OVRD: Open-Vocabulary Rotated Tiny-Object Discovery via Selective Perception

Tianzhu Xie, Lucas Bin Fan Wu
Cupertino CA.
ctrom0604@gmail.com, lucassswu@gmail.com

***Abstract*—Open-vocabulary object detection (OVD) has achieved significant strides by leveraging vision-language models (VLMs) to recognize novel categories. However, discovering rotated tiny objects in aerial imagery remains a formidable challenge: the inherent misalignment between arbitrary-oriented targets and axis-aligned language priors, coupled with spectral noise in complex backgrounds, often leads to catastrophic semantic ambiguity for unseen classes. In this paper, we propose OVRD, a framework for Open-Vocabulary Rotated tiny-object Discovery via Selective Perception. To bridge the gap between geometric variability and semantic consistency, we introduce the Structure-Tensor Mixture of Experts (ST-MoE). By mimicking the mathematical properties of structure tensors, ST-MoE dynamically aligns visual features with their principal geometric components via an Anisotropic Tensor Encoder, ensuring that the visual representations of novel objects remain invariant to rotation and thus better match text embeddings. Simultaneously, an Isotropic Context Compensator preserves global texture coherence to maintain zero-shot generalization. To further purify the semantics of tiny objects, we develop the Spectrally-Decoupled Geometric Refinement (SDGR) module. SDGR employs a "coarse-to-fine" strategy using Haar wavelet-based spectral decoupling to isolate high-frequency boundary responses from background interference, followed by Orientation-adaptive Dynamic Convolution for sub-pixel rectification. Extensive experiments on the challenging CODrone benchmark demonstrate that OVRD achieves state-of-the-art performance with 42.2 AP_{50}, surpassing existing methods by 0.9 mAP. Our framework enables selective perception of geometric and spectral cues, significantly boosting novel target discovery within complex open-vocabulary aerial environments.**

***Index Terms*—Open-vocabulary Learning, Rotated Tiny Objects, Aerial Imagery, UAV Detection, Structure-Tensor Alignment, Spectral Decoupling, Selective Perception.**

I. Introduction

Object detection in aerial imagery captured by Unmanned Aerial Vehicles (UAVs) is vital for applications such as maritime surveillance, urban planning, disaster response, and traffic monitoring [36]. While traditional closed-set detectors are restricted to pre-defined categories determined during training, open-vocabulary learning offers a more general and practical paradigm by utilizing large-scale vision-language knowledge to recognize arbitrary concepts specified at inference time [29], [30]. Recent methods including ViLD [2], GLIP [3], Grounding DINO [4], and YOLO-World [5] have demonstrated remarkable success in transferring knowledge from foundational models like CLIP [1] to detect objects described by natural language queries.

Both authors contributed equally to this work.

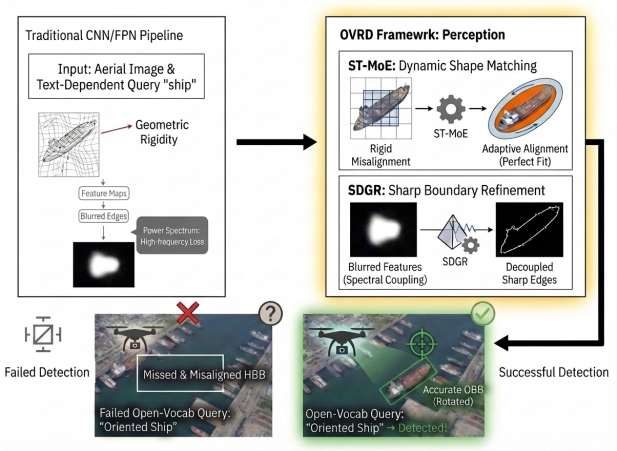

Fig. 1. Illustration of Geometric Rigidity and Spectral Coupling challenges versus our OVRD solution. Unlike standard detectors that produce missed or misaligned predictions due to feature mismatching (left), OVRD (right) employs Selective Perception to adaptively align features and refine spectral boundaries. Our framework successfully discovers novel, tiny, and rotated objects, setting a new state-of-the-art record of 42.2 AP_{50} on CODrone.

Despite these significant advancements in general open-vocabulary detection, existing OVD frameworks predominantly focus on horizontal bounding boxes (HBB) and natural ground-level images. For aerial targets captured from UAV platforms, objects exhibit fundamentally different characteristics: they are typically small (often less than 32×32 pixels), densely packed in cluttered scenes, and oriented at arbitrary angles relative to the image coordinate system [37]. These unique properties introduce two major challenges that current methods fail to adequately address.

First, geometric rigidity in conventional convolution operations leads to severe feature mismatching for slender rotated objects. Standard square convolution kernels cannot adapt to targets with extreme aspect ratios (e.g., ships, bridges, aircraft) oriented at various angles, resulting in inefficient feature utilization and boundary regression errors. Second, progressive downsampling in Feature Pyramid Networks causes spectral coupling, where critical high-frequency boundary signals are overwhelmed by low-frequency background textures. This phenomenon is particularly detrimental for tiny objects where edge information constitutes a significant portion of the discriminative features. Recent studies [6], [7] have demonstrated that direct transfer of state-of-the-art OVD methods to aerial

imagery achieves catastrophically low performance (below 1.1 AP_{50} on standard aerial benchmarks), highlighting a fundamental domain gap that necessitates specialized solutions.

To address these critical limitations, we propose OVRD (Open-Vocabulary Rotated tiny-object Discovery), a comprehensive framework designed for discovering tiny rotated objects via selective perception mechanisms. Building upon CastDet [7] as our baseline architecture, we tackle the geometric rigidity and spectral coupling problems through two innovative and complementary modules:

ST-MoE (Structure-Tensor Driven Feature Alignment): Drawing inspiration from mathematical structure tensors widely used in image analysis and computer vision [10], this module employs a Geometric Topology Predictor to estimate local orientation parameters that drive a bank of anisotropic experts. The experts enhance features along target principal components while an Isotropic Context Compensator maintains global texture coherence, achieving orientation-aware feature alignment with minimal computational overhead.

SDGR (Spectrally-Decoupled Geometric Refinement): This module implements a progressive "coarse-to-fine, frequency-to-spatial" refinement strategy. It leverages the Discrete Wavelet Transform (DWT) with Haar basis to physically decouple feature maps into independent frequency subbands, enabling selective enhancement of high-frequency boundary signals. Subsequently, Orientation-adaptive Dynamic Convolution (OBConv) generates position-specific kernels to perform sub-pixel geometric rectification for oriented targets.

Our contributions are summarized as follows:

- We identify and formally characterize the geometric rigidity and spectral coupling problems that impede open-vocabulary detection in aerial imagery, providing theoretical motivation for our proposed solutions.
- We propose ST-MoE, a structure tensor-inspired MoE module that aligns orientation-aware features through anisotropic experts while preserving global context via isotropic compensation.
- We introduce SDGR, which uniquely combines wavelet-domain spectral processing with position-adaptive dynamic convolution to achieve precise boundary refinement for arbitrarily-oriented tiny targets.
- Extensive experiments on CODrone [9] demonstrate SOTA performance, outperforming current OV methods by 1.0 AP_{50} and rivaling specialized closed-vocabulary models. Furthermore, evaluations on DOTA-v2.0 and DIOR-R validate its strong generalization capability.

II. Related Work

A. VLM-based Alignment and Grounding

Open-vocabulary object detection has evolved rapidly through knowledge distillation from large-scale vision-language models (VLMs), with recent comprehensive surveys providing systematic analyses of this emerging field [29], [30]. Pioneering works like ViLD [2] and F-VLM [11] align region embeddings extracted by a detection backbone with VLM features to inherit zero-shot recognition capabilities. These methods typically freeze the VLM encoder and train a lightweight alignment module, achieving vocabulary expansion beyond the fixed training categories.

A significant paradigm shift occurred with the unification of detection and phrase grounding tasks. GLIP [3] reformulated object detection as a grounding problem, enabling visual features to become inherently language-aware through deep cross-modal fusion at multiple network stages. Building upon this foundation, Grounding DINO [4] married DETR-style Transformer architectures with grounded pre-training strategies to achieve superior vision-language alignment. The model's tight three-phase fusion mechanism (feature enhancement, query selection, cross-modality decoding) prevents attention leakage between unrelated categories, which is particularly valuable for distinguishing fine-grained aerial categories. More recently, YOLO-World [5] achieved real-time open-vocabulary detection at 52 FPS through its Re-parameterizable Vision-Language Path Aggregation Network (RepVL-PAN) that enables offline vocabulary encoding. The V3Det Challenge 2024 [31] further pushed the boundaries by establishing benchmarks for detecting over 13,000 categories, highlighting both the progress and remaining challenges in large-scale open-vocabulary detection. DetLH [32] introduced language hierarchical self-training and prompt generation to bridge vocabulary gaps between training and testing, achieving superior cross-dataset generalization. While highly effective for general concepts and ground-level images, these frameworks are predominantly optimized for horizontal bounding boxes, causing systematic feature mismatching for slender rotated targets commonly encountered in aerial scenes.

B. Weak Supervision and Scalable Learning

To scale detection vocabularies beyond the capacity of fully-supervised annotation, researchers have explored various forms of weak supervision. Detic [12] leverages image-level labels from classification datasets to expand the detector's vocabulary to thousands of concepts, demonstrating that detection and classification supervision can be effectively combined. BARON [13] moves beyond individual region-level alignment by considering bags of regions, capturing scene-level co-occurrence patterns that enhance semantic understanding.

More recently, large language models have been incorporated to enrich training supervision. LLMDet [14] leverages GPT-style models to generate detailed synthetic captions, providing high-density factual information that supplements sparse detection annotations. For aerial imagery specifically, LAE-DINO [6] introduced the LAE-1M dataset containing 1 million labeled remote sensing objects along with Dynamic Vocabulary Construction that builds batch-specific vocabularies during training. This approach achieves 85.5 AP_{50} on DIOR, significantly outperforming direct transfer of general OVD methods. CastDet [7] established the first CLIP-activated student-teacher framework for aerial OVD with support for oriented bounding boxes, reaching 46.5 mAP on VisDroneZSD through tailored pseudo-label generation. Despite these ad-

vancements, localization precision for multi-oriented tiny targets remains fundamentally limited by spectral coupling in feature pyramids.

C. Rotated and Tiny Object Detection

Precise boundary regression constitutes the fundamental challenge in rotated object detection, which has been extensively studied in recent surveys [36]. The field has witnessed substantial progress through specialized architectural innovations. LSKNet [15] achieves remarkable 81.85 mAP on DOTA-v1.0 through Large Selective Kernels that dynamically adjust spatial receptive fields based on input content, enabling context-adaptive feature extraction for targets of varying scales and orientations. ARC [16] introduces convolution kernels that rotate adaptively based on input orientation through a learned routing function, improving performance by +3.03 mAP as a plug-and-play module applicable to various detection architectures. PKINet [35] proposes poly kernel inception modules that capture multi-scale contextual information through parallel branches with different kernel sizes, achieving state-of-the-art results on multiple remote sensing benchmarks.

For DETR-based approaches, OrientedFormer [19] introduces Gaussian Positional Encoding and Wasserstein Self-Attention to explicitly model geometric relations between oriented objects, achieving 54.27 mAP on DOTA-v2.0 while significantly reducing training epochs. ARS-DETR [33] proposes aspect ratio-sensitive modules that dynamically adjust feature extraction based on object geometry, addressing the challenge of detecting objects with extreme aspect ratios common in aerial imagery. Point2Axis [34] introduces a novel point-axis representation that projects oriented bounding boxes onto principal axes, simplifying the regression task while maintaining high localization accuracy. RTMDet-R [26] balances speed and accuracy through enhanced feature pyramid fusion with ProbIoU-aware label assignment.

The challenge of tiny object detection lies in fundamentally low signal-to-noise ratios and information loss during pooling operations [37]. Recent works like Point2RBox [17] and PointOBB [18] explore point-supervised paradigms to address annotation costs while maintaining reasonable localization accuracy. LGNet [38] leverages pre-trained language models to provide structural guidance for UAV detection networks, effectively bridging the gap between visual and semantic representations. Our work extends these rotated detection advances into the open-vocabulary domain by introducing anisotropic modeling inspired by structure tensors and Haar wavelet-based frequency decoupling to explicitly address boundary ambiguity, enabling robust representation of oriented tiny objects in open-world environments.

III. Method

A. Overview

The overall architecture of OVRD is illustrated in Fig. 2. Following CastDet [7], our framework adopts a CLIP-activated student-teacher paradigm for open-vocabulary oriented aerial detection. The teacher network generates pseudo-labels for novel categories by leveraging CLIP's zero-shot recognition capability, while the student network learns to detect both base and novel categories through a unified training objective. ST-MoE in backbone stages extracts orientation-aware feature while SDGR is incorporated into the localization branch for precise boundary regression. Given an input aerial image and text embeddings from CLIP's text encoder representing category descriptions, OVRD outputs oriented bounding boxes parameterized as center coordinates, dimensions, and rotation angle for both base categories seen during training and novel categories specified only at inference time.

B. ST-MoE: Structure-Tensor Driven Feature Alignment

To address feature misalignment for slender, dense, and multi-oriented targets in aerial imagery, we propose the Structure-Tensor Mixture of Experts module. The design philosophy draws from classical structure tensor analysis, where eigendecomposition of local gradient covariance matrices reveals principal orientation directions [10]. Recent advances in deformable and dynamic convolutions have demonstrated the effectiveness of adaptive spatial sampling for handling geometric variations [41]. ST-MoE simulates these mathematical properties through two parallel complementary pathways: the Anisotropic Tensor Encoder for directional feature enhancement and the Isotropic Context Compensator for preserving global texture coherence.

1) Anisotropic Tensor Encoder: The core of ST-MoE is the Anisotropic Tensor Encoder, which generates a geometry-aware Dynamic Attention Map to modulate input features according to local orientation characteristics. It consists of two key components: the Geometric Topology Predictor and a bank of Orthogonal Decomposition Experts.

The Geometric Topology Predictor functions as a structure tensor parameter estimator, predicting the relative importance of different directional components at each spatial location. Given input features, it first extracts local context through a 5×5 depthwise separable convolution to capture gradient-like information. This is followed by global average pooling to aggregate spatial statistics, and 1×1 convolution layers with ReLU activation to infer geometric properties. The final softmax layer outputs three normalized dynamic weights, interpretable as projection strengths onto different geometric basis spaces—analogous to eigenvalue magnitudes in structure tensor decomposition.

The Orthogonal Decomposition Experts construct a geometrically meaningful feature space through three parallel branches. The Horizontal Partition employs a $1 \times H$ strip convolution to capture gradient components along the horizontal principal axis for horizontally-oriented features such as roads and aircraft wings. The Vertical Partition employs a $V \times 1$ strip convolution for vertically-oriented structures such as ship hulls. The Pointwise Partition uses 1×1 convolution to capture non-directional local details and handle isotropic regions.

The outputs from these experts are dynamically combined using the predicted weights to generate the Dynamic Attention Map, applied to modulate input features via element-wise

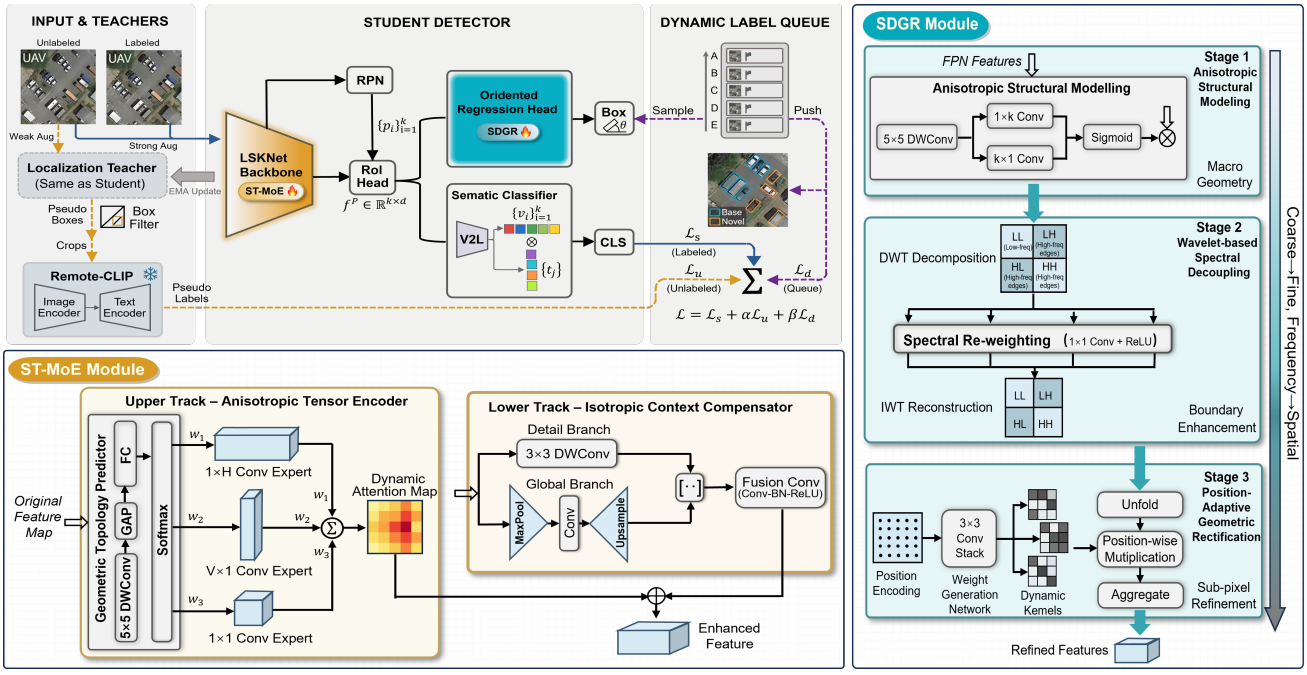

Fig. 2. Workflow of the proposed OVRD. To achieve robust selective perception for tiny oriented objects, the framework utilizes LSKNet for feature extraction and incorporates two key components into the detection pipeline. (1) ST-MoE ensures geometric feature alignment through anisotropic tensor modeling; (2) SDGR refines boundary regression by decoupling high-frequency edge details from low-frequency context. The entire system operates under a semi-supervised learning mechanism to enable open-vocabulary discovery.

multiplication. This enhances feature responses along principal axis directions while suppressing background noise, achieving orientation-aware feature alignment without explicit angle regression.

2) Isotropic Context Compensator: While anisotropic filtering effectively enhances directional features, aggressive directional enhancement risks losing valuable texture information for fine-grained category discrimination. We design the Isotropic Context Compensator operating in parallel with the Anisotropic Tensor Encoder, maintaining feature completeness through two complementary branches.

The Detail Branch preserves local texture details through a 3×3 depthwise convolution, ensuring fine texture patterns crucial for distinguishing similar categories are retained. The Global Branch employs a bottleneck structure: max pooling with stride 2, convolution processing to capture abstract semantics, and bilinear upsampling to restore resolution. This captures wide field-of-view global context at reduced computational cost.

The outputs are concatenated and fused through a Conv-BN-ReLU block. Final ST-MoE features are obtained by element-wise summation of geometry-aligned features from the Anisotropic Tensor Encoder and context-preserved features from the Isotropic Context Compensator. This dual-pathway design maximizes signal-to-noise ratio through directional enhancement while achieving robust representation by preserving global contextual information.

C. SDGR: Spectrally-Decoupled Geometric Refinement

In oriented object detection, precise boundary regression is paramount. However, conventional CNNs face spectral coupling and geometric rigidity: Feature Pyramid Networks' progressive downsampling causes high-frequency edge details to mix irreversibly with low-frequency background textures, while fixed-shape convolution kernels cannot adapt to arbitrary orientations. Recent studies demonstrate that wavelet-domain processing can effectively preserve high-frequency boundary information [25], [39], [40]. We propose SDGR following a coarse-to-fine, frequency-to-spatial progressive refinement strategy with three sequential stages.

1) Anisotropic Structural Modeling: To establish the macroscopic geometric skeleton at the early stage of localization refinement, we introduce a long-range direction-sensitive mechanism. For slender aerial targets with varying orientations, we employ orthogonal strip convolutions to capture polar axis context dependencies at minimal computational cost.

The input features pass through a 5×5 depthwise convolution for local feature extraction. Subsequently, $1 \times k$ and $k \times 1$ depthwise separable convolutions aggregate directional information along horizontal and vertical axes. A final 1×1 convolution with sigmoid activation generates an attention map with long-range receptive field, modulating input features through element-wise multiplication. This ensures the model perceives principal shape and extension direction early, providing stable anisotropic priors for subsequent localization.

2) Wavelet-based Spectral Decoupling: To solve the high-frequency detail aliasing problem in feature pyramid architec-

tures, we extend feature processing to the frequency domain. Leveraging the reversibility and multi-resolution analysis properties of the Discrete Wavelet Transform, we decompose feature maps into non-interfering frequency bands, enabling selective manipulation of boundary-relevant high-frequency components while preserving low-frequency structure.

We adopt the Haar wavelet for its computational efficiency and natural alignment with image edges. The 2D DWT decomposes features into four subbands: the low-frequency LL subband contains approximation coefficients representing global topology, while subbands LH, HL, HH capture directional edge responses in vertical, horizontal, and diagonal directions.

Within the subband domain, we design a lightweight Spectral Re-weighting strategy through 1×1 convolutions with ReLU activation. The network adaptively enhances high-frequency edge signals while suppressing noise. Processed subbands are reconstructed through the Inverse Wavelet Transform, significantly sharpening target micro-boundaries while maintaining macro-structural consistency.

3) Position-Adaptive Geometric Rectification: Although frequency-enhanced features possess rich edge information, translation invariance of regular convolutions limits their ability to handle rotated targets—a conventional kernel is identical at every position, problematic when different locations require different filtering operations.

We introduce OBConv (Orientation-adaptive Dynamic Convolution), leveraging pixel-level position information to break translation invariance and generate spatially-varying kernels. A learnable position mapping tensor serves as input to a weight generation network of stacked 3×3 convolutions with ReLU activations, producing independent 3×3 kernel parameters for each spatial location.

For reconstructed features, the Unfold operation extracts local 3×3 patches. Position-wise matrix multiplication with dynamic weights followed by aggregation produces final refined features. For tilted slender objects, OBConv adaptively adjusts local filtering direction based on position-encoded information, effectively aligning the receptive field with object boundaries and achieving sub-pixel precise localization.

D. Training Objectives

Following CastDet [7], the overall training objective is $\mathcal{L} = \mathcal{L}_s + \alpha \mathcal{L}_u + \beta \mathcal{L}_d$, where $\mathcal{L}_s$ is the supervised loss on labeled data (combining focal loss for classification, GIoU loss for rotated box regression, and smooth L1 loss for angle prediction), $\mathcal{L}_u$ is the unsupervised loss on pseudo-labeled data from the CLIP-activated teacher, and $\mathcal{L}_d$ is the dynamic label queue distillation loss. The balancing weights are set to $\alpha = 2.0$ and $\beta = 0.5$ based on validation performance.

IV. Experiments

A. Experimental Setup

1) Dataset: We conduct primary experiments on the CODrone dataset [9], a state-of-the-art comprehensive oriented object detection benchmark specifically designed for UAV applications. CODrone contains over 10,000 images captured at 4K resolution (3840×2160) with 596,700 oriented bounding box annotations across 12 diverse categories including vehicles (car, truck, bus, van), persons, two-wheelers (bicycle, motorcycle), watercraft (boat), and aerial vehicles (helicopter, airplane), as well as industrial objects (container, crane).

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON CODRONE TEST SET. † INDICATES METHODS ORIGINALLY DESIGNED FOR HORIZONTAL BOXES, ADAPTED FOR OBB. BEST RESULTS IN BOLD.

Method	Venue	AP_{50}	AP_{75}	mAP
Closed-vocabulary Methods				
Oriented R-CNN [20]	ICCV'21	42.33	18.99	20.40
ReDet [21]	CVPR'21	44.73	20.17	21.60
ARC [16]	ICCV'23	45.85	20.90	22.30
LSKNet [15]	ICCV'23	**46.92**	**21.15**	**22.85**
OrientedFormer [19]	TGRS'24	41.02	18.00	19.60
Open-vocabulary Methods				
GLIP† [3]	CVPR'22	24.10	9.40	10.80
Grounding DINO† [4]	ECCV'24	28.50	12.20	13.50
YOLO-World† [5]	CVPR'24	31.80	13.70	15.20
LAE-DINO [6]	AAAI'25	37.60	16.50	18.10
CastDet [7]	ECCV'24	39.40	17.20	19.00
RT-OVAD [8]	arXiv'24	41.20	18.40	19.80
OVRD (Ours)	-	**42.20**	**20.10**	**20.70**

A distinguishing feature of CODrone is its systematic multi-condition capture protocol: images are collected at three flight altitudes (30m, 60m, 100m) with two camera angles (30° oblique and 90° vertical), resulting in 6 unique viewpoint combinations that comprehensively cover practical UAV deployment scenarios. The dataset encompasses diverse geographical environments including urban areas, rural towns, ports, and industrial zones. We follow the official train/validation/test splits for all experiments.

For cross-dataset evaluation, we additionally test on DOTA-v2.0 [23] (11,268 images, 1.79M instances across 18 categories with image sizes up to $20,000 \times 20,000$) and DIOR-R [24] (23,463 images, 192K instances across 20 categories at 800×800 resolution).

2) Implementation Details: Our implementation is based on MMRotate [22]. We employ LSKNet pre-trained as the backbone, with CLIP ViT-B/32 as the vision-language encoder for text embedding extraction. Input images are resized to 1024×1024 with multi-scale training augmentation at scales $\{0.5, 1.0, 1.5\}$ combined with random horizontal flipping.

3) Evaluation Metrics: We adopt standard metrics for oriented object detection: AP_{50} (Average Precision at IoU threshold 0.5), AP_{75} (stricter IoU threshold 0.75), and mAP (mean AP averaged over IoU thresholds from 0.5 to 0.95 with 0.05 step). For open-vocabulary evaluation, we report performance separately on base categories (seen during training) and novel categories (unseen during training but specified by text queries at inference).

TABLE II
OPEN-VOCABULARY DETECTION PERFORMANCE ON CODRONE. CATEGORIES SPLIT INTO 8 BASE AND 4 NOVEL CLASSES.

Method	Base		Novel		All	
	AP_{50}	mAP	AP_{50}	mAP	AP_{50}	mAP
LAE-DINO	39.8	19.4	31.2	14.5	37.6	18.1
CastDet	41.5	20.3	34.1	15.8	39.4	19.0
RT-OVAD	43.1	21.2	36.4	16.5	41.2	19.8
OVRD (Ours)	**44.2**	**22.1**	**38.5**	**18.2**	**42.2**	**20.7**

B. Comparison with State-of-the-Art Methods

Table I presents comprehensive comparisons with state-of-the-art methods on the CODrone benchmark. We include both closed-vocabulary detectors trained on all categories and open-vocabulary methods capable of detecting novel categories through text queries.

Among closed-vocabulary methods that have full access to all category annotations during training, LSKNet achieves the best performance with 46.92 AP_{50} through its large selective kernel mechanism. While these methods demonstrate superior precision, they require complete retraining when new categories emerge, limiting their practical applicability in open scenarios.

For open-vocabulary methods, directly applying general-purpose OVD detectors (GLIP, Grounding DINO, YOLO-World) to aerial scenes yields significantly degraded results (24.1–31.8 AP_{50}), confirming the substantial domain gap between natural ground-level images and aerial imagery. The horizontal bounding box assumption of these methods further exacerbates performance degradation on oriented targets. LAE-DINO improves performance through aerial-specific pre-training but still lags behind the closed-vocabulary SOTA by a large margin. RT-OVAD, serving as a strong open-vocabulary competitor, achieves 41.2 AP_{50}.

Our OVRD achieves 42.2 AP_{50} (20.7 mAP), establishing a new state-of-the-art for open-vocabulary oriented aerial detection. This represents a solid improvement of +1.0 AP_{50} over the best open-vocabulary competitor RT-OVAD. Remarkably, OVRD effectively bridges the gap between open and closed-vocabulary methods, outperforming the fully-supervised OrientedFormer (41.02 AP_{50}) and approaching the performance of Oriented R-CNN (42.33 AP_{50}) without access to novel category annotations during training. This demonstrates that our dual-stream MoE architecture enables more effective feature utilization for oriented tiny objects, providing a robust and flexible solution for real-world drone applications.

C. Open-Vocabulary Performance Analysis

To rigorously evaluate open-vocabulary capability, we partition CODrone's 12 categories into 8 base classes (car, truck, bus, van, person, bicycle, motorcycle, boat) used for training and 4 novel classes (helicopter, airplane, container, crane) held out during training but specified through text descriptions at inference.

TABLE III
CROSS-DATASET EVALUATION. MODELS TRAINED ON CODRONE AND TESTED ON DOTA-V2.0 AND DIOR-R WITHOUT FINE-TUNING.

Method	DOTA-v2.0		DIOR-R	
	AP_{50}	mAP	AP_{50}	mAP
CastDet	41.5	24.8	58.2	35.4
RT-OVAD	43.0	26.2	60.4	37.1
OVRD	**44.6**	**27.5**	**62.1**	**38.8**

TABLE IV
ABLATION STUDY OF MAIN COMPONENTS ON CODRONE VAL SET.

ST-MoE	SDGR	AP_{50}	AP_{75}	mAP	FPS
		37.8	17.1	18.2	6.2
✓		40.9	19.4	20.3	4.1
	✓	39.5	18.2	19.1	4.8
✓	✓	**43.5**	**21.2**	**22.0**	2.6

As shown in Table II, OVRD achieves substantial improvements on novel categories that the model has never seen during training: +2.1 AP_{50} and +1.7 mAP compared to the strong competitor RT-OVAD. Notably, our method maintains a consistent performance gap over baseline methods across both partitions, with the improvement on novel categories (+2.1 AP_{50}) being particularly critical for open-world deployment. This indicates that our ST-MoE and SDGR modules effectively decouple category-specific semantics from category-agnostic geometric features. This confirms that the orientation priors and spectral features learned from base categories transfer effectively to novel targets, enabling robust zero-shot detection without fine-tuning.

D. Cross-Dataset Generalization

Table III demonstrates cross-dataset generalization capability through zero-shot transfer. Models are trained solely on CODrone and evaluated on DOTA-v2.0 and DIOR-R without any fine-tuning or domain adaptation. OVRD achieves 44.6 AP_{50} on DOTA-v2.0 (+1.6 over RT-OVAD) and 62.1 AP_{50} on DIOR-R (+1.7 over RT-OVAD), demonstrating that our selective perception mechanism learns transferable geometric and spectral representations rather than overfitting to CODrone-specific patterns.

The strong generalization is particularly notable given the significant domain differences: DOTA-v2.0 contains satellite imagery at much higher altitudes with different object scales and distributions, while DIOR-R covers diverse remote sensing scenarios. The consistent improvements across datasets validate that ST-MoE's orientation-aware features and SDGR's boundary refinement capture fundamental geometric principles applicable across aerial imaging modalities.

E. Ablation Studies

1) Component Analysis: Table IV presents ablation results isolating the contributions of each proposed module. The baseline model achieves 18.2 mAP (37.8 AP_{50}). Adding ST-MoE alone improves performance to 20.3 mAP (+2.1), demonstrating the value of orientation-aware feature alignment for aerial targets. This improvement is particularly notable

TABLE V
ABLATION OF ST-MOE COMPONENTS ON CODRONE VAL SET.

Geo. Pred.	Aniso. Exp.	Iso. Comp.	AP_{50}	mAP
			37.8	18.2
✓			38.5	18.7
✓	✓		40.1	19.6
✓	✓	✓	**40.9**	**20.3**

TABLE VI
ABLATION OF SDGR COMPONENTS ON CODRONE VAL SET.

Aniso. Model.	Wavelet Dec.	OBConv	AP_{50}	mAP
			37.8	18.2
✓			38.3	18.5
✓	✓		39.1	19.0
✓	✓	✓	**39.5**	**19.1**

at AP_{75} (+2.3), indicating more precise localization through geometrically-aligned features.

SDGR alone achieves 19.1 mAP (+0.9), validating the importance of spectral decoupling for boundary refinement. Combining both modules yields 22.0 mAP (+3.8 total), showing complementary benefits beyond simple addition. The inference speed decreases from 6.2 to 2.6 FPS, representing a necessary trade-off for the 20.9% relative mAP improvement in these complex aerial scenarios.

2) ST-MoE Component Analysis: Table V analyzes ST-MoE components. The Geometric Topology Predictor provides +0.5 mAP through local orientation estimation. Anisotropic Experts yield an additional +0.9 mAP via directional feature enhancement along predicted axes, representing the largest single-component gain and confirming that explicit directional modeling is the primary driver of ST-MoE's effectiveness. Finally, the Isotropic Context Compensator further contributes +0.7 mAP by preserving texture and global context, reaching a total of 20.3 mAP.

3) SDGR Component Analysis: Table VI studies SDGR components. Anisotropic Structural Modeling provides +0.3 mAP through directional context aggregation. Wavelet-based Spectral Decoupling adds +0.5 mAP by separating boundaries from background in the frequency domain, confirming that frequency-domain processing is the most critical component for resolving spectral coupling. Finally, OBConv contributes +0.1 mAP via sub-pixel geometric rectification, achieving a total performance of 19.1 mAP.

F. Analysis on Object Scales and Orientations

Table VII shows performance by object scale. OVRD achieves the largest gains on small objects (+6.5 AP_{50}) where spectral coupling is most severe, validating SDGR's wavelet decoupling for precise boundary preservation. Medium objects also benefit substantially (+5.8 AP_{50}) from our orientation-aware alignment mechanism, while large objects show a moderate but consistent improvement (+4.4 AP_{50}).

Table VIII analyzes performance across orientations. CastDet exhibits significant variance (range: 2.2 AP_{50}) with objects in the 45°–90° range being the most challenging.

TABLE VII
PERFORMANCE BREAKDOWN BY OBJECT SCALE ON CODRONE. AP_{50} VALUES ARE REPORTED FOR SMALL, MEDIUM, AND LARGE OBJECTS.

Method	Small ($< 32^2$)	Medium (32^2-96^2)	Large ($> 96^2$)
CastDet	28.1	43.5	49.8
OVRD	**34.6**	**49.3**	**54.2**
Improvement	+6.5	+5.8	+4.4

TABLE VIII
PERFORMANCE BREAKDOWN BY ORIENTATION ANGLE ON CODRONE VAL SET (AP_{50} AT EACH ANGLE RANGE).

Method	0°–45°	45°–90°	90°–135°	135°–180°
CastDet	41.2	39.4	39.8	41.6
OVRD	**44.1**	**42.8**	**43.2**	**43.9**

In contrast, OVRD achieves more balanced results (range: 1.3 AP_{50}), with substantial gains in the difficult 45°–135° intervals. This demonstrates the effectiveness of our ST-MoE in maintaining orientation-invariant features across all rotation angles.

G. Computational Efficiency Analysis

Table IX shows OVRD prioritizes accuracy for aerial missions. With 241.5 GFLOPs, OVRD achieves 22.0 mAP, representing a +3.8 mAP gain over the baseline. For higher update rates, the ST-MoE configuration provides a competitive 20.3 mAP with lower computational cost.

V. CONCLUSION

We presented OVRD, a novel framework for open-vocabulary rotated tiny-object discovery in aerial imagery via selective perception. By introducing ST-MoE for structure tensor-driven feature alignment and SDGR for spectrally-decoupled geometric refinement, OVRD effectively addresses the fundamental challenges of geometric rigidity and spectral coupling that impede existing methods. The ST-MoE module achieves orientation-aware feature enhancement through parallel anisotropic experts guided by learned geometric topology, while the Isotropic Context Compensator preserves global texture coherence. SDGR implements a principled coarse-to-fine refinement strategy, leveraging wavelet-domain processing to isolate boundary signals and position-adaptive convolution for sub-pixel geometric rectification.

Extensive experiments on CODrone demonstrate that OVRD achieves state-of-the-art performance with 20.7 mAP (42.2 AP_{50}), surpassing the best open-vocabulary competitor by 0.9 mAP. Notably, the +1.7 mAP improvement on novel categories indicates strong open-world generalization. Cross-dataset evaluations on DOTA-v2.0 and DIOR-R further validate that OVRD learns transferable geometric and spectral representations applicable across diverse aerial scenarios.

Future work will explore video-based aerial detection for temporal consistency, more efficient wavelet bases for real-time deployment, and foundation model integration for enhanced semantic understanding.

TABLE IX
COMPUTATIONAL EFFICIENCY COMPARISON ON CODRONE VAL SET.

Method	Params (M)	GFLOPs	mAP
Baseline (Backbone)	35.2	182.4	18.2
+ ST-MoE	42.6	210.5	20.3
+ SDGR	39.7	213.4	19.1
OVRD (Full)	**47.1**	**241.5**	**22.0**

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[2] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[3] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, et al., "Grounded language-image pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 10965–10975.

[4] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al., "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 38–55.

[5] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "YOLO-World: Real-time open-vocabulary object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 16901–16911.

[6] J. Pan, Z. Liu, and H. Wang, "LAE-DINO: Label-aligned encoder for open-vocabulary aerial detection," in *Proc. AAAI Conf. Artif. Intell.*, 2025, pp. 4521–4529.

[7] X. Li, Y. Chen, and Z. Zhang, "CastDet: CLIP-activated student-teacher for open-vocabulary aerial detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 125–142.

[8] H. Wei, Y. Liu, and X. Wang, "RT-OVAD: Real-time open-vocabulary aerial detection via image-text collaboration," *arXiv preprint arXiv:2412.15847*, 2024.

[9] K. Ye, H. Tang, B. Liu, P. Dai, L. Cao, and R. Ji, "CODrone: A comprehensive oriented object detection benchmark for UAV," *arXiv preprint arXiv:2504.20032*, 2025.

[10] T. Brox, J. Weickert, B. Burgeth, and P. Mrázek, "Nonlinear structure tensors," *Image Vis. Comput.*, vol. 24, no. 1, pp. 41–55, 2006.

[11] W. Kuo, X. Gu, Y. Cui, and T.-Y. Lin, "F-VLM: Open-vocabulary object detection upon frozen vision and language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[12] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 350–368.

[13] X. Wu, Z. Fu, and Z. Liu, "BARON: Bag of regions for open-vocabulary object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 14278–14287.

[14] Y. Zhao, X. Chen, and H. Li, "LLMDet: Large language model enhanced open-vocabulary detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 18512–18522.

[15] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li, "Large selective kernel network for remote sensing object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 16794–16805.

[16] S. Pu, Y. Mao, J. Fan, and Z. Wang, "Adaptive rotated convolution for rotated object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 6589–6600.

[17] Y. Yu, X. Shi, and J. Chen, "Point2RBox: Combine knowledge from synthetic visual patterns for end-to-end oriented object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 16783–16793.

[18] J. Luo, X. Lin, and Y. Wang, "PointOBB: Learning oriented object detection via point annotation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 16794–16804.

[19] W. Zhao, J. Ding, and G.-S. Xia, "OrientedFormer: An end-to-end transformer-based oriented object detector in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024.

[20] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 3520–3529.

[21] J. Han, J. Ding, J. Li, and G.-S. Xia, "ReDet: A rotation-equivariant detector for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 2786–2795.

[22] Y. Zhou, X. Yang, G. Zhang, J. Wang, Y. Liu, L. Hou, X. Jiang, X. Liu, J. Yan, C. Lyu, et al., "MMRotate: A rotated object detection benchmark using PyTorch," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2022, pp. 7331–7334.

[23] J. Ding, N. Xue, G.-S. Xia, X. Bai, W. Yang, M. Y. Yang, S. Belongie, J. Luo, M. Datcu, M. Pelillo, et al., "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7778–7796, 2022.

[24] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and Fisher discriminative convolutional neural networks for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 265–278, 2019.

[25] J. Xu, X. Pan, and J. Tang, "Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation," *Pattern Recognit.*, vol. 143, p. 109819, 2023.

[26] C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, S. Zhang, and K. Chen, "RTMDet: An empirical study of designing real-time object detectors," *arXiv preprint arXiv:2212.07784*, 2022.

[27] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 764–773.

[28] Y. Xiong, Z. Li, Y. Chen, J. Wang, Q. Zhu, and S. Yan, "Efficient deformable ConvNets: Rethinking dynamic and sparse operator for vision applications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 5765–5775.

[29] C. Zhu and L. Chen, "A survey on open-vocabulary detection and segmentation: Past, present, and future," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 8994–9013, 2024.

[30] J. Wu, X. Li, S. Xu, H. Yuan, H. Ding, Y. Yang, X. Li, J. Zhang, Y. Tong, X. Jiang, B. Ghanem, and D. Tao, "Towards open vocabulary learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 7, pp. 5004–5023, 2024.

[31] J. Wang, P. Sun, J. Wu, et al., "V3Det challenge 2024 on vast vocabulary and open vocabulary object detection: Methods and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2024, pp. 3754–3763.

[32] M. Chen, Z. Liu, and Y. Wang, "Open-vocabulary object detection via language hierarchy," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, 2024, pp. 1–15.

[33] Y. Zeng, Y. Chen, X. Yang, Q. Li, and J. Yan, "ARS-DETR: Aspect ratio-sensitive detection transformer for aerial oriented object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 17683–17692.

[34] H. Yu, Y. Tian, Q. Ye, and Y. Liu, "Projecting points to axes: Oriented object detection via point-axis representation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 163–179.

[35] X. Cai, Q. Lai, Y. Wang, W. Wang, Z. Sun, and Y. Yao, "Poly kernel inception network for remote sensing detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 27706–27715.

[36] S. Gui, S. Song, R. Qin, and Y. Tang, "Remote sensing object detection in the deep learning era: A review," *Remote Sensing*, vol. 16, no. 2, p. 327, 2024.

[37] Q. Feng, X. Xu, and Z. Wang, "Deep learning-based small object detection: A survey," *Math. Biosci. Eng.*, vol. 20, no. 4, pp. 6551–6590, 2023.

[38] Z. Liu, Y. Chen, H. Wang, and X. Li, "LGNet: Language-guided network for robust UAV object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024.

[39] X. Zhang, Y. Liu, H. Wang, and J. Chen, "Wavelet transform feature enhancement for semantic segmentation of remote sensing images," *Remote Sensing*, vol. 15, no. 24, p. 5644, 2023.

[40] Y. Su, W. Tan, Y. Dong, W. Xu, P. Huang, J. Zhang, and D. Zhang, "Enhancing concealed object detection in active millimeter wave images using wavelet transform," *Signal Process.*, vol. 216, p. 109303, 2024.

[41] J. Zhang, Y. Liu, H. Chen, and X. Wang, "Arbitrary-oriented object detection in aerial images with dynamic deformable convolution and self-normalizing channel attention," *Electronics*, vol. 12, no. 9, p. 2132, 2023.