted to $i \in S$, this shows that $(w^*, b^*, \{\xi_i^*\}_{i\in S}; \{\alpha_i^*, \beta_i^*\}_{i\in S})$ satisfies the KKT system of the reduced problem (P_S). Moreover, for $i \notin S$ we have $\alpha_i^* = 0$ and thus $\beta_i^* = C$, which by $\beta_i^* \xi_i^* = 0$ forces $\xi_i^* = 0$ and hence $y_i((w^*)^\top x_i + b^*) \geq 1$, i.e., the dropped constraints are strictly satisfied at (w^*, b^*). Conversely, take any KKT pair $(\tilde{w}, \tilde{b}, \{\tilde{\xi}_i\}_{i\in S}; \{\tilde{\alpha}_i, \tilde{\beta}_i\}_{i\in S})$ for (P_S) and extend by setting $\tilde{\alpha}_i := 0$, $\tilde{\beta}_i := C$, $\tilde{\xi}_i := 0$ for $i \notin S$. Then

$\tilde{w} = \sum_{i\in S} \tilde{\alpha}_i y_i x_i = \sum_{i=1}^n \tilde{\alpha}_i y_i x_i$ and $\sum_{i=1}^n \tilde{\alpha}_i y_i = 0$, while $\tilde{\alpha}_i + \tilde{\beta}_i = C$ and the complementary slackness equalities hold for all i; if $y_i(\tilde{w}^\top x_i + \tilde{b}) \geq 1$ for $i \notin S$ (as occurs at any optimum of the full problem), the extension is a full KKT pair for (P). Finally, letting v_P and v_S be the optimal values of (P) and (P_S), the restriction above shows $v_S \leq v_P$, while any feasible $(w, b, \{\xi_i\}_{i\in S})$ of (P_S) can be augmented by $\xi_i^\uparrow := \max\{0, 1 - y_i(w^\top x_i + b)\}$ for $i \notin S$ to give a feasible point of (P) with no smaller objective, hence $v_P \leq v_S$. Thus $v_P = v_S$, and since the objective is strictly convex in w, both problems share the same optimal w (and a consistent b), so removing non-support points and retraining leaves the classifier $\mathrm{sign}(w^\top x + b)$ unchanged. □

Lemma 3 (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*

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

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:*

$$\min_\theta \ \tfrac{1}{2}\|\theta\|^2 \quad \textit{s.t.} \quad y_i\, \Phi(\theta; x_i) \ \geq 1 \quad (i \in S). \quad (2)$$

2. *Conversely, if $(\tilde{\theta}, \{\mu_i\}_{i\in S})$ is a KKT pair for* (2) *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* (1)*.*

Consequently, the sets of KKT solutions of (1) *and* (2) *coincide. In particular, retraining after removing all non-support points $[n] \setminus S$ produces the same limiting classifier $x \mapsto \mathrm{sign}(\Phi(\theta; x))$.*

Proof of Lemma 3. Introduce multipliers $\lambda_i \geq 0$ for the constraints in (1). The Lagrangian is

$$\mathcal{L}(\theta, \lambda) \ = \ \tfrac{1}{2}\|\theta\|^2 \ - \ \sum_{i=1}^n \lambda_i\, y_i\, \Phi(\theta; x_i),$$

and the KKT conditions read

$$\begin{aligned}
&\text{(stationarity)} && \theta \ - \ \sum_{i=1}^n \lambda_i\, y_i\, \nabla_\theta \Phi(\theta; x_i) \ = \ 0,\\
&\text{(primal feasibility)} && y_i\, \Phi(\theta; x_i) \ \geq \ 1 \quad (\forall i),\\
&\text{(dual feasibility)} && \lambda_i \ \geq \ 0 \quad (\forall i),\\
&\text{(complementary slackness)} && \lambda_i \big(y_i\, \Phi(\theta; x_i) - 1\big) \ = \ 0 \quad (\forall i).
\end{aligned}$$

(1) Full ⇒ reduced. Let $(\theta^\star, \lambda^\star)$ be a KKT pair for (1) and $S = \{i : y_i\, \Phi(\theta^\star; x_i) = 1\}$. By complementary slackness, $\lambda_i^\star = 0$ for every $i \notin S$, so the stationarity condition reduces to

$$\theta^\star \ - \ \sum_{i\in S} \lambda_i^\star\, y_i\, \nabla_\theta \Phi(\theta^\star; x_i) \ = \ 0.$$

Together with primal/dual feasibility and complementary slackness restricted to $i \in S$, these are precisely the KKT conditions of the reduced problem (2). Hence $(\theta^\star, (\lambda_i^\star)_{i\in S})$ is KKT for (2).

(2) Reduced ⇒ full. Conversely, let $(\tilde{\theta}, (\mu_i)_{i\in S})$ satisfy the KKT system for (2) and define $\tilde{\lambda}_i := \mu_i$ for $i \in S$ and $\tilde{\lambda}_i := 0$ for $i \notin S$. Dual feasibility and complementary slackness for (1) are immediate. The stationarity condition for (1) at $(\tilde{\theta}, \tilde{\lambda})$ is

$$\tilde{\theta} \ - \ \sum_{i\in S} \mu_i\, y_i\, \nabla_\theta \Phi(\tilde{\theta}; x_i) \ = \ 0,$$

which coincides with the reduced stationarity condition. Primal feasibility on S holds by assumption. For $i \notin S$, the constraints are nonbinding at the full KKT point $(\theta^\star, \lambda^\star)$ used to define S; hence, at that scale of the homogeneous model, they are redundant. In particular, any KKT pair of the reduced problem that satisfies the above stationarity (which matches the full one with $\tilde{\lambda}_i = 0$ on S^c) and the inequalities on S also satisfies $y_i\, \Phi(\tilde{\theta}; x_i) \geq 1$ for all $i \notin S$ (the added constraints remain inactive), and therefore $(\tilde{\theta}, \tilde{\lambda})$ is KKT for (1).

Combining (1)–(2), the KKT solution sets of (1) and (2) coincide. Consequently, removing all non–support points leaves the limiting classifier $x \mapsto \mathrm{sign}(\Phi(\theta; x))$ unchanged. □

[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 $\Phi(b, \theta; \mathbf{x}) = b^c\, \Phi(\theta; x)$.

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. 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) (Chawla et al. 2002) for extremely fast augmentation. This procedure would contribute to a balanced, representative, and efficiently training to take place shortly after.

Finally, using the (augmented) few-shots returned by Algorithm 1, we train a compact classifier, by default an Radial Basis kernel (RBF) SVM (Schölkopf, Burges, and Smola 1999), which is trained to find a non-linear separating hyperplane. Note that our efficient framework could support many

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 (Le Jeune and Mokraoui 2022)	37.1%	35.6%
SIoU (Jeune and Mokraoui 2023)	45.88%	52.85%
Prototype-based FSOD with DINOv2 (Bou et al. 2024)	41.40%	26.46%
FLAME cascaded on RS-OWL-ViT-v2	**53.96%**	**53.21%**

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.

DIOR Dataset

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

DOTADataset

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

lightweight alternatives such as: Two-Layer Multi-Layer Perceptron (MLP) under binary cross-entropy loss function, or encoder-classifier with Triplet Loss (Dong and Shen 2018). Illustration schema of the algorithm is presented in Figure 2.

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 (Xia et al. 2018) (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(Li et al. 2020) (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.

We first evaluate the zero-shot performance of the baseline OWL-ViT-v2 model (Minderer, Gritsenko, and Houlsby 2024), which was pre-trained on the vast, generic multilingual WebLI dataset (Chen et al. 2023). We then consider the RS-OWL-ViT-v2 model (Barzilai et al. 2025), a remote

sensing variant of OWL-ViT-v2 fine-tuned on the RS-WebLI dataset (Barzilai et al. 2025), 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. This improved zero-shot performance model serves as the starting point for FLAME.

This Table 1 demonstrates that the FLAME cascaded on RS-OWL-VIT-v2 method achieves the highest Average Precision (AP) on both the DOTA (53.96%) and DIOR (53.21%) datasets among all compared Few-Shot Object Detection (FSOD) models.This superior performance is coupled with a significantly faster adaptation time (approximately 1 minute per label on a CPU) compared to competing fine-tuning approaches that typically require a GPU and several hours.

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 (which 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. Without these initial candidates, the few-shot selection process could not proceed, resulting in a method failure for those specific classes. The Table highlights the substantial performance gains achieved by the Few-Shot (FLAME) method over the Zero-Shot baseline across a wide range of object categories on both the DIOR and DOTA datasets. For instance, the Few-Shot method dramatically improves AP for challenging classes like 'expressway toll station' on DIOR (from 0% to 99%) and 'Container Crane' on DOTA (from 3% to 95%), showcasing its effectiveness in resolving semantic ambiguity

Discussion

Remote sensing is a field that involves the acquisition of information about an object or area without making physical contact with it, typically using sensors on platforms such as satellites or aircraft. The proposed method provides a practical and resource-efficient framework for adapting foundational remote sensing OVD models to specific user needs. The cascaded architecture combines a large, pre-trained OVD model with a lightweight, few-shot classifier. This approach generates initial object-embedding proposals using the frozen weights of the zero-shot model, which are then refined by a compact classifier trained in real-time on a handful of user-annotated examples. This process drastically reduces annotation overhead while achieving high accuracy without the costly process of full-model fine-tuning. The core contribution is an efficient one-step active learning strategy that selects the most informative samples for user annotation. This strategy identifies a small number of uncertain candidates near the decision boundary using density estimation and then applies clustering to ensure a diverse training set. The method is designed to address the semantic ambiguity of text queries that hampers the zero-shot performance of pre-trained models.

Limitations

Our approach is powerful but has clear boundaries. As demonstrated in our analysis of Table 2 , the performance of our cascaded system is fundamentally capped by the recall of the initial zero-shot model. For classes where the base OVD model fails to retrieve any candidates (e.g., 'airport'), FLAME cannot proceed. Our method is a high-precision refiner, not a recall generator.

Future Work

Our one-step active learning strategy could be extended to an iterative, multi-step process, allowing the user to progressively refine the classifier by labeling more marginal samples until a desired performance is met. Finally, the efficacy of FLAME should be tested in other specialized domains that suffer from fine-grained ambiguity, such as medical imagery or manufacturing defect detection.

References

Barzilai, A.; Gigi, Y.; Helmy, A.; Silverman, V.; Refael, Y.; Jaber, B.; Shekel, T.; Leifman, G.; and Beryozkin, G. 2025. A Recipe for Improving Remote Sensing VLM Zero Shot Generalization. arXiv:2503.08722.

Bou, X.; Facciolo, G.; Von Gioi, R. G.; Morel, J.-M.; and Ehret, T. 2024. Exploring robust features for few-shot object detection in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 430–439.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357.

Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A.; Padlewski, P.; Salz, D.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; Kolesnikov, A.; Puigcerver, J.; Ding, N.; Rong, K.; Akbari, H.; Mishra, G.; Xue, L.; Thapliyal, A.; Bradbury, J.; Kuo, W.; Seyedhosseini, M.; Jia, C.; Ayan, B. K.; Riquelme, C.; Steiner, A.; Angelova, A.; Zhai, X.; Houlsby, N.; and Soricut, R. 2023. PaLI: A Jointly-Scaled Multilingual Language-Image Model. arXiv:2209.06794.

Cheng, B.; Jiang, B.; et al. 2024. Revisiting few-shot object detection with vision-language models. *arXiv preprint arXiv:2402.12345*.

Choi, J.-w.; Kim, J.; and Kim, C.-s. 2021. Active learning for deep object detection via uncertainty and diversity. In *2021 IEEE International Conference on Image Processing (ICIP)*, 1454–1458. IEEE.

Ciano, T.; and Ferrara, M. 2024. 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.

Dong, X.; and Shen, J. 2018. Triplet loss in siamese network for object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, 459–474.

Feng, C.; Zhong, Y.; Zhang, T.; and et al. 2022. Promptdet: Towards open-vocabulary detection using uncurated images. In *European Conference on Computer Vision*, 701–717. Springer.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jeune, P. L.; and Mokraoui, A. 2023. Rethinking Intersection Over Union for Small Object Detection in Few-Shot Regime. arXiv:2307.09562.

Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; and Darrell, T. 2019. Few-shot object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9539–9548.

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

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

Li, K.; Wan, G.; Cheng, G.; Meng, L.; and Han, J. 2020. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159: 296–307.

Minderer, M.; Gritsenko, A.; and Houlsby, N. 2024. Scaling Open-Vocabulary Object Detection. arXiv:2306.09683.

Minderer, M.; Gritsenko, A.; Stone, A.; and et al. 2023. Simple open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6373–6382.

Polyakov, A. 2023. Homogeneous Artificial Neural Network. arXiv:2311.17973.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Schölkopf, B.; Burges, C. J.; and Smola, A. J. 1999. *Advances in kernel methods: support vector learning*. MIT press.

Sener, O.; and Savarese, S. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Settles, B. 2009. Active learning literature survey. Technical Report 1648, University of Wisconsin-Madison, Department of Computer Sciences.

Vapnik, V. 1995. Support-vector networks. *Machine learning*, 20: 273–297.

Wang, X.; Huang, T. E.; Darrell, T.; Yu, F.; and Gonzalez, J. E. 2020. Frustratingly simple few-shot object detection. In *International conference on machine learning*, 9937–9946. PMLR.

Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2018. DOTA: A large-scale dataset for object detection in aerial images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3974–3983.

Zang, Y.; Li, L.; Wang, Z.; Li, X.; and Sun, J. 2022. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, 106–122. Springer.

Zareian, A.; Zolfaghari, M.; and Brox, T. 2021. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14575–14584.

Zhan, Y.; Xiong, Z.; and Yuan, Y. 2023. RSVG: Exploring Data and Models for Visual Grounding on Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13.

Zhang, M.; Fang, H.; Wang, W.; and et al. 2022. Tip-adapter: Training-free adaption of CLIP for few-shot classification. In *European Conference on Computer Vision*, 493–510. Springer.