OPEN-VOCABULARY OBJECT DETECTION FOR LOW-ALTITUDE SCENARIOS USING RGB-INFRARED DATA: A BENCHMARK AND A NEW METHOD

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-altitude UAV perception has attracted increasing attention in applications such as urban management and emergency response. However, existing object detection methods mainly rely on predefined categories and struggle to recognize novel objects in open-world environments. Although open-vocabulary object detection (OVD) enables knowledge transfer through vision-language models, its application to low-altitude aerial scenarios remains challenging due to two key issues: **(I)** existing UAV benchmarks are primarily designed for closed-set object detection, lacking multimodal and multi-category evaluation protocols for open-vocabulary learning.**(II)**Existing OVD methods are mainly trained on natural or remote sensing images, whereas low-altitude aerial images exhibit significant differences in object scale, background environment, and imaging conditions, making existing methods difficult to directly adapt to low-altitude object detection. To address above issues, we introduce **UAVOVD**, the first visible–infrared open-vocabulary oriented object detection benchmark for low-altitude scenarios, and introduce a base–novel category split for evaluating open-vocabulary generalization. Based on this benchmark, we propose **LAOVD**, a multimodal open-vocabulary oriented object detection framework equipped with a **Structure-Guided Asymmetric Gated Fusion (SGAGF)** module, which effectively exploits complementary structural and semantic information from infrared and visible modalities. Extensive experiments on UAVOVD, DroneVehicle and ATR-UMOD demonstrate that LAOVD achieves superior performance on novel-category detection and cross-domain generalization, validating the effectiveness of multimodal open-vocabulary learning for low-altitude aerial perception.

1 INTRODUCTION

Table 1: Comparison of existing UAV datasets. OVD denotes whether the dataset supports open-vocabulary object detection.

Dataset	OVD	Modality	Images	Classes	Instances	Resolution	Rotated Box	Year
VEDAIRazakarivony & Jurie (2016)	×	RGB	1.2K	9	41K	1024×1024	×	2016
DOTAXia et al. (2018)	×	RGB	2.8K	15	188K	2000×1500	✓	2018
UAVDTDu et al. (2018)	×	RGB	80K	3	280K	1080×540	×	2018
VisDroneZhu et al. (2018)	×	RGB	10K	10	54K	2000×1500	×	2018
DroneVehicleSun et al. (2022)	×	RGB-IR	56K	5	190K	840×712	✓	2022
RGBT-TinyYing et al. (2025)	×	RGB-IR	93K	7	1.2M	640×512	×	2025
AerialAIDLiu et al. (2025)	×	RGB	9K	60	60K	–	×	2025
ATR-UMODChen et al. (2025)	×	RGB-IR	13K	11	162K	640×512	✓	2025
UAVOVD (Ours)	✓	RGB-IR	25K	88	368K	640×512	✓	2026

Existing UAV object detection methods are typically developed under a closed-set assumption, where training and testing share a fixed set of predefined categoriesRen et al. (2016); Varghese & Sambath (2024); Carion et al. (2020). This limits their ability to recognize novel object categories that emerge in real-world environments. Open-vocabulary detection (OVD) offers a promising way to overcome this limitation by using knowledge from vision–language pretraining to associate vi-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

sual regions with textual semantics Zhu & Chen (2024). However, directly applying existing OVD methods to low-altitude UAV scenarios remains challenging for two main reasons Tian et al. (2025).

First, multimodal benchmarks tailored to low-altitude open-vocabulary detection are lacking. As shown in Table 1, most existing UAV datasets are designed for closed-set detection and contain only a limited number of categories, such as vehicles and pedestrians. They therefore cannot support evaluation of generalization to novel classes. Although several visible–infrared UAV datasets include multimodal information, they still use fixed category vocabularies and lack the semantic category organization and base–novel splits required for open-vocabulary learning. Moreover, few benchmarks jointly support multimodal fusion, oriented object localization, and novel-category recognition, hindering the development of low-altitude perception in open environments. **Second, existing OVD methods are poorly adapted to low-altitude multimodal imagery.** Current OVD methods are primarily developed for natural imagesShao et al. (2019); Lin et al. (2014); Gupta et al. (2019) or remote-sensing imageryPan et al. (2025). Natural images typically contain rich textures and relatively large objects, whereas remote-sensing datasets focus on large-scale ground objects viewed from high altitudes. Both domains differ substantially from low-altitude UAV imagery. In addition, most existing OVD methods use only RGB images and cannot exploit the stable structural cues provided by infrared imagery under low-light, nighttime, or low-contrast conditions. Effectively integrating complementary visible and infrared information while preserving the semantic transferability of open-vocabulary models therefore remains a key challenge.

To address these challenges, we **First** construct **UAVOVD**, the first visible–infrared open-vocabulary object detection benchmark for low-altitude aerial scenarios. The dataset contains 25,292 spatially registered bimodal image pairs, covering 88 object categories and 360k oriented bounding-box instances. A base–novel category split is further introduced to support the evaluation of open-vocabulary generalization. Compared with existing UAV detection datasets, UAVOVD jointly considers multimodal imaging, oriented object localization, and novel-category recognition, providing a unified benchmark for low-altitude open-vocabulary detection. **Second,** we propose **LAOVD** (**L**ow-**A**ltitude **O**pen-**V**ocabulary **D**etection), an OVD framework for visible-infrared low-altitude imagery. To leverage complementary advantages of both modalities, we design a **Structure-Guided Asymmetric Gated Fusion (SGAGF)** module. It takes infrared features as structural priors for stable location, contour and scale cues, and supplements texture and fine-grained semantics from visible features. A cross-modal gating mechanism adaptively regulates information flow between modalities. This asymmetric fusion strategy mitigates illumination and background interference while improving semantic transfer to novel classes.In summary, our main contributions are as follows:

- **Large-scale dual-modal open-vocabulary benchmark:** We construct the UAVOVD dataset, the first open-vocabulary object detection dataset based on spatially registered RGB–infrared image pairs. It fills the data gap for dual-modal open-vocabulary detection in low-altitude UAV scenarios.
- **First dual-modal open-vocabulary detection baseline:** We propose LAOVD, the first model for open-vocabulary object detection in dual-modal aerial scenes. LAOVD effectively fuses complementary information from the two modalities and improves generalization to novel categories.
- **Superior performance:** Extensive experimental results show that LAOVD outperforms existing open-vocabulary object detection methods on base categories and novel categories. Experiments on UAVOVD also demonstrate that the benchmark can effectively evaluate the challenges of low-altitude dual-modal open-vocabulary detection.

2 Related Work

2.1 Low-Altitude UAV Scene Datasets

Low-altitude UAV datasets have been widely developed for object detection and tracking in real-world aerial scenarios. Datasets such as VisDroneZhu et al. (2018),UAVDTDu et al. (2018) and CODroneChen et al. (2026) provide large-scale benchmarks with diverse urban scenes and common object categories, facilitating research on UAV perception under complex environments. With the increasing demand for robust multimodal perception, visible–infrared UAV datasets have also been introduced. For example, DroneVehicleSun et al. (2022) provides large-scale aligned visible–infrared images with oriented bounding-box annotations, while RGBT-TinyYing et al. (2025) and

ATR-UMODChen et al. (2025) further improve diversity in scenes, viewpoints, and imaging conditions.

However, existing UAV datasets are mainly designed for closed-set detection and rely on fixed category vocabularies, limiting their capability for open-world recognition. They generally lack large-scale semantic category descriptions, base–novel evaluation protocols, and unified settings that combine visible–infrared fusion, oriented object localization, and open-vocabulary recognition. To address these limitations, we introduce **UAVOVD**, the first visible–infrared open-vocabulary object detection benchmark for low-altitude scenarios, providing a unified platform for evaluating multimodal open-world UAV perception.

2.2 Open-Vocabulary Object Detection

Open-vocabulary object detection (OVD) aims to extend traditional detectors beyond predefined categories by leveraging vision-language knowledge. Early approaches, such as ViLDGu et al. (2021), RegionCLIPZhong et al. (2022), and DetCLIPYao et al. (2022), explored CLIP-based knowledge transfer and region-text alignment, while Grounding DINOLiu et al. (2024) and recent efficient frameworks including YOLO-WorldCheng et al. (2024) and DOSODHe et al. (2025) further improved open-vocabulary localization and practical deployment. Meanwhile, several studies,such as CastDetLi et al. (2024), LAE-DINOPan et al. (2025), and DescRegZang et al. (2024), have investigated OVD in remote sensing imagery by incorporating vision-language pre-training and aerial-specific representation learning.

Despite these advances, existing OVD methods are primarily developed for natural images or RGB remote sensing imagery and face substantial challenges when transferred to low-altitude UAV scenarios. The unique characteristics of UAV imagery, including small objects, arbitrary orientations, dense layouts, and significant illumination variations, require specialized multimodal modeling. Moreover, existing methods rarely exploit complementary infrared information or address oriented object detection in open-vocabulary settings. In this work, we propose **LAOVD**, a visible–infrared open-vocabulary oriented object detection framework that integrates multimodal fusion and vision-language alignment for robust low-altitude UAV perception.

3 UAVOVD Dataset

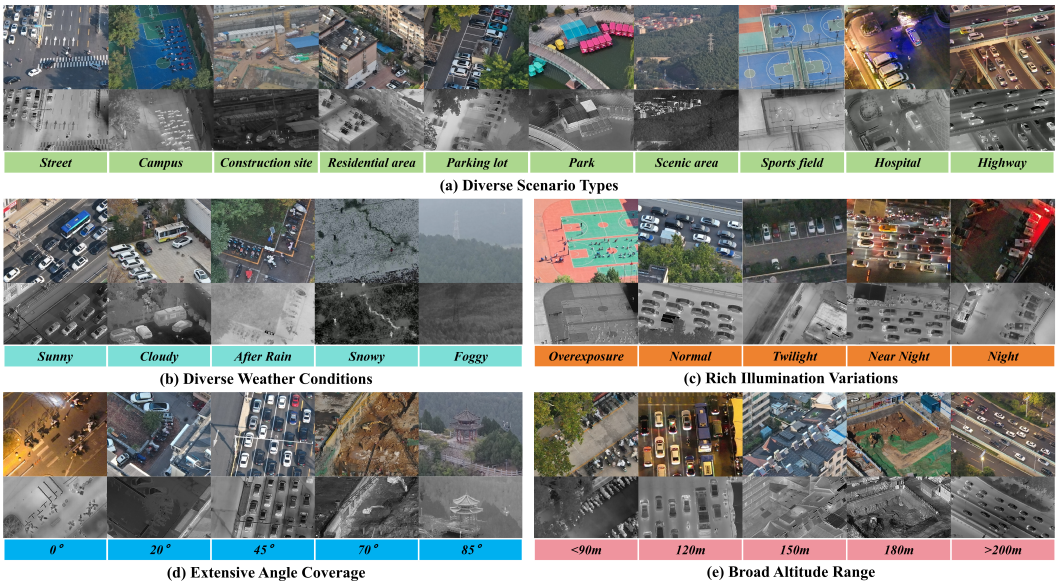

Figure 1: The UAVOVD dataset covers diverse scenes and imaging conditions. (a)–(e) illustrate the variations in scenes, weather conditions, illumination conditions, viewpoints, and flight altitudes, respectively.

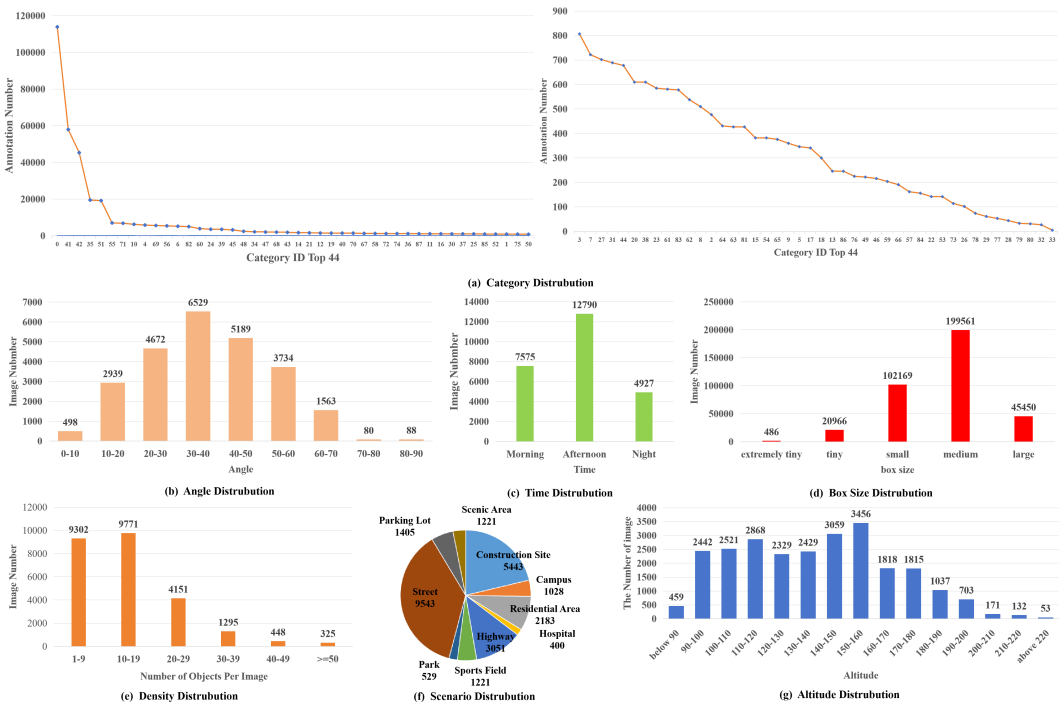

Figure 2: (a)–(g) illustrate the distributions of object categories, viewpoints, capture times, object sizes, object densities, scene types, and flight altitudes in the UAVOVD dataset, respectively.

3.1 Dataset Construction

We construct **UAVOVD**, a visible–infrared open-vocabulary object detection dataset for low-altitude aerial scenarios. The data are collected using a DJI Matrice 350 RTK UAV equipped with a DJI Zenmuse H20N multi-sensor gimbal camera. As shown in Fig. 1, UAVOVD covers diverse low-altitude environments, including urban roads, residential areas, construction sites, parking lots, campuses, and parks, with various flight altitudes, camera viewpoints, and acquisition periods.

To obtain spatially aligned visible–infrared image pairs, we perform cross-modal feature matching followed by homography transformation to compensate for geometric offsets between modalities. All image pairs are cropped and resized to 640×512. For annotation, we adopt a semi-automatic pipeline consisting of manual annotation, model-assisted refinement, and manual verification to ensure annotation quality. More details are provided in the Appendix. A.3.

After quality control, UAVOVD contains 25,292 aligned visible–infrared image pairs, 88 object categories, and 368,632 object instances. Each category is associated with a standardized text label for open-vocabulary learning. Following the open-vocabulary evaluation protocol, the categories are divided into 68 base classes and 20 novel classes. The dataset is split into 15,923 training image pairs with base-class annotations and 9,369 test image pairs with annotations from all categories, enabling systematic evaluation of both base and novel category detection.

3.2 Dataset Statistics

As shown in Fig. 2, UAVOVD exhibits diverse distributions in category, viewpoint, illumination, object scale, density, and scene conditions.

Category distribution. As shown in Fig. 2(a), UAVOVD presents a long-tailed category distribution. The three dominant categories, *passenger car*, *pedestrian*, and *rider*, account for 58.92% of all instances, while rare categories contain substantially fewer samples, making open-vocabulary recognition more challenging.

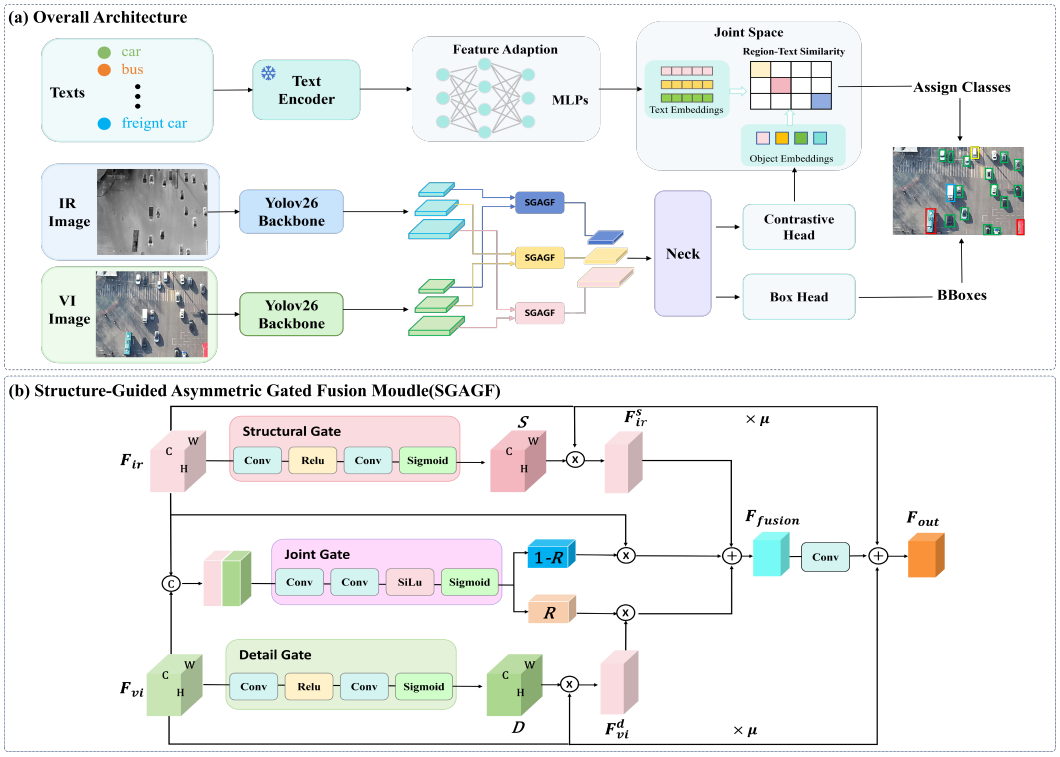

Figure 3: Architecture of the proposed LAOVD framework. (a) shows the overall pipeline, while (b) details the structure of the SGAGF module and its intra-modal and cross-modal gating mechanisms.

Flight condition distribution. As shown in Fig. 2(b) and Fig. 2(g), the dataset covers a wide range of flight altitudes and camera viewpoints. Most samples are collected at altitudes of 90–200 m (96.78%) and camera pitch angles of 20°–70° (85.75%), introducing considerable variations in object scale and appearance.

Acquisition time and object scale. As shown in Fig. 2(c) and Fig. 2(d), UAVOVD contains both daytime and nighttime samples, leading to significant illumination and thermal variations across visible and infrared modalities. Meanwhile, small and medium objects dominate the dataset (81.85%), posing challenges for multi-scale representation and accurate localization.

Object density and scene diversity. As shown in Fig. 2(e) and Fig. 2(f), most image pairs contain fewer than 30 objects, while crowded scenes are also included. The dataset covers 10 representative low-altitude environments, such as streets, construction sites, highways, and residential areas, enabling comprehensive evaluation under diverse real-world conditions.

4 Method

4.1 Task Definition

Given a paired infrared-visible training dataset

$$\mathcal{D} = \left\{ (I_i^{ir}, I_i^{vi}, A_i) \right\}_{i=1}^{N}, \tag{1}$$

where I_i^{ir} and I_i^{vi} denote the i-th pair of spatially aligned infrared and visible images, and A_i represents the region-text annotations. Each annotation is formulated as

$$A_i = \{(r_{ij}, t_{ij})\}_{j=1}^{M_i}, \tag{2}$$

where r_{ij} denotes the rotated bounding box of the j-th object and t_{ij} is the corresponding category text. Let the category space be divided into base and novel categories:

$$\mathcal{C} = \mathcal{C}_{base} \cup \mathcal{C}_{novel}, \quad \mathcal{C}_{base} \cap \mathcal{C}_{novel} = \emptyset. \tag{3}$$

During training, only annotations from $\mathcal{C}_{base}$ are available. During inference, text prompts from both base and novel categories are provided, and the model predicts object bounding boxes together with their category-text matching scores.

4.2 Overall Architecture of LAOVD

As shown in Figure 3(a), the model consists of four core components:

Text semantic encoding: A frozen CLIP text encoder extracts category semantics, and a lightweight MLP adapter maps text features into the joint visual-text space. It preserves the open-vocabulary capability from large-scale pre-training while adapting to the feature distribution of low-altitude targets.

Dual-stream visual feature extraction: Infrared and visible images are fed into two YOLO26 backbones with independent parameters to extract multi-scale features. The infrared branch models thermal radiation structural features of targets, while the visible branch extracts texture and fine-grained semantic features.

Structure-Guided Asymmetric Gated Fusion (SGAGF): The core innovation of this work. Fusion modules are embedded at corresponding scales of the two branches, which dynamically inject visible details with infrared structure as the anchor. Details are elaborated in Section4.3.

Region-text alignment detection head: After cross-scale aggregation via the neck, the fused features are fed into a YOLO26-OBB detection head, which outputs rotated box parameters and region visual embeddings simultaneously. The normalized region visual embeddings compute cosine similarity with text features for open-vocabulary classification.

4.3 Structure-Guided Asymmetric Gated Fusion (SGAGF)

Infrared and visible images have distinct imaging mechanisms and information distributions. Infrared images reflect thermal radiation differences between targets and backgrounds, and maintain stable target responses under low-light, shadow and night conditions, thus are suitable for characterizing structural information such as position, contour and scale. Visible images contain rich color, texture, edge and fine-grained semantic information, but their quality is vulnerable to illumination changes, shadows, glare and complex backgrounds.

Direct feature addition or channel concatenation would introduce noise from both modalities and ignore the functional difference between infrared structure and visible details. Conventional symmetric fusion treats both modalities equally; when the visible modality is severely disturbed, its noise tends to corrupt the stable target structure in the infrared modality.

As shown in Figure 3(b), to address this issue, we propose SGAGF, which forms an asymmetric fusion paradigm of *infrared structure dominated, visible detail supplemented*. The process is divided into three steps.

4.3.1 Intra-Modal Feature Filtering

Two gated branches with independent parameters perform adaptive filtering on each modality. The infrared structural gate enhances target thermal response regions and suppresses background thermal noise, outputting structure-filtered features F_{ir}^{s}. The visible detail gate retains texture and edge information with category discriminability, outputting detail-filtered features F_{vi}^{d}. The process is formulated as:

$$S = \sigma\left(\phi_r\left(F_{\text{ir}}\right)\right), \quad F_{\text{ir}}^{s} = S \odot F_{\text{ir}} \tag{4}$$

$$D = \sigma\left(\phi_v\left(F_{\text{vi}}\right)\right), \quad F_{\text{vi}}^{d} = D \odot F_{\text{vi}} \tag{5}$$

where S and D are the feature importance weight matrices for infrared and visible modalities respectively; $\sigma(\cdot)$ denotes the Sigmoid function; $\odot$ denotes element-wise multiplication. $\phi_r(\cdot)$ and $\phi_v(\cdot)$ share the same network structure but have independent parameters, each composed of two 1×1 convolution layers followed by a SiLU activation.

4.3.2 CROSS-MODAL JOINT RELIABILITY GATING

Single-modal gating can only evaluate feature importance within each modality, and cannot measure the relative reliability across modalities at the same spatial position and channel. We further construct a joint reliability gate. After channel-wise concatenation of original dual-modal features, shared convolutions learn reliability weights R across channel and spatial dimensions:

$$R = \sigma\left[\phi_{\text{joint}}\left(\text{Concat}(F_{\text{vi}}, F_{\text{ir}})\right)\right] \in [0,1]^{B\times C\times H\times W} \tag{6}$$

where $\text{Concat}(\cdot)$ denotes channel concatenation; $\phi_{\text{joint}}(\cdot)$ consists of two 1×1 convolution layers and a SiLU activation, with output channel number C.

Dual-modal features are dynamically fused based on R:

$$F_{\text{fusion}} = F_{\text{ir}}^{s} + R \odot F_{\text{vi}}^{d} + (1-R) \odot F_{\text{ir}} \tag{7}$$

When R is large (high-quality visible imaging), the contribution of visible details is enhanced to improve category discriminability. When R is small (visible modality disturbed by illumination or shadows), unreliable visible information is suppressed and the original infrared compensation path dominates, falling back to stable infrared structure representation.

4.3.3 LOW-INTENSITY RESIDUAL CONNECTION

To alleviate original information loss caused by continuous gating, we introduce a low-intensity residual path to preserve underlying information and improve gradient propagation:

$$F_{\text{base}} = \mu\left(F_{\text{ir}} + F_{\text{vi}}\right) \tag{8}$$

$$F_{\text{out}} = \text{Conv}\left(F_{\text{fusion}}\right) + F_{\text{base}} \tag{9}$$

where μ controls the retention degree of original raw features.

5 EXPERIMENTS

5.1 DATASETS

We train and evaluate models on the self-built UAVOVD dataset to verify detection performance on both base and novel classes. For cross-domain generalization, we perform zero-shot evaluation on DroneVehicle and ATR-UMOD using models trained on UAVOVD, without any training or fine-tuning on target datasets.

UAVOVD Dataset Collected with zoom cameras, UAVOVD presents rich object details and textures. It differs from DroneVehicle and ATR-UMOD in imaging scale and detail level, serving as a test bed for evaluating transfer ability under varying imaging conditions.

DroneVehicle Dataset This large-scale drone visible-infrared vehicle dataset covers day and night scenarios, with over 400k rotated bounding box annotations for 5 categories. We use infrared annotations as unified labels for dual-modal images, and remove white borders and adjust coordinates during preprocessing.

ATR-UMOD Dataset This spatiotemporally aligned bimodal dataset contains 13,353 image pairs of 640×512 resolution, covering 11 categories across various flight altitudes, angles and lighting conditions.

For both cross-domain datasets, annotations are converted to YOLO-OBB format, and infrared annotations are adopted as the unified evaluation benchmark.

5.2 EXPERIMENTAL SETUP

We select DOSOD, YOLO-World v1 and YOLO-World v2 as baselines. All methods are adapted for rotated box detection based on the Ultralytics framework to support the open-vocabulary rotated detection task.

Table 2: Detection results on the UAVOVD dataset.

Modality	Method	Venue	Backbone	Base Classes		Novel Classes	
				mAP50	mAP95	mAP50	mAP95
IR	YOLO-World v1-s	CVPR 2024	YOLOv8-s	0.4354	0.3192	0.1157	0.0841
	YOLO-World v2-s	CVPR 2024	YOLOv8-s	0.4369	0.3253	0.102	0.0740
	DOSOD-s	ICRA 2025	YOLOv8-s	0.4240	0.3066	0.1025	0.074
	YOLO-World v1-m	CVPR 2024	YOLOv8-m	0.4834	0.3645	0.1204	0.0905
	YOLO-World v2-m	CVPR 2024	YOLOv8-m	0.4195	0.3093	0.1285	0.0893
	DOSOD-m	ICRA 2025	YOLOv8-m	0.4247	0.308	0.1119	0.0824
Visible	YOLO-World v1-s	CVPR 2024	YOLOv8-s	0.5307	0.4116	0.1463	0.1134
	YOLO-World v2-s	CVPR 2024	YOLOv8-s	0.5002	0.3878	0.142	0.1105
	DOSOD-s	ICRA 2025	YOLOv8-s	0.5177	0.3986	0.1366	0.1013
	YOLO-World v1-m	CVPR 2024	YOLOv8-m	0.5560	0.4334	<u>0.156</u>	<u>0.124</u>
	YOLO-World v2-m	CVPR 2024	YOLOv8-m	0.5291	0.4103	0.1556	0.1215
	DOSOD-m	ICRA 2025	YOLOv8-m	<u>0.5571</u>	<u>0.4433</u>	0.1466	0.1166
Visible+IR	Ours (small)	–	YOLO26-s	0.4768	0.3678	0.1861	0.1359
	Ours (medium)	–	YOLO26-m	**0.5652**	**0.4495**	**0.189**	**0.145**

Our method takes paired visible-infrared images as joint input, while baselines are trained and evaluated separately on visible and infrared modalities. All models are trained for 100 epochs on UAVOVD on a platform with 4 NVIDIA RTX 5090 GPUs.

5.3 Comparative Experiments

The experimental results are summarized in Table 2. Our method achieves competitive performance on both base and novel classes, with particularly significant improvements on novel class detection. The medium-sized model obtains the best overall performance, improving the novel-class mAP50/mAP95 by 3.30/2.10 percentage points compared with YOLO-World-v1-m and by 4.24/2.84 percentage points compared with DOSOD-m. These results demonstrate the effectiveness of our approach in improving open-vocabulary generalization. Although the small model achieves slightly lower base-class performance than some visible-only baselines, it consistently surpasses them on novel classes, indicating that our method maintains strong novel-category detection capability even with limited model capacity. Furthermore, scaling the model from small to medium mainly benefits base-class recognition, while novel-class performance remains relatively stable, suggesting that open-vocabulary generalization is primarily attributed to the proposed multimodal fusion and semantic alignment rather than model size alone.

Overall, visible-only baselines generally outperform their infrared counterparts, indicating the stronger standalone detection capability of visible imagery in this dataset. We further visualize the detection results to analyze the effectiveness of our method. Detailed visualization comparisons are provided in the Appendix. A.

5.4 Generalization Evaluation

We further evaluate the zero-shot generalization ability of our method on two novel datasets, ATR-UMOD and DroneVehicle, as reported in Table 3. Our method consistently achieves superior performance on both datasets, demonstrating strong transferability across different low-altitude infrared-visible scenarios.

On ATR-UMOD, our medium model achieves the highest training performance and maintains competitive zero-shot results on the test categories, while the small model also outperforms all single-modal baselines on novel classes. Similarly, on DroneVehicle, the proposed visible–infrared fusion framework achieves the best overall generalization performance, surpassing visible-only and infrared-only baselines.

These results indicate that the proposed multimodal fusion and region–text alignment effectively improve cross-domain open-vocabulary recognition, enabling the model to generalize beyond the training distribution.

Table 3: Zero-shot generalization results on ATR-UMOD and DroneVehicle datasets.

Modality	Method	Backbone	ATR-UMOD				DroneVehicle			
			Train Set		Test Set		Train Set		Test Set	
			mAP50	mAP95	mAP50	mAP95	mAP50	mAP95	mAP50	mAP95
IR	YOLO-World v1-s	YOLOv8-s	0.0505	0.0408	0.0643	0.0511	0.1916	0.149	0.2154	0.166
	YOLO-World v2-s	YOLOv8-s	0.0519	0.0422	0.0534	0.0441	0.158	0.125	0.1807	0.1419
	DOSOD-s	YOLOv8-s	0.0475	0.0382	0.0563	0.0462	0.1848	0.1452	0.2146	0.1673
	YOLO-World v1-m	YOLOv8-m	0.0558	0.0456	0.0586	0.048	0.2038	0.1627	0.2256	0.1786
	YOLO-World v2-m	YOLOv8-m	0.0516	0.0421	0.0595	0.0485	0.1522	0.1241	0.1649	0.1321
	DOSOD-m	YOLOv8-m	0.0489	0.0392	0.0595	0.0479	0.1804	0.142	0.208	0.1614
Visible	YOLO-World v1-s	YOLOv8-s	0.0763	0.0494	0.0894	0.0597	0.2036	0.1495	0.2211	0.1581
	YOLO-World v2-s	YOLOv8-s	0.0809	0.0523	0.0829	0.0534	0.2143	0.1575	0.2368	0.1698
	DOSOD-s	YOLOv8-s	0.0712	0.0468	0.0791	0.0529	0.202	0.1483	0.2189	0.1558
	YOLO-World v1-m	YOLOv8-m	0.0743	0.0493	0.0938	0.0601	0.22	0.1611	0.2393	0.1703
	YOLO-World v2-m	YOLOv8-m	0.0785	0.0513	0.0955	0.0611	0.2196	0.1611	0.2341	0.1672
	DOSOD-m	YOLOv8-m	0.0722	0.0484	0.0842	0.0559	0.2031	0.1499	0.2205	0.1584
Visible+IR	Ours (small)	YOLO26-s	0.0866	0.0628	**0.1065**	**0.0787**	0.205	0.1626	0.2287	0.1794
	Ours (medium)	YOLO26-m	**0.1074**	**0.0777**	0.1014	0.0752	**0.2243**	**0.1766**	**0.2401**	**0.1863**

5.5 Ablation Study

Table 4: Ablation study of fusion strategies.

Fusion Strategy	Base Classes		Novel Classes	
	mAP50	mAP95	mAP50	mAP95
Element-wise addition	0.5823	0.4660	0.1686	0.1382
Cross-modal cross-attention	0.5847	0.4735	0.1665	0.1364
Visible-dominant gated fusion	**0.5991**	**0.4839**	0.1687	0.1398
Infrared-dominant gated fusion (SGAGF)	0.5652	0.4495	**0.189**	**0.145**

To verify the impact of dual-modal fusion strategies, we compare four approaches under identical network and training settings: element-wise addition, cross-modal cross-attention, visible-dominant gated fusion, and infrared-dominant gated fusion (SGAGF). As shown in Table 4, visible-dominant fusion achieves better base-class performance, while infrared-dominant SGAGF yields the best novel-class results. This indicates that modality role allocation directly affects the performance trade-off between base and novel classes. Element-wise addition aggregates modalities with fixed equal weights and lacks adaptive filtering. Cross-attention enables information interaction via feature correlation. Asymmetric gated fusion further clarifies information transmission paths by combining intra-modal filtering and inter-modal modulation. The infrared-dominant design preserves infrared representations via an independent gated branch and a direct compensation path, while regulating visible information injection via the joint gate. It reduces the model's sole reliance on visible appearance, which is more conducive to novel class recognition.

Overall, the ablation validates the effectiveness of infrared-dominant asymmetric gating for novel class detection. Aligned with our focus on novel category detection in open-vocabulary settings, we adopt infrared-dominant SGAGF as the final fusion module.

6 Conclusion

In this paper, we introduce UAVOVD, a large-scale visible-infrared open-vocabulary object detection dataset for low-altitude scenarios, featuring diverse conditions, long-tailed categories, and abundant small-medium objects. We further propose LAOVD, a dual-modal open-vocabulary rotated object detector with the SGAGF module for adaptive infrared-visible fusion. Comparative experiments and cross-dataset zero-shot evaluation demonstrate that our method outperforms single-modal baselines in novel class detection and cross-domain generalization. The ablation study further validates the effectiveness of the infrared-dominant fusion strategy. In the future, we will release the dataset and code to support research on multi-modal open-vocabulary detection in low-altitude drone scenarios.

REFERENCES

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Chen Chen, Kangcheng Bin, Ting Hu, Jiahao Qi, Xingyue Liu, Tianpeng Liu, Zhen Liu, Yongxiang Liu, and Ping Zhong. Fusion meets diverse conditions: A high-diversity benchmark and baseline for uav-based multimodal object detection with condition cues. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27958–27967, 2025. doi: 10.1109/ICCV51701.2025.02595.

Pengyu Chen, Tao Ouyang, Ke Luo, Weijie Hong, and Xu Chen. Codrone: Autonomous drone navigation assisted by edge and cloud foundation models. *IEEE Internet of Things Journal*, 13 (4):5593–5609, 2026. doi: 10.1109/JIOT.2025.3648721.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16901–16911. IEEE, 2024.

Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *European conference on computer vision*, pp. 375–391. Springer, 2018.

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

Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5351–5359, 2019. doi: 10.1109/CVPR.2019.00550.

Yonghao He, Hu Su, Haiyong Yu, Cong Yang, Wei Sui, Cong Wang, and Song Liu. A light-weight framework for open-set object detection with decoupled feature alignment in joint space. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13590–13596, 2025. doi: 10.1109/ICRA55743.2025.11127317.

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*, pp. 431–448. Springer, 2024.

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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Junli Liu, Qizhi Chen, Zhigang Wang, Yiwen Tang, Yiting Zhang, Chi Yan, Dong Wang, Xuelong Li, and Bin Zhao. Aerialvg: A challenging benchmark for aerial visual grounding by exploring positional relations. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5177–5187. IEEE, 2025.

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*, pp. 38–55. Springer, 2024.

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*, volume 39, pp. 6281–6289, 2025.

Sebastien Razakarivony and Frederic Jurie. Vehicle detection in aerial imagery : A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34:187–203, 2016. ISSN 1047-3203. doi: https://doi.org/10.1016/j.jvcir.2015.11.002. URL https://www.sciencedirect.com/science/article/pii/S1047320315002187.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8429–8438, 2019. doi: 10.1109/ICCV.2019.00852.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, Oct 2022. ISSN 1558-2205. doi: 10.1109/TCSVT.2022.3168279.

Yonglin Tian, Fei Lin, Yiduo Li, Tengchao Zhang, Qiyao Zhang, Xuan Fu, Jun Huang, Xingyuan Dai, Yutong Wang, Chunwei Tian, et al. Uavs meet llms: Overviews and perspectives towards agentic low-altitude mobility. *Information Fusion*, 122:103158, 2025.

Rejin Varghese and M Sambath. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International conference on advances in data engineering and intelligent computing systems (ADICS)*, pp. 1–6. IEEE, 2024.

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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3974–3983, Los Alamitos, CA, USA, June 2018. IEEE Computer Society. doi: 10.1109/CVPR.2018.00418. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00418.

Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in neural information processing systems*, 35:9125–9138, 2022.

Xinyi Ying, Chao Xiao, Wei An, Ruojing Li, Xu He, Boyang Li, Xu Cao, Zhaoxu Li, Yingqian Wang, Mingyuan Hu, Qingyu Xu, Zaiping Lin, Miao Li, Shilin Zhou, Li Liu, and Weidong Sheng. Visible-thermal tiny object detection: A benchmark dataset and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):6088–6096, 2025. doi: 10.1109/TPAMI.2025.3544621.

Zhengqing Zang, Chenyu Lin, Chenwei Tang, Tao Wang, and Jiancheng Lv. Zero-shot aerial object detection with visual description regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6926–6934, 2024.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16772–16782, 2022. doi: 10.1109/CVPR52688.2022.01629.

Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):8954–8975, 2024.

Peng Fei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. Vision meets drones: A challenge. *ArXiv*, abs/1804.07437, 2018. URL https://api.semanticscholar.org/CorpusID:5048008.

AI USE STATEMENT

No generative AI tools or large language models were used in any stage of this research, including manuscript writing, translation, figure creation, code implementation, dataset processing, experimental analysis and result interpretation. The authors are fully responsible for all content of this paper.

A APPENDIX

A.1 VISUAL COMPARISON

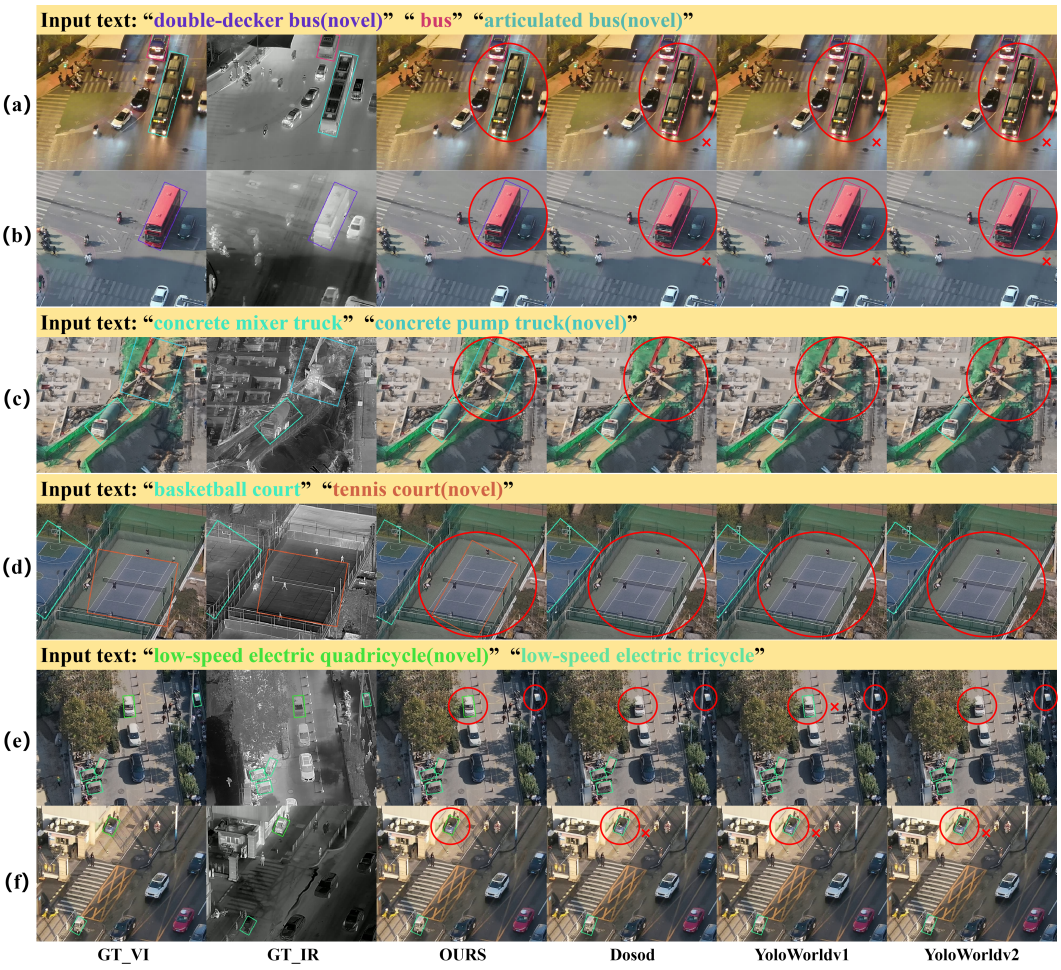

Figure 4: We provide the model with multiple text prompts that include both base and novel classes. Each prompt set contains semantically similar categories, allowing us to assess the model's ability to detect unseen classes. Red circles highlight regions where the methods produce different detection results, while red crosses indicate incorrect detections. Some methods also miss objects. In Figures(a) and (b), the input categories are "double-decker bus (novel)", "bus," and "articulated bus(novel)". These examples test whether the model can distinguish double-decker and articulated bus from bus, despite their similar textual and visual features. Figure(c) evaluates detection of the novel construction vehicle "concrete pump truck" at a construction site, while Figure(d) evaluates detection of the novel category "tennis court" in a sports scene. Figures(e) and (f) show detection results for "low-speed electric quadricycle" and "low-speed electric tricycle" in road scenes, where the two categories are similar in both text and appearance but represent distinct vehicle types.

Qualitative comparison of detection results from different methods is presented in Figure 4, with 6 pairs of sample images.

In Figures 4(a) and 4(b), we use "double-decker bus", "bus" and "articulated bus" as text prompts. Our method can accurately locate and identify novel classes double-decker bus and articulated bus, while other methods can locate the targets but fail to effectively distinguish fine-grained categories, and misclassify them as the base class *bus*. Figure 4(c) shows detection results of concrete mixer truck and concrete pump truck. These two categories have semantically similar names but distinct visual appearances. Our method correctly locates and identifies the novel class concrete pump truck,

while other methods recognize it as the base class concrete mixer truck. In Figure 4(d), with "tennis court" and "basketball court" as text prompts, our method also accurately identifies the novel class tennis court. Figures 4(e) and 4(f) further demonstrate small objects low-speed electric quadricycle and low-speed electric tricycle, which have similar appearances and category names. Our method can effectively distinguish the two categories, while other methods suffer from missed detection or misclassifying novel classes as base classes.

These qualitative results support the advantages of our method in novel class detection and fine-grained category discrimination, consistent with the quantitative results. It verifies that the overall model combining dual-modal input and SGAGF fusion can reduce missed detection and misclassification of novel class objects in these scenarios.

A.2 Category Statistics of UAVOVD

Table 5: Category statistics of the UAVOVD dataset.

ID	Category	Instances	ID	Category	Instances	ID	Category	Instances
0	passenger car	113916	29	forklift	61	58	sunshade canopy	1234
1	gondola lift	903	30	truck crane	1048	59	pop-up canopy	204
2	sightseeing vehicle	477	31	concrete pump truck	689	60	boat	3913
3	police car	807	32	road sweeper	27	61	litter bin	581
4	taxi	5858	33	lawn mower	5	62	wheelie bin	538
5	off-road vehicle	346	34	bicycle	2182	63	U-turn arrow marking	427
6	bus	5203	35	electric bicycle	19504	64	straight-or-left arrow marking	431
7	double-decker bus	722	36	motorcycle	1164	65	straight-or-right arrow marking	376
8	articulated bus	510	37	tricycle	1041	66	speed reduction marking	191
9	minibus	360	38	electric courier tricycle	610	67	left-turn arrow marking	1292
10	van	6281	39	low-speed electric tricycle	3530	68	right-turn arrow marking	1989
11	ambulance	1094	40	low-speed electric quadricycle	1454	69	straight-going arrow marking	5605
12	box truck	1486	41	pedestrian	57970	70	no-parking marking	1444
13	dump truck	246	42	rider	45330	71	pedestrian crossing	6851
14	flatbed truck	1721	43	traffic light	1875	72	basketball court	1194
15	stake-bed truck	382	44	solar street light	678	73	badminton court	114
16	light-duty truck	1077	45	solar water heater	3242	74	basketball goal	1191
17	pickup truck	341	46	solar panel	216	75	table tennis table	889
18	canopy-equipped pickup truck	300	47	standard street light	2044	76	soccer goal	225
19	small water tank trailer	1454	48	bollard light	2448	77	tennis court	53
20	tank truck	610	49	Chinese classic street light	222	78	lattice telecom tower	74
21	concrete mixer truck	1607	50	high-mast light	813	79	monopole telecom tower	33
22	garbage truck	142	51	manhole cover	19181	80	transmission tower	31
23	rotary drilling rig	585	52	traffic sign	925	81	utility pole	427
24	excavator with bucket	3562	53	pedestrian bridge	142	82	outdoor ac unit	5013
25	hammer excavator	1035	54	flag	382	83	bus stop	578
26	hydraulic compactor	102	55	stone bollard	7027	84	bench	156
27	loader	702	56	road barrier	5427	85	traffic surveillance pole	947
28	road roller	44	57	pavilion	162	86	charging pile	246
						87	two-wheeler parking area	1113

Table 6: Novel categories in UAVOVD.

Unseen Classes			
police car	double-decker bus	articulated bus	minibus
box truck	stake-bed truck	light-duty truck	canopy-equipped pickup truck
tank truck	hammer excavator	concrete pump truck	motorcycle
electric courier tricycle	low-speed electric quadricycle	standard street light	straight-or-left arrow marking
straight-or-right arrow marking	left-turn arrow marking	tennis court	monopole telecom tower

A.3 Dataset Construction Details

Data Collection and Preprocessing. We collect the data using a DJI Matrice 350 RTK UAV equipped with a DJI Zenmuse H20N multi-sensor gimbal camera. The device is capable of simultaneously capturing visible and long-wave infrared images during the same flight mission, ensuring consistent acquisition timestamps and similar spatial viewpoints between the two modalities.

Due to the differences in imaging mechanisms, field of view, and sensor mounting positions between visible and infrared cameras, the raw image pairs still contain certain geometric misalignments. To establish accurate cross-modal spatial correspondence, we develop a visible–infrared image registration pipeline. Specifically, synchronized video streams are first decomposed into visible and infrared frame pairs. To increase data diversity, one image pair is sampled every second. Then, cross-modal

feature correspondences between visible and infrared images are extracted using the MINIMA algorithm. Based on the detected correspondences, a homography matrix is estimated through robust geometric estimation with RANSAC, establishing the perspective transformation between the two modalities. Finally, the valid overlapping regions after registration are cropped and resized to obtain spatially aligned visible–infrared image pairs with a resolution of 640×512.

Object Annotation. To balance annotation quality and efficiency, we develop a semi-automatic annotation pipeline consisting of three stages: manual annotation, model-assisted annotation, and manual verification. First, 5,000 visible–infrared image pairs are carefully annotated using the X-AnyLabeling platform. Then, these manually annotated samples are used to fine-tune a pre-trained YOLO26-OBB model, which is subsequently applied to generate candidate annotations for the remaining images. Finally, all model-generated category labels and bounding boxes are manually inspected and refined to correct missing annotations, false detections, category errors, and inaccurate localization results.

After data cleaning and quality control, UAVOVD contains 25,292 spatially aligned visible–infrared image pairs, covering 88 object categories and 368,632 object instances. Each category is associated with a standardized vocabulary-level text label, enabling semantic correspondence between visual targets and category descriptions, and providing a foundation for open-vocabulary object detection research.

A.4 Loss Function

The training objective consists of region-text contrastive loss, rotated box ProbIoU loss, rotated distance L1 loss and angle loss. The total training loss is defined as:

$$L_{\text{total}} = \alpha_{\text{con}} L_{\text{con}} + \alpha_{\text{prob}} L_{\text{prob}} + \alpha_{\text{l1}} L_{\text{l1}} + \alpha_{\text{angle}} L_{\text{angle}} \tag{10}$$

where: L_{con} optimizes semantic alignment between region visual features and category texts; L_{prob} constrains the overall geometric consistency between predicted and ground-truth rotated boxes; L_{l1} refines the regression of rotated boundary distances; L_{angle} improves the stability of target orientation prediction; α_{con}, α_{prob}, α_{l1} and α_{angle} are the balance coefficients of each loss term.