FLAME: On-the-Fly OVD Adaptation through Active Selection of Marginal Samples

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

Introduction & Motivation

Open-Vocabulary Detection (OVD) allows flexible text-based detection but suffers from semantic ambiguity (e.g., the dual meaning of "bat"). This is exacerbated in Remote Sensing, where the severe domain shift from eye-level pre-training to overhead imagery hinders fine-grained distinction.

Key Challenges:

- Zero-shot models struggle to distinguish fine-grained classes (e.g., "fishing boat" vs. "yacht").
- Acquiring dense labels in RS is labor-intensive and costly.
- Full fine-tuning is computationally prohibitive for rapid deployment.

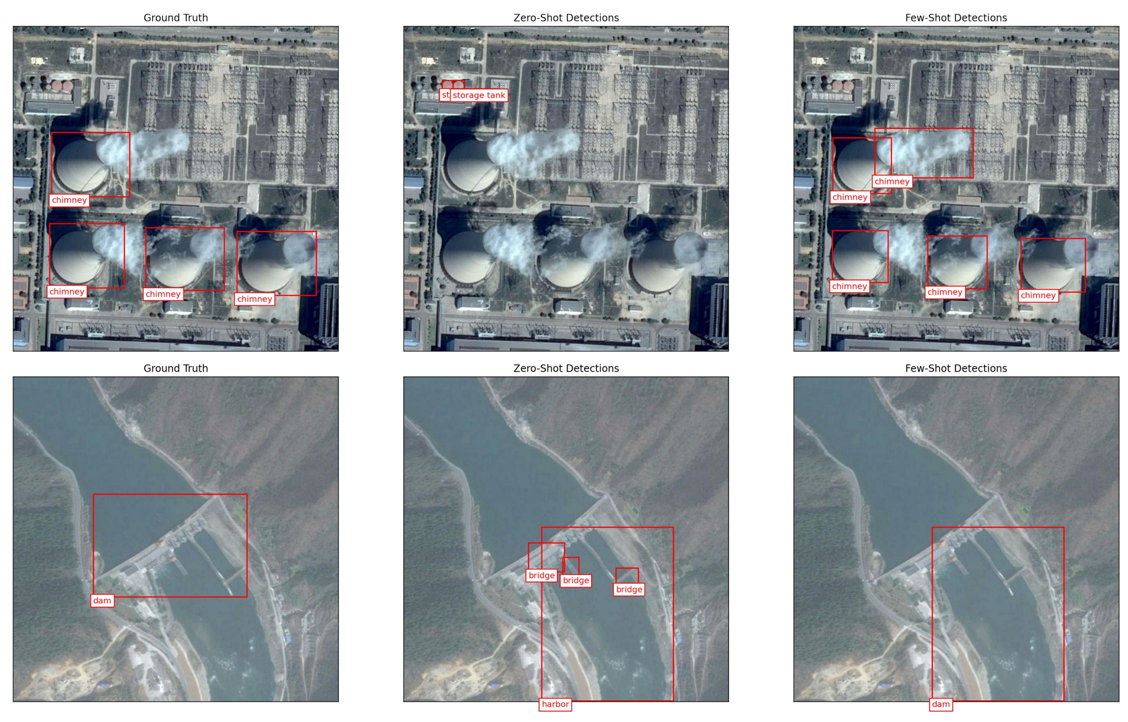

Figure 1. Visual demonstration on the DIOR dataset: 'chimney' (top) and 'dam' (bottom). **Left:** Ground Truth. **Center:** Zero-shot detection (noisy with false positives). **Right:** Our few-shot method refines the output, matching the ground truth.

Theoretical Foundation

Our approach relies on the principle that a binary decision boundary is defined solely by its **marginal (support) examples**.

Lemma (Support-Determination): Retraining a hard-margin SVM (or homogeneous Neural Network) after discarding all non-support training points leaves the decision boundary invariant.

$$\alpha_i^* \left(y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 \right) = 0$$
$$\Downarrow$$
$$\alpha_i^* > 0 \iff \mathbf{x}_i \text{ is a support vector}$$

Consequently, only samples near the decision boundary (uncertainty region) are informative. We leverage this insight to minimize user annotation effort by exclusively querying these marginal samples.

The FLAME Framework

Few-shot **L**ocalization via **A**ctive **M**arginal-Sample **E**xploration.

We propose a three-stage cascaded framework:

1. **Zero-Shot Proposal:** Generate high-recall proposals using a frozen OVD model (OWLViT-v2).
2. **Active Selection:** FLAME identifies the most informative samples.
3. **On-the-Fly Training:** Train a lightweight classifier (SVM/MLP) on the user-labeled support set.

Method Description

Algorithm Steps:

1. **Zero-Shot Retrieval:** Initial filtering via text-to-image similarity.
2. **Density Estimation:** Estimate embedding distribution density using Gaussian KDE to identify high-confidence regions.
3. **Marginal Retrieval:** Retrieve samples at the boundaries (low density relative to the peak) that represent semantic ambiguity.
4. **Diversity Clustering:** Cluster marginal samples and select centroids to ensure annotation diversity.
5. **User Annotation + Few-Shot Classifier:** Expert labels the K selected centroids. Few-Shot Classifier
6. **Cascaded Inference:** Sequential application of Zero-Shot and Few-Shot classifiers.

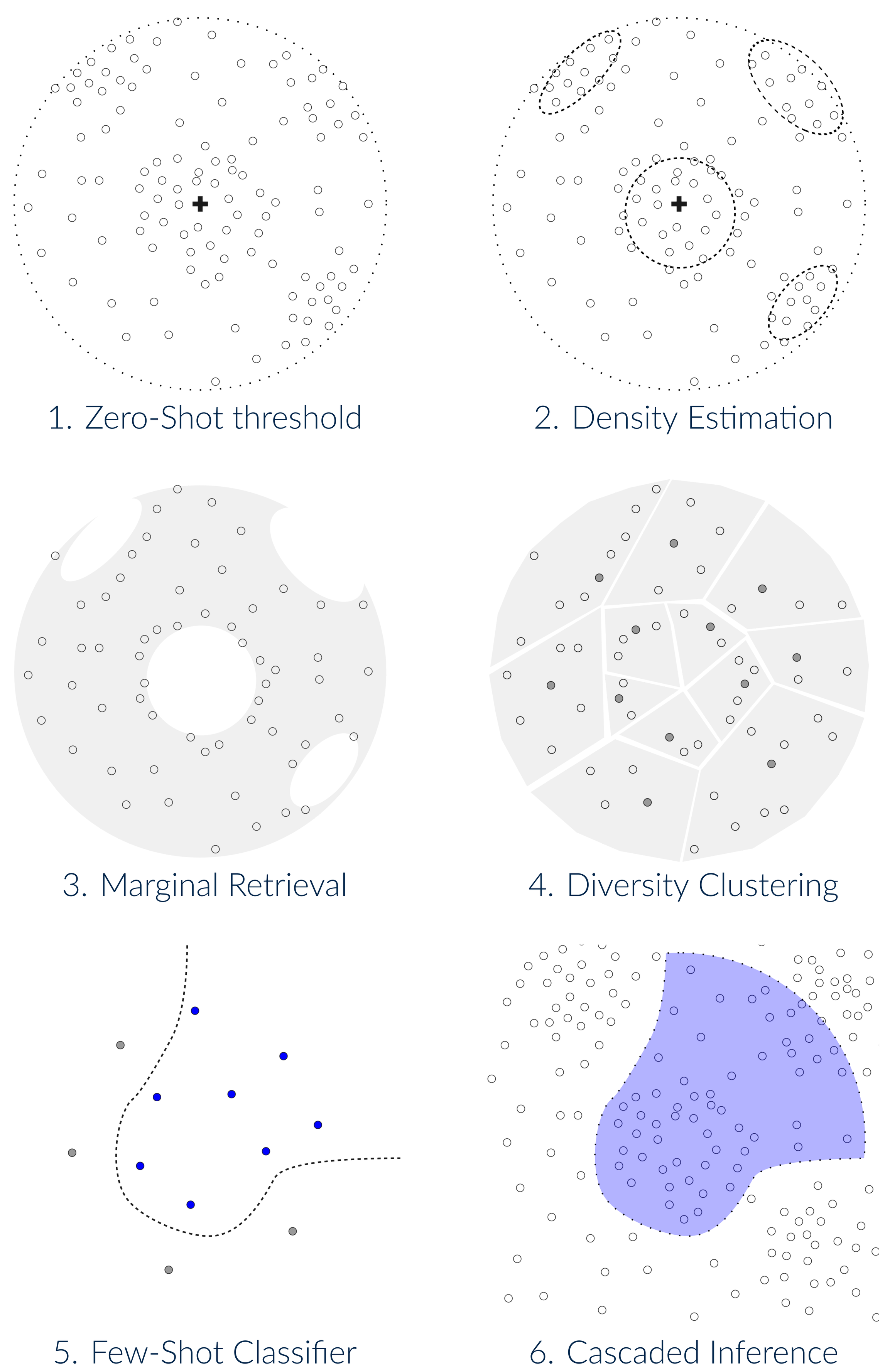

Figure 2. **Method overview**. Image embeddings are visualized as small circles; the text embedding is represented by the bold plus sign.

Algorithm Pseudocode

Input: Embeddings X, Text query t, Zero-shot threshold τ_{ZS}, and Few-shot budget K.
Zero-Shot: Compute cosine similarities $c_i \leftarrow \frac{x_i^\top t}{\|x_i\| \, \|t\|}$
Filter $X_{ZS} \leftarrow \{x_i \ s.t. \ c_i > \tau_{ZS}\}$
Augment $\tilde{x}_i \leftarrow [x_i, c_i]$.
Marginal Retrieval: Project X_{ZS} to low-dim via PCA.
Fit KDE $\hat{f}$.
Identify margins: $s \in [s_L, s_U]$ based on density ratios.
Diversity Clustering: Cluster samples into K groups.
Select centroids X_{FS}.
User Loop: User labels X_{FS}.
Train: Lightweight Classifier (SVM/MLP) on X_{FS}.

Experimental Results

We evaluated on **DOTA** [1] and **DIOR** [2] datasets using a 30-shot protocol.

Method	DOTA AP	DIOR AP
Zero-shot OWLViT-v2 [3]	13.77%	14.98%
Zero-shot (RS-WebLI fine-tuned)	31.83%	29.39%
Le Jeune et al. [4]	37.10%	35.60%
Prototype-based FSOD (DINOv2) [5]	41.40%	26.46%
SIoU [6]	45.88%	52.85%
Ours (FLAME + RS-WebLI)	**53.96%**	**53.21%**

Table 1. Few-shot detection performance (Average Precision). Our method significantly outperforms zero-shot and few-shot baselines.

Key Statistics:

- **Adaptation Latency:** $\approx$ **1 minute** per class on a standard CPU.
- **DOTA Improvement: +22.1%** over Zero-shot RS-WebLI baseline.
- **DIOR Improvement: +23.8%** over Zero-shot RS-WebLI baseline.

Discussion & Conclusion

FLAME offers a resource-efficient framework for adapting foundation models to specialized domains:

- **Data Efficiency:** Implicitly selects mathematical "support vectors" through active learning.
- **Operational Viability:** Enables real-time human-in-the-loop workflows, essential for time-sensitive RS applications (e.g., illegal fishing monitoring).
- **Computational Efficiency:** Avoids costly fine-tuning by restricting training to a lightweight head.

Future Work: Extending FLAME to multi-class active selection and video-based RS anomaly detection.

References

[1] Xia et al. DOTA: A large-scale dataset for object detection in aerial images. CVPR 2018.
[2] Zhan et al. RSVG: Exploring Data and Models for Visual Grounding on Remote Sensing Data. IEEE TGRS 2023.
[3] Minderer et al. Scaling Open-Vocabulary Object Detection. NeurIPS 2024.
[4] Le Jeune et al. Improving few-shot object detection through a performance analysis on aerial and natural images. EUSIPCO 2022.
[5] Bou et al. Exploring robust features for few-shot object detection in satellite imagery. CVPR EarthVision Workshop 2024.
[6] Le Jeune et al. SIoU Loss for Few-Shot Object Detection. 2023.