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

Abdullah Azeem[1] Ruisheng Wang[1†] Qingquan Li[1] Abubakar Siddique[2]
[1]Shenzhen University [2]Chongqing University
abdullahazeem06@outlook.com, ruiswang@szu.edu.cn, liqq@szu.edu.cn, abubakar@cqu.edu.cn

Abstract

Autonomous object detection in remote sensing requires systems that can discover new categories and assign them usable labels during deployment. Existing Open-World Object Detectors identify unknown objects but leave them unnamed until manual annotation. In contrast, Open-Vocabulary Detectors recognize unseen categories only with provided prompts at test time, lacking autonomous discovery or naming. This work presents HSGDet, a detector that achieves both discovery and semantic assignment at test time without external prompts. This method introduces DHGA that navigates a hierarchical semantic graph to perform scene-conditioned coarse-to-fine classification of detected objects. It leverages spatial co-occurrence patterns from surrounding scene context to produce classification confidence scores. High-scoring regions are identified as known objects, while low-scoring regions are flagged as unknown detections. Unknown regions pass to CR2T, which synthesizes text embeddings by fusing visual features, hierarchical parents, and scene context, enabling prompt-free labeling and vocabulary expansion. This approach enables prompt-free semantic labeling and supports autonomous vocabulary expansion without requiring external models. Results demonstrate that HSGDet outperforms state-of-the-art methods by a large margin of 6.6 points in Known mAP and 9.9 points in Unknown Recall. It also reduces Wilderness Impact by 36%, enabling scalable and autonomous aerial monitoring.

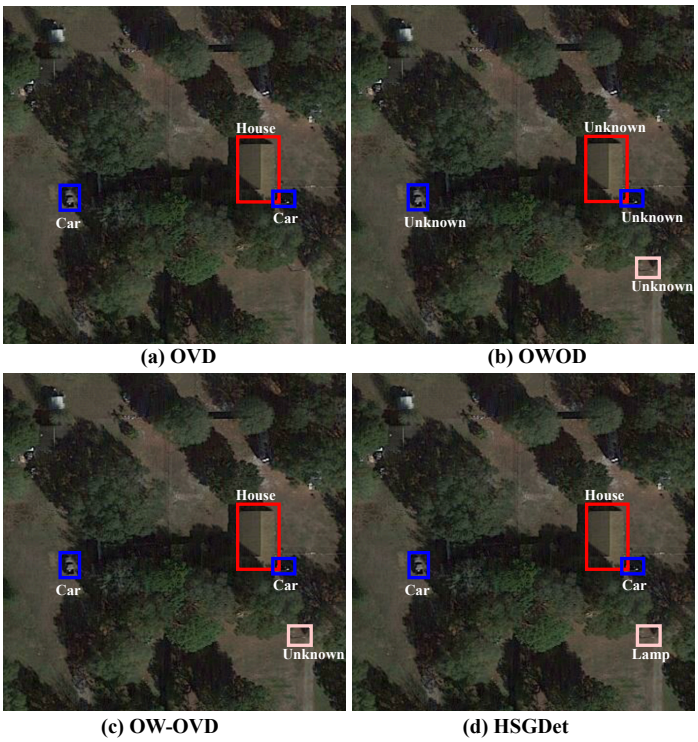

Figure 1. Comparison of OVD and OWOD paradigms on a aerial scene with prompted classes ("car" and "house") and an unprompted object ("lamp"). (a) OVD detects only prompted categories. (b) OWOD identifies novel classes as "unknown." (c) Unified OW-OVD detects the prompted classes but flags "lamp" as "unknown". (d) HSGDet autonomously discovers and labels it as "lamp" via contextual reasoning.

1. Introduction

Object detection in remote sensing (RS) enables rapid damage assessment for disaster response [10] and infrastructure monitoring [18]. Traditional closed-set detectors [34, 39, 48] assume a fixed label space defined at training time, which critically limits their deployment when unexpected categories emerge in operational scenarios.

Open-Vocabulary Detection (OVD) [23, 26, 47] leverages large-scale vision-language pretraining to enable zero-shot recognition of categories specified at test time. However, OVD strictly requires predefined vocabularies and explicit prompts. This makes it incapable of discovering or labeling truly unanticipated objects, such as the "lamp" (Fig. 1(a)), unless manually specified. Building on OVD's strengths, Open-World Object Detection (OWOD) shifts focus to autonomous discovery of instances outside the

training set. It extends closed-set detectors with mechanisms like energy-based scoring [16] or feature orthogonality [38] to flag low-confidence regions as "unknowns". This approach enables incremental learning without retraining. However, OWOD provides no semantic grounding for these unknowns (Fig. 1(b)), leaving them as anonymous placeholders that require manual annotation for meaningful use [12, 41]. Unified models [24, 42] further evolve this paradigm by merging OVD's zero-shot capabilities with OWOD's discovery mechanisms into hybrid frameworks. These systems recognize prompted known classes while flagging unprompted instances (Fig. 1(c)). Although unified models partially mitigate prompt dependency and anonymous labeling, they still rely on external aids like LLM-generated attributes or foundation model pseudo-labels for semantic assignment to unknowns. This reliance limits true autonomy in deployment.

RS imagery presents unique challenges, including context-dependent semantics and unknown ategories that existing paradigms fail to address. Identical visual patterns carry different meanings depending on surroundings, complicating reliable scene-level understanding in RS scenarios. Unknown categories arise routinely in operational deployments such as disaster response and infrastructure monitoring. As illustrated across Fig. 1(a)–(c), these paradigms address recognition and discovery but converge on a common shortcoming. The lack of autonomous semantic generation and continual vocabulary expansion hinders fully autonomous labeling in RS applications. To address this, we propose HSGDet (Fig. 1(d)), which achieves prompt-free discovery and labeling of unknowns. HSGDet leverages learned contextual co-occurrence patterns within a hierarchical semantic graph for autonomous RS object recognition. Specifically, Deformable Hierarchical Graph Attention (DHGA) enables scene-conditioned coarse-to-fine classification by navigating the hierarchical semantic graph. DHGA routes queries toward contextually appropriate branches, ensuring semantics are resolved by surroundings rather than visual features alone. Complementing this, the Context-Aware Region-to-Text module (CR2T) synthesizes text embeddings for flagged unknowns using learned contextual patterns. CR2T fuses refined visual queries, scene cues, and the nearest hierarchical parent for immediate semantic grounding. New categories are registered in the graph, enabling continual vocabulary expansion across evolving RS operational scenarios. This end-to-end mechanism ensures discovered objects gain usable labels autonomously, bypassing manual annotation or language models In summary, these are the following contributions:

- To the best of our knowledge, this is the first OWOD method with autonomous semantic generation and continual vocabulary expansion, enabling discovered unknowns to be semantically grounded without manual annotation or external language models.
- We propose DHGA that learns spatial co-occurrence patterns from the scene. A learnable context token aggregates co-occurrence signals across queries, enabling scene-conditioned hierarchical navigation to improve classification confidence and known-unknown separation in aerial imagery.
- We propose Context-Aware Region-to-Text (CR2T) that synthesizes semantic embeddings for unknown objects by leveraging the learned co-occurrence patterns. This achieves 0.79 alignment with ground-truth text embedding without requiring external language models.
- State-of-the-art results across three remote sensing and natural images benchmarks (DOTA-v2, FAIR1M, DIOR, and COCO) with consistent 6-10 points improvements in both Known mAP and Unknown Recall. It also reduces Wilderness Impact by 36%, enabling scalable and autonomous aerial monitoring.

2. Related Work

2.1. Open-Vocabulary Object Detection

Open-Vocabulary detection (OVD) leverages vision-language models to detect objects beyond training categories using text prompts [1, 14, 21, 31, 40]. Methods employ knowledge distillation from vision-language models [6, 29], region-text pre-training [17, 20], and prompt engineering [2, 4] to enable zero-shot generalization. Recent advances improve cross-modal alignment through deformable attention and key feature matching [15], while others exploit language hierarchies for structured vocabulary organization [11]. Scene graph-based approaches [36] leverage relational structures to enhance object relationships. In remote sensing, domain-specific foundation models provide visual-linguistic representations [23, 47], enabling open-vocabulary detection through domain-specific pretraining and region-text alignment [14, 21, 31]. Despite these advances, OVD methods require predefined text vocabularies at inference time, cannot autonomously discover unanticipated objects, and operate on flat category structures without hierarchical semantic reasoning [5, 25].

2.2. Open-World Object Detection

Open-World Object Detection (OWOD) framework addresses progressive learning scenarios where detectors encounter previously unseen objects [16]. Early methods develop generalized objectness through energy-based classification [8, 16] and probabilistic modeling [49]. Subsequent works enhance unknown identification via cascade decoding architectures [27], unknown-classified frameworks [41], and feature orthogonality constraints [38]. Recent advances introduce partial optimal transport for attribute-based clas-

sification [44], and view-consistent learning for multi-view scenarios [45]. In remote sensing contexts, methods leverage multimodal large language models for unknown discovery [35], self-adaptive language models for semantic grounding [13], and foundation model distillation for enhanced objectness in aerial imagery [12]. Distinguishing remote sensing objects remains challenging due to intricate spatial distributions and complex contextual dependencies across scenes. Learned contextual patterns provide critical disambiguation signals that are unavailable through standard flat vision-language alignment methods. When unknowns are identified, existing methods lack semantic grounding mechanisms and depend entirely on manual annotation. This reliance on manual vocabulary expansion [12, 41] limits deployment in autonomous open-world monitoring applications for remote sensing.

2.3. Unified Open-Vocabulary and Open-World Detection

Recent work OW-OVD [42] combines OVD and OWOD through Visual Similarity Attribute Selection and Hybrid Attribute-Uncertainty Fusion, improving unknown recall. Foundation model approaches [12, 13] leverage large language models for unknown labeling. However, these methods rely on predefined attribute vocabularies or external LLMs at inference time and employ flat classification without hierarchical semantic reasoning. Unlike these approaches, HSGDet learns a fully autonomous semantic generation mechanism during training, requiring no external models or manual labeling at deployment.

3. Method

3.1. Problem Formulation

We address open-world object detection with autonomous semantic generation and continual vocabulary expansion. Given an image $\mathcal{I}$ and known class labels $\mathcal{Y}$ for K classes, the detector fulfills three tasks. First, accurately detect known objects $\hat{\mathcal{D}}_{\text{known}} = \{(b_i, \hat{y}_i)\}$ with bounding boxes b_i and class $\hat{y}_i \in \mathcal{K}$. Second, flag low-confidence regions as unknown $\hat{\mathcal{D}}_{\text{unknown}} = \{(b_j, \text{unknown})\}$ outside the training distribution. Third, autonomously generate semantic labels ℓ_{new} for unknowns without prompts or annotations. The vocabulary expands to $\mathcal{K}' = \mathcal{K} \cup \{\ell_{\text{new}}\}$ for continual learning.

3.2. Base Model

Our model builds upon the Deformable DETR architecture [48] as shown in Fig. 2. We extract multi-scale image features $\{F_1, F_2, F_3, F_4\}$ using frozen CLIP's vision encoder [33]. These are refined by a transformer decoder processing N learnable object queries $Q \in \mathbb{R}^{N \times d}$ through self-attention, deformable spatial attention, Deformable Hierarchical Graph Attention (DHGA), and feed-forward networks. To enable semantic understanding beyond known classes, we organize categories into a hierarchical semantic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ derived from WordNet IS-A taxonomies. Nodes represent categories embedded in CLIP text space [33], and edges encode parent-child relationships for coarse-to-fine classification. Building on this, we propose two novel modules enabling autonomous semantic generation and continual vocabulary expansion. DHGA navigates $\mathcal{G}$ conditioned on scene context to classify known and unknown objects. Context-Aware Region-to-Text (CR2T) synthesizes semantic embedding for flagged unknowns, expanding $\mathcal{G}$ without external prompts or manual annotation.

3.3. Hierarchical Semantic Graph

OWOD requires discovering novel objects using contextual relationships between known and unknown regions within scenes. The hierarchical semantic graph provides this structured taxonomy by encoding parent-child relationships among object categories across granularity levels. This organization enables autonomous semantic generation by grounding unknowns through scene context derived from detected known objects. It also supports dynamic vocabulary expansion as new categories are discovered and integrated into the hierarchy.

We define the hierarchical semantic graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, representing semantic relationships among categories. Each node $v \in \mathcal{V}$ corresponds to an object category with a CLIP embedding $t_v \in \mathbb{R}^d$. Additionally, each node has a learnable key embedding $e_v \in \mathbb{R}^d$ for query-dependent attention during navigation. Edges $(u, v) \in \mathcal{E}$ encode directed "is-a" parent-child relationships, forming a directed hierarchical taxonomy.

The adjacency matrix $A \in \{0, 1\}^{|\mathcal{V}| \times |\mathcal{V}|}$ formally represents the graph structure as:

$$A_{uv} = \begin{cases} 1, & \text{if } (u, v) \in \mathcal{E}, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

New nodes v_{new} with embeddings t_{new} and e_{new} are added under parent v_p during deployment. This enables seamless vocabulary expansion while preserving the hierarchical semantic structure for consistent open-world operation.

3.4. Deformable Hierarchical Graph Attention

While hierarchical semantic graph provides the structured taxonomy for semantic reasoning during detection, effectively navigating this graph requires specialized mechanisms. We introduce DHGA, which enables scene-conditioned coarse-to-fine classification. DHGA extends Deformable Graph Attention (DGA) [32], which provides sparse query-dependent node sampling for efficient feature aggregation. However, standard DGA treats all graph nodes

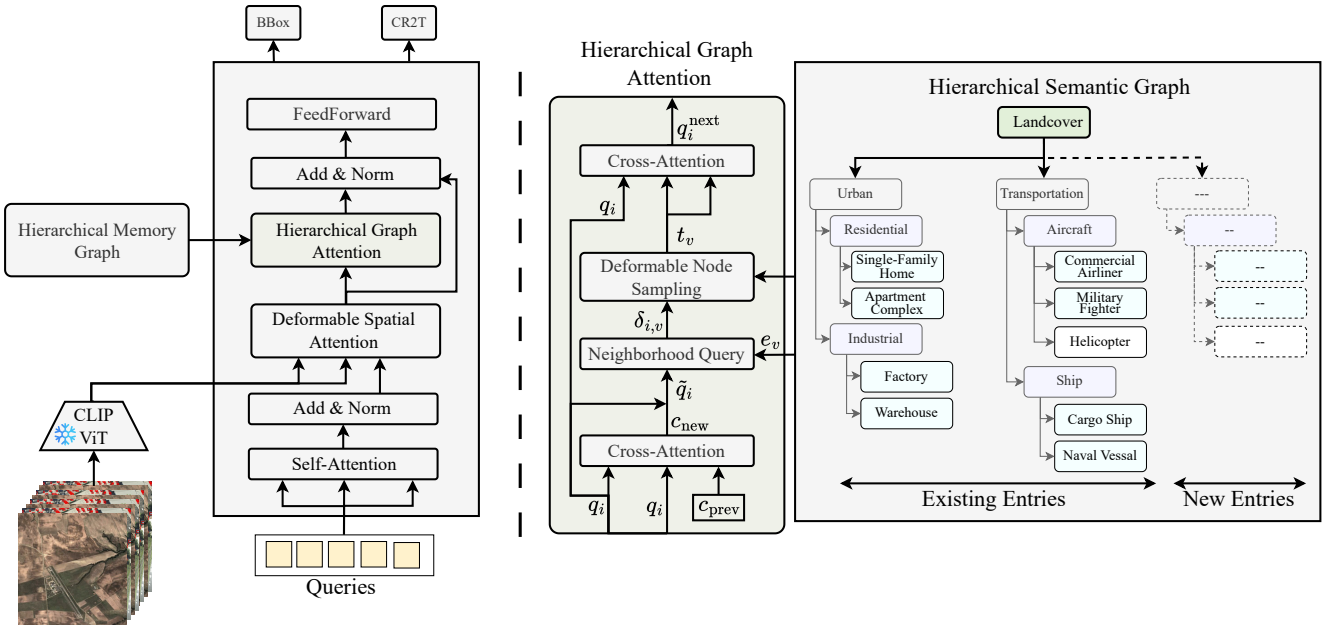

Figure 2. HSGDet architecture. The decoder refines object queries through transformer layers incorporating DHGA and scene context. DHGA performs hierarchical classification while CR2T synthesizes embedding for unknown objects, expanding the semantic graph autonomously.

uniformly without leveraging parent-child hierarchical relationships. DHGA addresses this by incorporating hierarchical semantic navigation with learnable keys for adaptive path selection conditioned on scene context.

To enable hierarchical navigation, we leverage co-occurrence patterns among detected objects to guide navigation toward contextually relevant categories. A learnable Scene Context Token (SCT) $c \in \mathbb{R}^d$ aggregates co-occurrence information across all queries at each decoder layer. The SCT updates via cross-attention to current queries and enriches each through residual addition (Eq. 3). This injects shared scene cues conditioning coarse-to-fine refinement and classification without requiring external prompts.

$$c_{\text{new}} = \text{CrossAttn}(c_{\text{prev}}, q_i, q_i) \tag{2}$$

$$\tilde{q}_i = q_i + c_{\text{new}} \tag{3}$$

where q_i represents i_{th} query, c_{prev} denotes context from the previous layer with $c^{(0)}$ initialized as a trainable parameter.

To navigate the semantic graph $\mathcal{G}$ efficiently, we employ Top-K sampling to select semantically relevant nodes. For each context-enriched query $\tilde{q}_i$, we compute relevance scores to all node key embeddings e_v. We then identify the Top-K candidates via Eq. 5. This selection reduces complexity from $\mathcal{O}(N|\mathcal{V}|)$ to $\mathcal{O}(NK)$ while ensuring contextually relevant semantic fusion.

$$\delta_{i,v} = \tilde{q}_i^T e_v \quad \forall v \in \mathcal{V}, \tag{4}$$

$$\mathcal{S}_i = \text{Top-K}(\text{softmax}(\delta_i), \text{K}) \tag{5}$$

where e_v denotes semantic node embedding $\delta_i = [\delta_{i,1}, \ldots, \delta_{i,|\mathcal{V}|}]$ and $\mathcal{S}_i \subset \mathcal{V}$ contains the K highest-scoring nodes.

For sampled nodes in $\mathcal{S}_i$, we retrieve their CLIP text embeddings t_v for semantic alignment. Context-enriched queries $\tilde{q}_i$ compute attention weights determining each selected node's contribution to query refinement (Eq. 6, 7). Since weights depend on $\tilde{q}_i$ rather than raw queries, semantic fusion naturally biases toward context-appropriate categories.

$$\beta_{i,v} = \text{softmax}\left(\frac{\tilde{q}_i^T t_v}{\sqrt{d}}\right) \tag{6}$$

$$q_i^{\text{next}} = q_i + \sum_{v \in \mathcal{S}_i} \beta_{i,v} t_v \tag{7}$$

where t_v is clip embedding related to node e_v.

DHGA enables coarse-to-fine semantic navigation over graph $\mathcal{G}$ through L decoder layers. Early layers l use context-enriched queries $\tilde{q}_i^{(l)}$ to attend broadly to parent nodes with coarse categories, guided by scene context c_{new}. Top-K sampling prioritizes semantically broad nodes. As l increases, $\tilde{q}_i^{(l)}$ captures finer visual details, shifting attention to child nodes. This traversal emerges from layer-wise refinement of $\tilde{q}_i^{(l)}$, hierarchical loss $\mathcal{L}_{\text{hier}}$ ensuring ancestral

path consistency, and scene conditioning biasing towards contextually relevant descendants. At the final layer, refined query $q_i^{(\text{final})}$ produces bounding boxes, with classification via attention weights $\beta_{i,v}$, obviating separate heads. Queries with max $\beta_{i,v} < \tau_{\text{unk}} = 0.4$ route to CR2T for unknown embedding synthesis. This process achieves semantic specialization without explicit path planning.

3.5. Context-Aware Region-to-Text

When a query exhibits low confidence across all known categories, CR2T synthesizes a semantic embedding to characterize this unknown object. Although the query does not confidently match any known class, the hierarchical graph structure still provides valuable grounding through the parent node with the highest attention score.

Unknown Embedding Generation. For an unknown query, we first identify its most likely parent node in the semantic hierarchy. This parent node represents the closest known category in the taxonomy, providing semantic context about what general category the unknown object resembles. For example, if an unknown object visually resembles ships but does not match any specific ship type in the training set, the parent might be "ship" or "vehicle." Specifically, we select the parent node v_p as the node with the maximum attention weight among the sampled nodes from DHGA, as defined in Eq. 8. Although the maximum attention score falls below the unknown threshold, indicating that the query does not confidently belong to any known class, v_p still captures valuable hierarchical information about where this unknown fits within the taxonomy.

$$v_p = \arg\max_{v \in \mathcal{S}_i} \beta_{i,v} \tag{8}$$

Next, we synthesize the semantic text embedding for this unknown by fusing three complementary information sources. The embedding is generated as shown in Eq. 9, where f is a learnable fusion function (implemented as a multi-layer perceptron).

$$t_{\text{new}} = f([q_i^{(\text{final})}; c; t_{v_p}]) \tag{9}$$

where $q_i^{(\text{final})}$ is the refined visual query from DHGA, c is the scene context token and t_{v_p} is the text embedding of the parent node. This multi-source fusion ensures that discovered unknowns are not merely flagged as generic "unknowns" but receive meaningful semantic embeddings that position them appropriately within the taxonomy based on visual features, contextual relationships, and hierarchical grounding. ince t_{new} resides directly in CLIP's embedding space, it serves as the semantic label itself.

Continual Vocabulary Expansion. HSGDet employs a buffer-and-cluster strategy to validate that only coherent, recurring object types are added to the vocabulary, thereby preventing noise and spurious detections from polluting the taxonomy. Unknown embeddings t_{new} are accumulated in a buffer during inference. When a cluster of $M = 5$ embeddings forms, where each pair exhibits cosine similarity exceeding the threshold $\tau_{\text{sim}} = 0.7$, HSGDet automatically creates a new category node in the semantic graph. The new node is formally defined in Eq. 10. Once created, this new category becomes immediately available for detection in subsequent images, enabling continuous vocabulary growth that is grounded in the semantic hierarchy and validated through consistent observations.

$$\begin{aligned} v_{\text{new}} \leftarrow \{t_{\text{new}} &= \frac{1}{M}\sum_{i=1}^{M} t_i, \\ e_{\text{new}} &= e_{v_p},\ p(v_{\text{new}}) = v_p\} \end{aligned} \tag{10}$$

where t_{new} is the averaged text embedding of the cluster members, $e_{\text{new}} = e_{v_p}$ inherits the visual prototype embedding from the parent node, and $p(v_{\text{new}}) = v_p$ sets the parent-child relationship.

Since t_{new} resides directly in CLIP's embedding space, it serves as the semantic label itself, enabling vocabulary-free semantic grounding within the graph. For human-readable output at inference, we perform a hierarchically-constrained nearest-neighbor search against CLIP text embeddings of categories under parent v_p, converting the synthesized embedding to an interpretable label. This retrieval step is a post-hoc output conversion for user interpretation only,all internal detection, classification, and graph expansion operate entirely on continuous CLIP embeddings without any fixed vocabulary constraint.

3.6. Training Objectives

The model is trained end-to-end with a multi-task loss combining detection objectives and hierarchical semantic losses:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_1 \mathcal{L}_{\text{hier}} + \lambda_2 \mathcal{L}_{\text{CR2T}} \tag{11}$$

where $\lambda_1 = 0.5$ and $\lambda_2 = 0.3$ are loss weights. The detection loss $\mathcal{L}_{\text{det}}$ includes only bounding box regression (L1 + GIoU), with classification replaced by hierarchical graph navigation.

The hierarchical navigation loss $\mathcal{L}_{\text{hier}}$ supervises semantic path traversal. For each ground-truth object of class $\hat{c}$, the model is supervised to assign high attention to all ancestor nodes on the path from root to $\hat{c}$:

$$\mathcal{L}_{\text{hier}} = -\frac{1}{N_{\text{gt}}} \sum_{i=1}^{N_{\text{gt}}} \sum_{v \in \mathcal{P}(c_i)} \log \beta_{i,v} \tag{12}$$

where $\mathcal{P}(c_i)$ denotes the ancestral path and N_{gt} is the number of ground-truth objects.

The CR2T loss $\mathcal{L}_{\text{CR2T}}$ supervises text embedding synthesis. By masking known classes as pseudo-unknowns with

combined L1 alignment and contrastive sibling guidance (Eq. 13). The L1 loss constrains synthesized embedding to valid CLIP space, while contrastive guidance prevents memorization of individual classes by enforcing separation from siblings. Together, they enable the network to learn compositional patterns of how objects relate to parent categories, transferring this knowledge to novel unknowns at inference.

$$\mathcal{L}_{\text{CR2T}} = \frac{1}{N_{\text{mask}}} \sum_{i=1}^{N_{\text{mask}}} \Bigg[\left(1 - \cos(t_i^{\text{pred}}, t_i^{\text{gt}})\right) + \lambda_c \mathcal{L}_{\text{contrast}}^{i} \Bigg] \tag{13}$$

where t_i^{pred} is the synthesized embedding and t_i^{gt} is the ground-truth CLIP text embedding. The contrastive term prevents semantic collapse:

$$\mathcal{L}_{\text{contrast}}^{i} = \max\Big(0, \cos(t_i^{\text{pred}}, t_i^{\text{neg}}) - \cos(t_i^{\text{pred}}, t_i^{\text{gt}}) + \gamma\Big) \tag{14}$$

where t_i^{neg} is sampled from sibling classes, $\lambda_c = 0.1$, and $\gamma = 0.2$ is the margin.

4. Experimental Results

4.1. Datasets, Metrics, and Implementation

We evaluate on DOTA-v2 [43] (18 classes), FAIR1M [37] (37 categories), DIOR [19] (20 classes), and COCO [22] (80 categories). We adopt standard metrics [16], i.e. Known mAP, Unknown Recall, and Wilderness Impact. DOTA-v2 and COCO use four-task incremental evaluation [16], while others use single-task evaluation. We use Deformable DETR with a frozen CLIP ViT-B/32 encoder [33]. The decoder has $L = 4$ layers with $N = 300$ queries. The CR2T module uses a 3-layer MLP [768→512→256] with dropout 0.1. Training employs AdamW optimizer (lr=1×10^{-4}, weight decay=1×10^{-5}) with cosine annealing over 50 epochs and 5-epoch warmup. During training, 30% of known classes are randomly masked as pseudo-unknowns to supervise the CR2T module's embedding generation capability.

4.2. Ablation Studies

4.2.1. Component Analysis

Table 1 ablates HSGDet's core components on DOTA-v2 Task 1, starting from a baseline defined as a CLIP Vision Deformable DETR decoder with ORE-style energy-based unknown detection, which flags unknowns but cannot assign names. Adding DHGA enables queries to navigate the semantic graph hierarchically, improving U-R by +7.5 through coarse-to-fine classification. The SCT further boosts U-R by +5.8 by integrating scene-level co-occurrence information, conditioning classification decisions on global context (e.g., recognizing vehicles near runways). Incorporating the CR2T module synthesizes text embedding for unknowns via fusion of visual features, semantic hierarchy, and scene context. The full model achieves 54.8 K-mAP and 41.2% U-R, marking +10.1 and +20.7 absolute gains over baseline, while reducing WI to 5.8, indicating improved known-unknown discrimination.

Table 1. Ablation study of HSGDet components on DOTA-v2. K-mAP: Known mAP, U-R: Unknown Recall, WI: Wilderness Impact.

DHGA	SCT	CR2T	K-mAP ↑	U-R ↑	WI ↓
✗	✗	✗	44.7	20.5	14.3
✓	✗	✗	48.9	28.0	11.2
✓	✓	✗	52.1	33.8	08.4
✓	✓	✓	54.8	41.2	05.8

Table 2. CR2T ablation on DOTA-v2. Text Alignment (TA): cosine similarity to ground-truth CLIP embedding. Semantic Coherence (SMC): intra-class visual similarity. VF: Visual Features, SC: Scene Context, HP: Hierarchical Parent, VO: Visual Only, SC:Sibling Context, Full: Full CR2T.

CR2T	VF	SC	HP	TA ↑	SMC ↑	U-R ↑	WI ↓
CR2T Configurations							
+VO	✓	✗	✗	0.56	0.67	37.4	07.5
+HP	✓	✗	✓	0.63	0.71	38.6	06.7
+SC	✓	✗	✓	0.66	0.74	39.1	06.4
Full	✓	✓	✓	0.79	0.82	41.2	05.8
Training Data Sensitivity							
10%	✓	✓	✓	0.74	0.79	39.2	06.5
20%	✓	✓	✓	0.77	0.82	40.3	06.1
30%	✓	✓	✓	0.79	0.84	41.2	05.8

4.2.2. CR2T Analysis

The CR2T mechanism synthesizes text embeddings for unknown objects by fusing visual region features with hierarchical guidance from semantic graph parent and sibling nodes, along with scene context from co-occurring tokens. This fusion produces embeddings compatible with CLIP text space, enabling immediate semantic interpretation. Table 2 presents ablation results on DOTA-v2 Task 1, evaluated using text alignment (cosine similarity to ground-truth CLIP embeddings) and semantic coherence (intra-class visual similarity). The visual-only variant (VO), lacking structural guidance, achieves a 0.56 cosine alignment and 37.4% U-R. Introducing hierarchical parent (HP) guid-

Table 3. Comparison of OWOD methods on DOTA v2

Method	Venue	Task 1			Task 2		Task 3		Task 4
		K-mAP ↑	U-R ↑	WI ↓	K-mAP ↑	U-R ↑	K-mAP ↑	U-R ↑	K-mAP ↑
ORE [16]	CVPR-21	42.3	18.5	15.2	35.8	21.3	28.4	23.7	22.1
OW-DETR [7]	CVPR-22	43.8	20.1	13.8	37.2	23.4	29.6	25.8	23.4
UC-OWOD [41]	ECCV-22	44.2	21.8	12.5	38.1	24.6	30.7	27.1	24.5
PROB [49]	CVPR-23	45.7	24.3	10.8	39.4	27.2	32.1	29.5	25.8
CAT [27]	CVPR-23	46.1	25.1	11.2	40.2	28.1	33.4	30.3	27.1
SS-OWOD [30]	AAAI-24	44.9	23.2	11.8	39.8	26.9	32.7	28.9	26.4
OrthogonalDet [38]	CVPR-24	47.8	28.6	09.3	42.6	31.4	36.9	34.8	31.2
OW-OVD [42]	CVPR-25	48.3	29.5	08.9	43.1	32.4	37.8	35.8	32.5
OWOBJ [46]	CVPR-25	48.5	30.2	08.7	43.4	33.1	38.2	36.5	32.9
SkySense-O† [47]	CVPR-25	50.2	31.5	08.1	44.8	34.2	39.6	37.1	34.3
HSGDet		**54.8**	**41.2**	**05.8**	**57.9**	**44.7**	**60.1**	**47.8**	**62.3**

Table 4. Comparison of OWOD methods performance on FAIR1M and DIOR.

Method	Venue	FAIR1M				DIOR			
		K-mAP ↑	U-R ↑	WI ↓	A-mAP ↑	K-mAP ↑	U-R ↑	WI ↓	A-mAP↑
ORE [16]	CVPR-21	38.5	15.2	18.7	35.1	45.2	19.3	14.8	41.5
OW-DETR [7]	CVPR-22	39.8	17.3	16.9	36.4	46.7	21.5	13.2	43.1
UC-OWOD [41]	ECCV-22	40.5	18.9	15.8	37.2	47.3	23.1	12.3	44.0
PROB [49]	CVPR-23	42.1	21.4	13.5	38.9	49.1	26.2	10.5	46.3
CAT [27]	CVPR-23	42.8	22.7	12.8	39.5	49.8	27.5	9.9	47.1
SS-OWOD [30]	AAAI-24	-	-	-	-	48.5	24.8	11.2	45.6
OrthogonalDet [38]	CVPR-24	44.2	25.9	11.2	41.1	51.5	30.8	8.4	49.2
OW-OVD [42]	CVPR-25	45.3	27.9	10.1	42.5	52.7	33.1	7.6	50.5
OWOBJ [46]	CVPR-25	45.1	27.5	10.4	42.2	52.3	32.4	7.9	50.1
SkySense-O† [47]	CVPR-25	45.8	28.6	9.7	42.9	52.9	33.6	7.4	50.8
HSGDet		**52.4**	**38.5**	**6.2**	**49.7**	**59.3**	**42.8**	**5.2**	**57.1**

ance anchored in the taxonomy improves alignment to 0.63 and U-R to 38.6 (+1.2). Adding sibling context (SC) relational information further enhances performance to 0.66 alignment and 39.1 U-R. The full CR2T module, integrating hierarchical parents, sibling context, and scene context, obtains the highest scores of 0.79 alignment, 0.84 semantic coherence, and 41.2 U-R (+3.8 over VO), significantly improving disambiguation of unknowns.

We also study training data sensitivity by varying the ratio of masked pseudo-unknown classes. At 10%, alignment is 0.74 with 39.2 U-R, 20% improves to 0.77 and 40.3 U-R, 30% achieves optimal performance with 0.79 alignment and 41.2 U-R. This indicates that balanced pseudo-unknown masking during training prevents known-class performance degradation while maximizing unknown discovery accuracy.

4.3. Comparisons Against SOTA

Table 3, 4, 5 summarize HSGDet's performance across DOTA-v2, FAIR1M, DIOR, and COCO datasets under the open-world detection protocol. HSGDet consistently outperforms SOTA models across all benchmarks in K-mAP, U-R, and WI metrics. On DOTA-v2 Task 1, HSGDet achieves 54.8 K-mAP and 41.2 U-R, surpassing SkySense-O by +4.6 and +9.7 respectively, with K-mAP and U-R further improving to 62.3 and 47.8 by later tasks. On FAIR1M and DIOR, HSGDet attains gains of +6.6 and +6.4 K-mAP, with unknown recall improvements of +9.9 and +9.2 over baselines. WI scores remain consistently lower than competing methods (e.g., 5.8 vs. 8.1 on DOTA-v2 Task 1), indicating superior known-unknown discrimination. Unlike SkySense-O, which experiences a K-mAP degradation of -15.9 points across tasks, HSGDet steadily improves throughout incremental learning by organizing new categories within a semantic taxonomy rather than treating them as independent additions. On COCO, HSGDet generalizes effectively to natural images, achieving 81.3 K-mAP and 78.8 U-R on Task 1, with K-mAP varying only between 80.9 and 82.4 across all tasks, while OW-OVD suffers a severe drop of -13.8 points. These results demonstrate that

Table 5. Comparison of OWOD methods performance on COCO dataset. The comparison results are from OW-OVD [42].

Method	Venue	Task 1		Task 2		Task 3		Task 4
		K-mAP ↑	U-R ↑	K-mAP ↑	U-R ↑	K-mAP ↑	U-R ↑	K-mAP ↑
OW-DETR [7]	CVPR-22	71.5	05.7	62.8	06.2	45.2	06.9	38.2
CAT [27]	CVPR-23	74.2	24.0	67.6	23.0	51.2	24.6	45.4
PROB [49]	CVPR-23	73.4	17.6	66.3	22.0	47.8	24.8	42.6
OrthogonalDet [38]	CVPR-24	71.6	24.6	64.0	27.9	52.1	31.9	48.7
MEPU-FS [3]	TNNLS-25	74.3	37.9	68.0	35.8	50.2	35.7	43.7
SGROD [9]	TIP-25	73.2	48.0	64.7	48.9	47.4	47.7	42.5
SKDF [28]	TPAMI-25	69.4	60.9	63.8	60.0	46.2	58.6	41.8
OW-OVD [42]	CVPR-25	78.6	76.2	78.5	79.8	69.6	78.4	64.8
HSGDet		**81.3**	**78.8**	**80.9**	**85.6**	**82.4**	**84.3**	**81.2**

Figure 3. Comparison between OWOBJ and HSGDet on DOTA-v2. OWOBJ detects the unseen objects without semantic grounding, whereas HSGDet detect the unseen object with proper class labels.

HSGDet effectively combines hierarchical semantic graph navigation with context-aware learning for robust open-world detection

5. Conclusion

Autonomous object detection in remote sensing requires handling dense distributions, arbitrary orientations, and complex contextual dependencies. Current OWOD methods identify unknown objects but cannot autonomously name them, leaving discovered instances as unlabeled placeholders. This reliance on manual annotation for vocabulary expansion creates a fundamental scalability barrier for operational deployment. We presented HSGDet, the first open-world object detector to autonomously generate semantic labels for discovered unknowns without external models or human intervention. The method combines hierarchical graph navigation with real-time text embedding synthesis to expand its vocabulary during deployment, enabling immediate semantic grounding of novel categories without manual annotation. HSGDet achieves substantial improvements across benchmarks. On remote sensing datasets (DOTA-v2, FAIR1M, DIOR), it gains +6.6 K-mAP and +9.9 U-R while reducing WI by 36%. COCO experiments validate cross-domain generalization, maintaining stable performance (80.9-82.4 K-mAP) where competing methods degrade significantly.

Acknowledgement This work is supported by the Key Technological Innovation Program of Ningbo City under Grant No. 2024Z297.

References

[1] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16901–16911, 2024. 2

[2] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14084–14093, 2022. 2

[3] Ruohuan Fang, Guansong Pang, Wenjun Miao, Xiao Bai, Jin Zheng, and Xin Ning. Unsupervised recognition of unknown objects for open-world object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):11340–11354, 2025. 8

[4] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *European conference on computer vision*, pages 701–717. Springer, 2022. 2

[5] Shenghao Fu, Junkai Yan, Qize Yang, Xihan Wei, Xiaohua Xie, and Wei-Shi Zheng. A hierarchical semantic distillation framework for open-vocabulary object detection. *IEEE Transactions on Multimedia*, 2025. 2

[6] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations (ICLR)*, 2022. 2

[7] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9235–9244, 2022. 7, 8

[8] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9235–9244, 2022. 2

[9] Yulin He, Wei Chen, Siqi Wang, Tianrui Liu, and Meng Wang. Recalling unknowns without losing precision: An effective solution to large model-guided open world object detection. *IEEE Transactions on Image Processing*, 34:729–742, 2025. 8

[10] Victor Hertel, Christian Geiß, Marc Wieland, and Hannes Taubenböck. Rapid domain adaptation for disaster impact assessment: Remote sensing of building damage after the 2021 germany floods. *Science of Remote Sensing*, page 100287, 2025. 1

[11] Jiaxing Huang, Jingyi Zhang, Kai Jiang, and Shijian Lu. Open-vocabulary object detection via language hierarchy. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[12] Yangyang Huang, Jie Hu, and Ronghua Luo. Fmdl: Enhancing open-world object detection with foundation models and dynamic learning. *Expert Systems with Applications*, 275: 127050, 2025. 2, 3

[13] Yangyang Huang, Linhua Ye, and Ronghua Luo. Sallm: Open world object detection empowered by self adaptive learning and large model. *Expert Systems with Applications*, page 129375, 2025. 3

[14] Ziyue Huang, Yongchao Feng, Ziqi Liu, Shuai Yang, Qingjie Liu, and Yunhong Wang. Openrsd: Towards open-prompts for object detection in remote sensing images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8384–8394, 2025. 2

[15] Yunqing Jiang, Sunyuan Qiang, Wuchen Li, Zhao Huijia, and Yanyan Liang. Ov-kfa: Open-vocabulary object detection via key feature alignment. *Neurocomputing*, page 131790, 2025. 2

[16] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5830–5840, 2021. 2, 6, 7

[17] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11144–11154, 2023. 2

[18] Nadiia Kopiika, Andreas Karavias, Pavlos Krassakis, Zehao Ye, Jelena Ninic, Nataliya Shakhovska, Sotirios Argyroudis, and Stergios-Aristoteles Mitoulis. Rapid post-disaster infrastructure damage characterisation using remote sensing and deep learning technologies: A tiered approach. *Automation in Construction*, 170:105955, 2025. 1

[19] Ke Li, Gang Wan, Gong Cheng, Liqiu 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. 6

[20] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10965–10975, 2022. 2

[21] Yan Li, Weiwei Guo, Xue Yang, Ning Liao, Dunyun He, Jiaqi Zhou, and Wenxian Yu. Toward open vocabulary aerial object detection with clip-activated student-teacher learning. In *European Conference on Computer Vision*, pages 431–448. Springer, 2024. 2

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[23] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote

sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. 1, 2

[24] Lihao Liu, Juexiao Feng, Hui Chen, Ao Wang, Lin Song, Jungong Han, and Guiguang Ding. Yolo-uniow: Efficient universal open-world object detection. *arXiv preprint arXiv:2412.20645*, 2024. 2

[25] Mingxuan Liu, Tyler L Hayes, Elisa Ricci, Gabriela Csurka, and Riccardo Volpi. Shine: Semantic hierarchy nexus for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16634–16644, 2024. 2

[26] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 1

[27] Shuailei Ma, Yuefeng Wang, Ying Wei, Jiaqi Fan, Thomas H Li, Hongli Liu, and Fanbing Lv. Cat: Localization and identification cascade detection transformer for open-world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19681–19690, 2023. 2, 7, 8

[28] Shuailei Ma, Yuefeng Wang, Ying Wei, Enming Zhang, Jiaqi Fan, Xinyu Sun, and Peihao Chen. Skdf: a simple knowledge distillation framework for distilling open-vocabulary knowledge to open-world object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 8

[29] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14074–14083, 2022. 2

[30] Sahal Shaji Mullappilly, Abhishek Singh Gehlot, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Hisham Cholakkal. Semi-supervised open-world object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4305–4314, 2024. 7

[31] Jiancheng Pan, Yanxing Liu, Yuqian Fu, Muyuan Ma, Jiahao Li, Danda Pani Paudel, Luc Van Gool, and Xiaomeng Huang. Locate anything on earth: Advancing open-vocabulary object detection for remote sensing community. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6281–6289, 2025. 2

[32] Jinyoung Park, Seongjun Yun, Hyeonjin Park, Jaewoo Kang, Jisu Jeong, Kyung-Min Kim, Jung-Woo Ha, and Hyunwoo J Kim. Deformable graph transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5385–5396, 2025. 3

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini 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. 3, 6

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[35] Nandini Saini, Ashudeep Dubey, Debasis Das, and Chiranjoy Chattopadhyay. Advancing open-set object detection in remote sensing using multimodal large language model. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 451–458, 2025. 3

[36] Hengcan Shi, Munawar Hayat, and Jianfei Cai. Open-vocabulary object detection via scene graph discovery. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4012–4021, 2023. 2

[37] Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, et al. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022. 6

[38] Zhicheng Sun, Jinghan Li, and Yadong Mu. Exploring orthogonality in open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17302–17312, 2024. 2, 7, 8

[39] Yunjie Tian, Qixiang Ye, and David Doermann. YOLOv12: Attention-centric real-time object detectors. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[40] Guanqun Wang, Jianlin Xie, Tong Zhang, Yikang Sun, He Chen, Yin Zhuang, and Jun Li. Llama-unidetector: An llama-based universal framework for open-vocabulary object detection in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–18, 2025. 2

[41] Zhiheng Wu, Yue Lu, Xingyu Chen, Zhengxing Wu, Liwen Kang, and Junzhi Yu. Uc-owod: Unknown-classified open world object detection. In *European conference on computer vision*, pages 193–210. Springer, 2022. 2, 3, 7

[42] Xing Xi, Yangyang Huang, Ronghua Luo, and Yu Qiu. Owovd: Unified open world and open vocabulary object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25454–25464, 2025. 2, 3, 7, 8

[43] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3974–3983, 2018. 6

[44] Muli Yang, Gabriel James Goenawan, Huaiyuan Qin, Kai Han, Xi Peng, Yanhua Yang, and Hongyuan Zhu. Detecting open world objects via partial attribute assignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20318–20328, 2025. 3

[45] Chang-Bin Zhang, Jinhong Ni, Yujie Zhong, and Kai Han. v-clr: View-consistent learning for open-world instance segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20307–20317, 2025. 3

[46] Shan Zhang, Yao Ni, Jinhao Du, Yuan Xue, Philip Torr, Piotr Koniusz, and Anton van den Hengel. Open-world objectness modeling unifies novel object detection. In *Proceedings of*

the Computer Vision and Pattern Recognition Conference, pages 30332–30342, 2025. 7

[47] Qi Zhu, Jiangwei Lao, Deyi Ji, Junwei Luo, Kang Wu, Yingying Zhang, Lixiang Ru, Jian Wang, Jingdong Chen, Ming Yang, et al. Skysense-o: Towards open-world remote sensing interpretation with vision-centric visual-language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14733–14744, 2025. 1, 2, 7

[48] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021. Oral Presentation. 1, 3

[49] Orr Zohar, Kuan-Chieh Wang, and Serena Yeung. Prob: Probabilistic objectness for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11444–11453, 2023. 2, 7, 8