Prompt-Free Unknown Label Generation for Open World Detection in Remote Sensing

Supplementary Material

A. Dataset Construction for OWOD

We adapt three remote sensing datasets (DOTA-v2 [43], FAIR1M [37], DIOR [19]) for OWOD following the incremental learning protocol [16]. DOTA-v2 uses a four-task incremental protocol where new classes are progressively introduced, requiring the model to discover and label unknowns autonomously at each task. FAIR1M and DIOR employ single-task evaluation, where a fixed subset of classes serves as known during training while the remainder are held out as unknowns during test-time discovery. Table 6 details the known/unknown class assignments for each task. We follow ORE [16] for experimentation on COCO dataset.

B. Injection Strategy Sensitivity

The Scene Context Token (SCT) aggregates scene-level co-occurrence patterns across all object queries and injects this contextual information into the detection pipeline (Eq. 3). We test five injection strategies to determine where to apply SCT: encoder only, different subsets of decoder layers (first 2, last 2, or all 4), and encoder + decoder combined.

Table 7 shows results on DOTA-v2 Task 1. Encoder-only injection achieves minimal gains (48.9 K-mAP, 28.0% U-R) over NO SCT configuration (44.7 K-mAP, 20.5% U-R) because early feature extraction lacks object-specific information. Progressive decoder injection shows steady improvements: first 2 layers achieve 49.8 K-mAP and 30.1% U-R, while last 2 layers reach 51.2 K-mAP and 32.4% U-R as context is applied where queries are more refined. Full decoder injection (all 4 layers) achieves best performance (52.1 K-mAP, 33.8% U-R), representing +7.4 K-mAP and +13.3% U-R over baseline. Combining encoder and decoder underperforms due to noisy early-stage conditioning. This validates that SCT should be applied during decoder refinement where object-level reasoning occurs.

C. Buffer-and-Cluster Ablation

Buffer-and-Cluster controls vocabulary expansion by requiring M clustered embeddings with pairwise cosine similarity exceeding threshold τ_{sim} before creating new category nodes (Eq. 10). We ablate these parameters to analyze trade-offs between vocabulary quality and expansion rate (Table 8). Low cluster size ($M = 3$) enables rapid vocabulary expansion but admits noisy detections, increasing WI to 9.2 and reducing alignment to 0.72. This aggressive expansion allows spurious visual patterns to be incorrectly registered as categories. Increasing to $M = 5$ substantially improves quality, alignment reaches 0.79, WI drops to 5.8, and 42 categories are discovered. This represents the optimal balance between expansion rate and quality. Stricter requirements ($M = 7$) further reduce WI to 4.1 but significantly limit expansion to 28 categories, potentially missing valid rare categories. For similarity threshold, loose requirements ($\tau_{\text{sim}} = 0.6$) produce 54 categories but suffer from semantic drift, where unrelated objects cluster together, reducing alignment to 0.74 and increasing WI to 7.3. The $\tau_{\text{sim}} = 0.7$ balances cluster cohesion and expansion. Overly strict thresholds ($\tau_{\text{sim}} = 0.8$) fragment valid categories into multiple clusters, reducing discovered vocabulary to 31 categories while only marginally improving quality. These results validate that $M = 5$ and $\tau_{\text{sim}} = 0.7$ provide robust vocabulary expansion without compromising semantic coherence.

D. Failure Case Categorization

We analyze failure modes to identify when CR2T produces low-quality semantic assignments (Text Align < 0.6). Table 9 categorizes 150 failure cases from the test set. The dominant failure mode is severe occlusion ($\geq$50%, 37.2% of failures), where incomplete visual features prevent accurate semantic synthesis. Tiny objects ($\leq$32px, 30.4%) lack sufficient spatial context for CR2T's region features. DHGA parent prediction errors (13.6%) propagate through CR2T, causing misalignment, when the hierarchical parent is wrong, synthesized embeddings diverge from ground truth. Ambiguous visual appearance (16.3%), objects equally resembling multiple categories, produces moderate alignment (0.55-0.65) but fails to reach the 0.79 target. Low scene context (2.5%) occurs in generic environments lacking spatial co-occurrence cues.

E. Per-Category Context Impact

In remote sensing imagery, objects exhibit varying degrees of visual distinctiveness. Ships and bridges possess distinctive signatures enabling identification through appearance alone, while large vehicles such as fuel trucks and cargo trucks appear nearly identical from aerial views, requiring spatial context for differentiation, fuel trucks concentrate near refineries while cargo trucks cluster near warehouses. Small vehicles and helicopters occupy an intermediate position with moderate distinguishing features that benefit from contextual disambiguation.

Table 6. Known and unknown class splits for each dataset. DOTA-v2 uses incremental discovery, FAIR1M and DIOR use single-task evaluation.

Dataset/Task	Known Classes	Unknown Classes	Total
DOTA-v2			
Task 1	plane, ship, storage-tank, baseball-diamond, tennis-court, basketball-court, ground-track-field, harbor, bridge, large-vehicle	small-vehicle, helicopter, swimming-pool, roundabout, soccer-ball-field, container-crane, airport, helipad	10 K 8 U
Task 2	Task 1 classes + small-vehicle, helicopter	swimming-pool, roundabout, soccer-ball-field, container-crane, airport, helipad	12 K 6 U
Task 3	Task 2 classes + swimming-pool, roundabout, soccer-ball-field	container-crane, airport, helipad	15 K, 3 U
Task 4	All 18 classes	None (closed-set evaluation)	18 K, 0 U
FAIR1M			
Training	Boeing737, Boeing747, A220, A321, A330, ARJ21, C919, other-airplane, passenger-ship, motorboat, fishing-boat, tugboat, engineering-ship, liquid-cargo-ship, dry-cargo-ship, warship, small-car, bus, cargo-truck, dump-truck	Boeing777, Boeing787, A350, other-cargo-plane, military-aircraft, oil-tanker, bridge, tower-crane, container-crane, reach-stacker, straddle-carrier, mobile-crane, truck-tractor, excavator, tractor, trailer, van	20 K 17 U
DIOR			
Training	airplane, airport, baseball-field, basketball-court, bridge, chimney, dam, expressway-service-area, expressway-toll-station, golf-field, ground-track-field, harbor	overpass, ship, stadium, storage-tank, tennis-court, train-station, vehicle, windmill	12 K 8 U

Table 7. SCT injection point sensitivity on DOTA-v2 Task 1.

Injection Strategy	K-mAP	U-R	WI
NO SCT	44.7	20.5	14.3
Encoder only	48.9	28.0	11.2
First 2 decoder layers	49.8	30.1	10.5
Last 2 decoder layers	51.2	32.4	09.3
All 4 decoder layers (ours)	52.1	33.8	08.4
Encoder + all decoder	51.5	32.9	08.9

Table 8. Buffer-and-cluster parameter ablation on DOTA-v2 Task 1.

M	τ_{sim}	Text Align	Coherence	U-R	WI	Categories
3	0.7	0.72	0.76	38.4	9.2	68
5	0.7	0.79	0.84	41.2	5.8	42
7	0.7	0.81	0.86	39.8	4.1	28
10	0.7	0.82	0.88	37.2	3.4	19
5	0.6	0.74	0.78	39.7	7.3	54
5	0.7	0.79	0.84	41.2	5.8	42
5	0.8	0.80	0.86	38.6	4.9	31
5	0.9	0.81	0.87	36.1	4.2	22

Table 9. Failure mode analysis for CR2T on DOTA-v2 (150 sampled failure cases with Text Align < 0.6).

Failure Type	Freq. (%)	Text Align
Severe occlusion ($\geq$50%)	37.2	0.52
Tiny objects ($\leq$32px)	30.4	0.48
Ambiguous appearance	16.3	0.61
DHGA parent error	13.6	0.45
Low scene context	2.5	0.63

Table 10 presents per-category ablation results measuring performance with and without SCT to isolate context contributions within DHGA's hierarchical navigation. High-context categories (large vehicles, storage tanks) show substantial gains of +11.6 to +12.1 points because SCT's spatial co-occurrence signals guide DHGA toward contextually appropriate branches when visual features underdetermine the hierarchical path. Medium-context categories (small vehicles, helicopters) show moderate gains of +5.2 to +7.5 points, benefiting from scene-conditioned graph attention when appearance alone is insufficient. Low-context categories (ships, bridges, tennis courts) show minimal gains of +0.9 to +1.7 points since their distinctive visual signatures enable accurate hierarchical navigation without contextual cues. This pattern validates that SCT provides semantically meaningful scene-conditioned reasoning: gains directly correlate with visual ambiguity, confirming the mechanism enhances DHGA's graph navigation for contextually dependent categories rather than providing uniform performance boosts.

F. Cross-Dataset Generalization

To validate CR2T's generalization capability, we test whether the learned visual-to-semantic mapping transfers across datasets or merely memorizes training patterns. Table 11 shows three settings: in-domain, cross-dataset transfer, and mixed-dataset training. In-domain performance establishes baselines of 0.76-0.79 CLIP alignment across DOTA-v2, FAIR1M, and DIOR. Cross-dataset transfer degrades to 0.68-0.73 alignment, a 7-9% drop, but reveals

Table 10. Per-category SCT impact on DOTA-v2 Task 1. Context dependence reflects how strongly categories rely on scene co-occurrence for disambiguation.

Category	w/o SCT	w/ SCT	ΔmAP	Context Dep.	Scene Type
Large vehicle	46.8	58.9	+12.1	Very High	Roads, industrial yards
Storage tank	46.5	58.1	+11.6	Very High	Refineries, tank farms
Small vehicle	52.3	59.8	+7.5	Medium	Roads, parking lots
Helicopter	58.2	63.4	+5.2	Medium	Air bases, open aprons
Ship	68.4	70.1	+1.7	Low	Maritime surfaces
Bridge	71.2	72.3	+1.1	Low	Linear crossings
Tennis court	80.1	81.0	+0.9	Low	Sports facilities

Table 11. CR2T transfer across remote sensing datasets. In-domain uses single-dataset training, cross-dataset evaluates zero-shot generalization, mixed-dataset samples from multiple sources with identical data volume.

Train Dataset	Test Dataset	Text Align	Coherence	U-R	WI
In-Domain Performance					
DOTA-v2	DOTA-v2	0.79	0.84	41.2	5.8
FAIR1M	FAIR1M	0.77	0.82	38.5	6.2
DIOR	DIOR	0.76	0.81	39.1	6.4
Cross-Dataset Transfer					
DOTA-v2	FAIR1M	0.72	0.78	36.8	7.1
DOTA-v2	DIOR	0.70	0.76	35.4	7.8
FAIR1M	DOTA-v2	0.71	0.77	37.2	7.0
FAIR1M	DIOR	0.73	0.79	38.1	6.6
DIOR	DOTA-v2	0.68	0.75	34.9	7.9
Multi-Dataset Training (Mixed)					
DOTAv2+FAIR1M	DOTA-v2	0.80	0.85	42.3	5.3
DOTAv2+FAIR1M	FAIR1M	0.79	0.84	41.1	5.6
DOTAv2+FAIR1Mv2+DIOR	DOTA-v2	0.81	0.86	43.2	5.0

Table 12. HSGDeT vs traditional/ OVD methods (mAP/UR). †: text-interactive HSGDeT.

Method	Venue	DOTA2	DIOR	FAIR1M
Deformable-DETR[48]	ICLR'21	0.9/-	1.1/-	0.1/-
YOLOv12 [39]	NIPS'25	1.4/-	3.2/-	0.4/-
GroundingDINO[26]	ECCV'24	44.7/-	51.8/-	39.4/-
YOLO-World[1]	CVPR'24	43.4/-	51.2/-	38.1/-
RemoteClip[23]	TGRS'24	45.4/-	48.5/-	40.3/-
LAE-DINO[31]	AAAI'25	49.8/-	52.4/-	43.1/-
LLaMA-Unidetector[40]	TGRS'25	52.7/-	54.2/-	46.7/-
HSGDet	-	54.8/41.2	59.3/42.8	52.2/38.5
HSGDet †	-	58.1/42.2	65.6/43.7	60.7/44.8

Table 13. Computational efficiency on 1024×1024 image size.

Method	Params (M)	FLOPs (G)	FPS
Open Vocabulary Detectors (OVD)			
GroundingDINO[26]	180	290	11
YOLO-World[1]	65	127	18
LAE-DINO[31]	190	300	9
LLaMA-Unidetector[40]	400	670	2
Open World Detectors (OWOD)			
OW-DETR[7]	55	290	13
UC-OWOD[41]	61	260	15
CAT[27]	68	254	16
OWOBJ[46]	105	207	19
SkySense-O[47]	497	1082	4
HSGDet	167	261	15

that CR2T learns generalizable remote sensing semantics (spatial co-occurrence patterns like vehicles near infrastructure, ships in harbors) rather than dataset-specific artifacts. Multi-dataset training recovers this gap, achieving 0.79-0.81 alignment and nearly matching or exceeding single-dataset performance. These results demonstrate domain-agnostic semantic reasoning without requiring external language models or manual annotation.

G. Text Interactivity:

HSGDet supports optional text prompts via $t_{\text{prompt}} = \text{CLIP}_{\text{text}}(\text{“query”})$ to bias detection toward user-specified categories while preserving autonomous discovery capabilities. Table 12 compares against OVD methods using Task 1 protocol: classes are partitioned into known (text-prompted) and unknown, enabling evaluation of both prompted detection (mAP) and autonomous discovery (UR). With identical text prompts, HSGDet achieves 58.1% mAP versus LLaMA-Unidetector's 52.7% on DOTA-v2, while simultaneously discovering unknowns at 42.2% recall. OVD methods report only mAP (“-” for UR) as they lack mechanisms to autonomously discover unknown objects.

H. Efficiency Comparison

HSGDet demonstrates competitive efficiency (167M/261G/15 FPS) across OWOD/OVD paradigms (Table 13). Compared to OVD methods, HSGDet operates with significantly fewer parameters and lower computational cost than language-heavy detectors such as LLaMA-Unidetector (400M/670G/2 FPS) and LAE-DINO (190M/300G/9 FPS), while achieving a higher frame rate than GroundingDINO (180M/290G/11 FPS). Among OWOD methods, the most direct performance competitor SkySense-O requires $3\times$ more parameters (497M), $4\times$ more FLOPs (1082G), and runs at only 4 FPS, nearly $4\times$ slower than HSGDet. While HSGDet introduces additional parameters over standard OWOD detectors (e.g., CAT at 68M) due to the hierarchical semantic graph and CR2T module, this overhead is modest and justified by the added capability of autonomous semantic generation and continual vocabulary expansion, which no prior OWOD method provides. HSGDet thus strikes a practical balance between semantic richness and deployment efficiency.

I. Qualitative Detection Results

Fig. 4—5 presents visual comparisons between HSGDet and comparison methods on DOTA-v2 and COCO datasets. To ensure valid evaluation of label quality, we visualize only discovered objects whose ground-truth categories are present in the original dataset annotations. This allows direct comparison between CR2T-synthesized labels and true semantics, verifying that generated names accurately reflect the underlying object categories. On DOTA-v2 Task 1, HSGDet successfully labels discovered unknowns with semantically meaningful category-level names, for example, identifying a *small-vehicle* as “car” or a *swimming-pool* as “pool”, while OWOBJ mark them only as generic ”unknown” without semantic grounding. Similarly, HSGDet generalizes to natural images as shown in 5, producing semantically interpretable category-level labels for unknown objects, while OWOBJ only provide generic ”unknown” detections. The synthesized labels capture appropriate semantic granularity, enabling human interpretability and downstream use without requiring manual annotation.

Figure 4. Visualizations on remote sensing images.

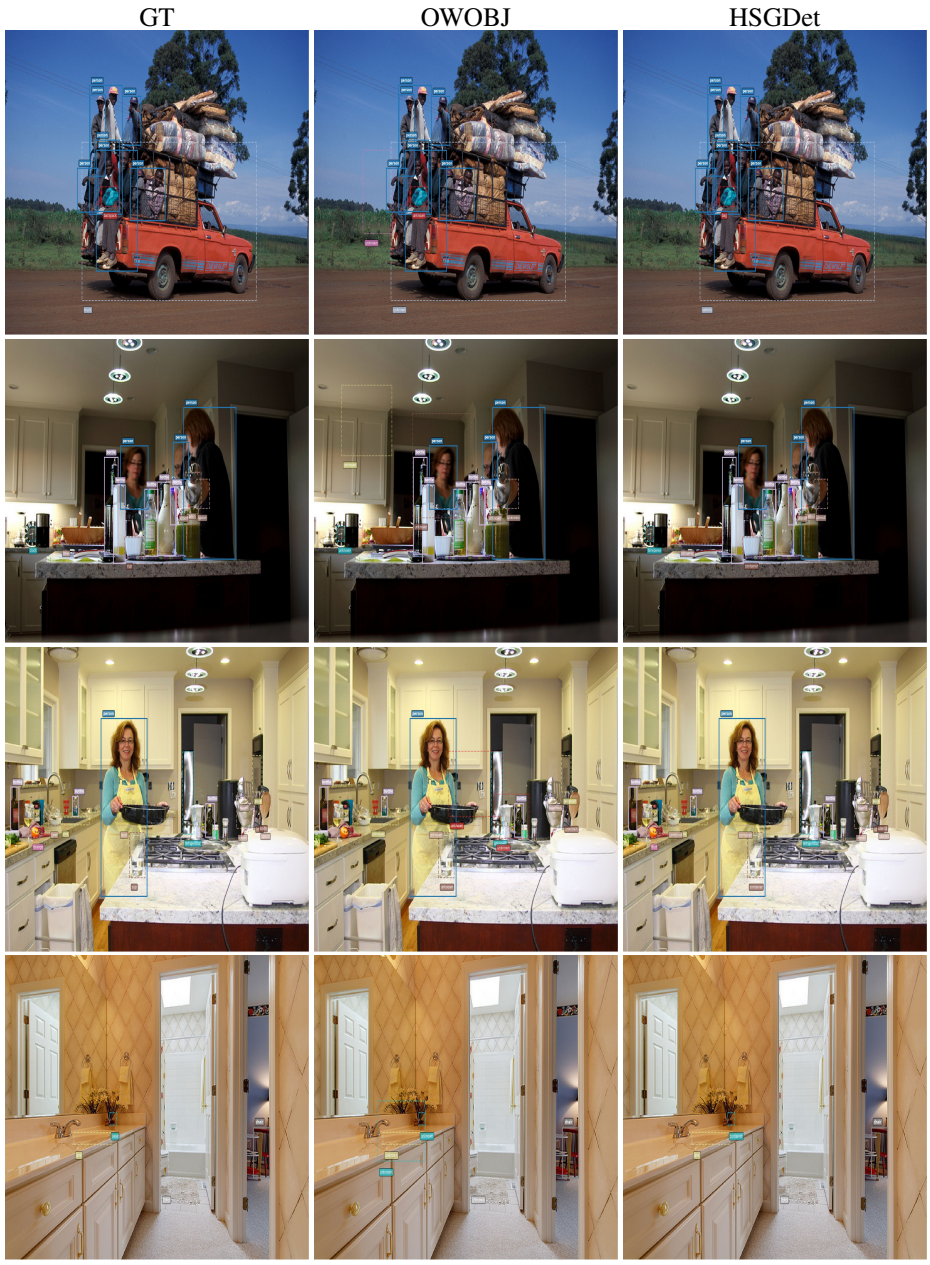

Figure 5. Visualizations on Natural images.