KARYA ILHAM PUBLISHING
CONNECTING MINDS, BRIDGING IDEAS, INSPIRING READERS

Journal of Advanced Research in Computing and Applications

Journal homepage:
https://karyailham.com.my/index.php/arca/index
ISSN: 2462-1927

Lightweight Open-Vocabulary Object Detection: A Systematic Taxonomy, Comparative Evaluation, and Future Outlook

Zainal Rasyid Mahayuddin[1,*], Liu Youlin[1], Mohammad Faidzul Nasrudin[1]

[1] Center for Artiffcial Intelligence Technology (CAIT), Universiti Kebangsaan Malaysia, 43600 Bangi Selangor, Malaysia

ARTICLE INFO

Article history:
Received 12 October 2025
Received in revised form 2 November 2025
Accepted 21 November 2025
Available online 5 December 2025

ABSTRACT

Open Vocabulary Object Detection (OVD) aims to break through the limitations of traditional detectors. Traditional detectors can only rely on a fixed set of categories. Although vision-language models like CLIP offer zero-shot recognition capabilities, there are still many problems in transferring their global semantics to regional-level detection. These problems include inaccurate spatial positioning, semantic bias and high computational overhead, etc., which greatly affect the actual deployment. This review, with the core perspective of "lightweight and deployable", systematically organized 46 OVD studies and classified the existing methods into three categories: false label learning, knowledge distillation, and architecture optimization. These three types of methods were compared under a unified evaluation setting, and the trade-offs among annotation cost, model compression and inference speed were analyzed. Afterwards, we summarized the key challenges that the OVD field still face, including semantic bias, weak ability to detect small targets, and insufficient cross-domain generalization. Finally, we discussed several new trends like dynamic prompts of Large Language Models (LLMs), adaptive distillation, and collaboration with Segment Anything Model (SAM), etc. It is hoped that this can provide a clear reference framework and research direction for building scalable and resource-friendly OVD systems.

Keywords:
Open-Vocabulary Object Detection; lightweight deployment; knowledge distillation; Pseudo-label learning; lightweight architecture

1. Introduction

Object detection technology has developed rapidly in recent years. This technology, moving from the foundational work of Fast R-CNN [1] and Faster R-CNN's RPN [2] toward highly efficient end-to-end systems like YOLO [3], has significantly improved the detection speed and accuracy [4]. But these models rely on fixed category vocabularies. When they encounter new classes or specific targets, they need to be re-labeled and retrained. This process is very time-consuming and memory-consuming, and it is also difficult to scale, for example, category updates are expensive in autonomous driving [5], and long training and maintenance procedures pose a challenge within security systems [6].

[*] *Corresponding author.*
E-mail address: zainalr@ukm.edu.my

https://doi.org/10.37934/arca.40.1.209222

This problem becomes particularly evident when it comes to professional scenarios such as remote sensing and unmanned aerial vehicles. Due to the large differences in target size, numerous occlusions, and complex backgrounds, the performance of the model on categories that do not appear will decline rapidly [7,8]. To solve this difficult problem, researchers proposed Open Vocabulary Object Detection (OVD), with the aim of enabling the detector to recognize new concepts that have never appeared in the training set [9].

With the rise of Vision–Language Models (VLMs), the function of learning semantics in the joint space of images and text has been realized, providing a new foundation for open vocabulary detection. Gan *et al*., [10] analyzed the cross-scene generalization behavior of VLM. However, VLM generally has strong semantics but weak spatial perception [11]. If embedding it directly into the detection architecture, it will cause positioning errors and performance degradation. The core issue in that field then becomes how to reduce the gap between semantic understanding and spatial positioning. Lightweight object detection has long explored structural simplification, for example, by reducing the structural complexity of the detector through a proposal-free fully convolutional framework, an early architectural reference for lightweight object detection is provided [12]. A large number of lightweight OVD studies aim to achieve low computational overhead and high deployment efficiency while maintaining semantic generalization, main including:

1. Pseudo-label learning: Utilize CLIP [13] to generate automatic annotations to reduce labor costs, including improving the quality of region pseudo-labels [14] and improving region prompting strategies [15].
2. Knowledge distillation: Transferring the semantic knowledge of large models to small detectors, such as cluster-instance discriminative distillation [16] and semantic transfer methods that enhance anomaly scenarios [17].
3. Efficient structural design: Model compression and architecture optimization to achieve real-time detection. For example, optimizing the Detection Transformer (DETR) architecture to handle small targets [18], lightweight mobile design [19], and OVD frameworks that do not require region proposals [20].

Representative methods such as RegionCLIP [14], CORA [15], OWL-ViT [21], and CASTDet [22] mark the gradual shift of research from large-scale semantic alignment to efficient and deployable systems. Figure 1 shows challenges and directions of OVD.

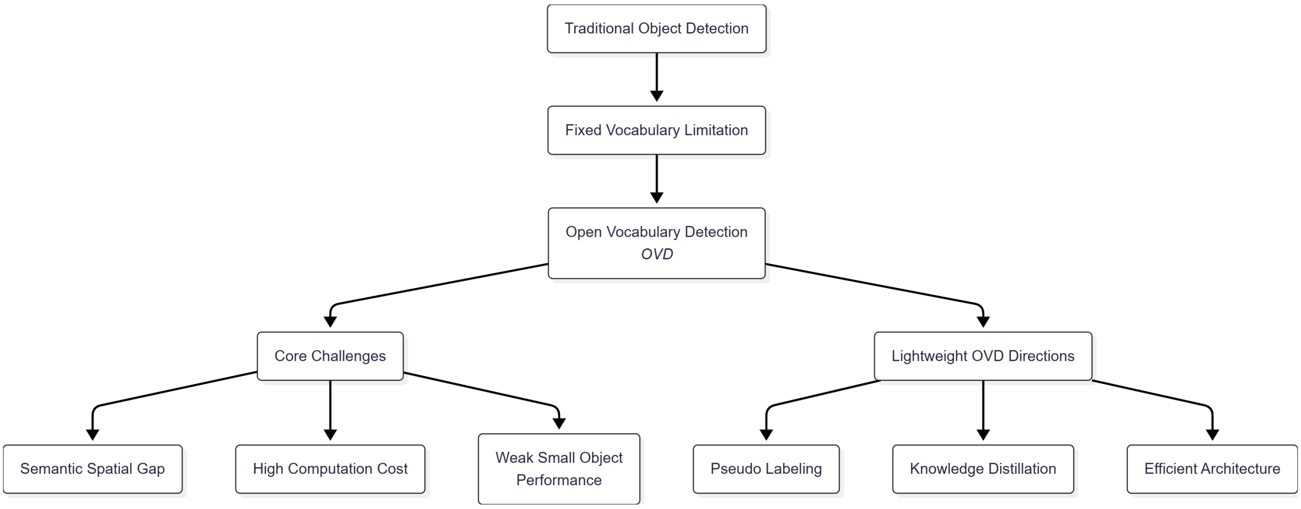
Traditional Object Detection
Fixed Vocabulary Limitation
Open Vocabulary Detection
OVD
Core Challenges
Lightweight OVD Directions
Semantic Spatial Gap
High Computation Cost
Weak Small Object Performance
Pseudo Labeling
Knowledge Distillation
Efficient Architecture

Fig. 1. Lightweight OVD challenges and directions

However, most of the existing reviews only focus on a single dimension. For instance, some only discuss semantic transfer, while others only focus on lightweight implementation. They lack a unified framework to systematically analyze the balance relationship of "accuracy - scalability - efficiency". This has created a clear research gap: we still lack a systematic understanding. This understanding is about the balance between semantic generalization, inference speed and deployment cost of lightweight OVD. Based on this situation, the goal of this article is to sort out the main methods in this direction. It needs to clarify the capabilities and costs of different technical routes. It needs to build a reusable comparison and evaluation framework. Such an analysis not only helps everyone grasp the overall context of the field, but also provides a reference for how to design deployable OVD models in actual systems.

1. Propose a systematic method classification: Summarize the algorithm from five aspects: pseudo-labeling, distillation, structural optimization, semantic enhancement, and hybrid strategies.
2. Conduct cross-benchmark comparisons: Uniformly analyze the performance of the algorithm in terms of accuracy, inference speed, and model size.
3. Constructing a general process and evaluation matrix: This paper summarizes the general process, representative algorithms, key improvement directions of lightweight OVD, and generalizes the relationship between datasets, tasks, and metrics.

This review provides a clear perspective for the understanding of the overall relationship between "semantics - efficiency - scalability", and a reference and direction for building scalable and deployable open vocabulary awareness systems.

2. Survey Methodology

To ensure that the collection and classification process of literature on lightweight open vocabulary object detection is scientific and transparent, this paper presents a systematic and reproducible research process. We retrieved papers from major academic databases (IEEE Xplore, ACM Digital Library, SpringerLink and arXiv) between 2019 and 2025. And use the keyword combinations "open-vocabulary detection", "VLMs" and "lightweight architectures". Among the initially identified 625 studies, after undergoing operations such as removing duplicates and eliminating irrelevant ones, 46 representative studies were screened out and included in the review analysis. The overall process is shown in Figure 2.

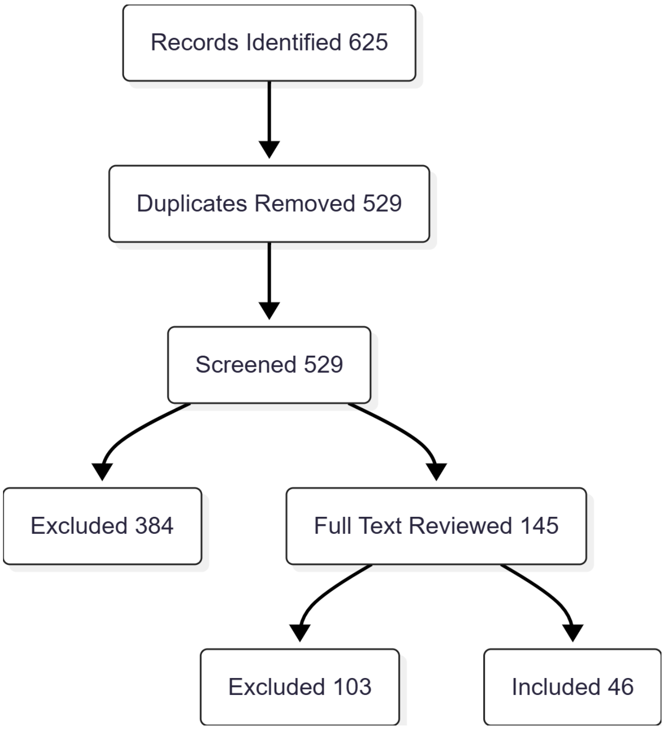

Fig. 2. Literature selection procedure (PRISMA style)

In addition, Fig.2 presents the development trends in this field in detail by year and research type. It can be seen that since 2021, related research has grown rapidly, especially in the directions of lightweighting and semantic distillation. After classification and comparison based on a unified technical paradigm, this paper constructs five types of methodological frameworks ：

1. Pseudo-label learning: Utilize vision-language models to generate automatic labels to reduce manual annotation.
2. Knowledge distillation: Transferring semantic knowledge from large-scale multimodal models to lightweight detectors.
3. Architecture optimization: Real-time inference is achieved through model pruning, re-parameterization, and accelerated structure.
4. Semantic enhancement: Utilize prompt learning or attribute modeling to improve semantic generalization ability.
5. Hybrid strategy: Integrating multiple technologies to balance accuracy and efficiency.

Based on systematic selection and classification methods, this paper can compare the performance and cost of different research methods under a unified standard.

3. Method Pattern and Evolution of Lightweight Open Vocabulary Target Detection

Research on open vocabulary object detection initially focused on openness, and then gradually evolved into lightweighting, that is, enabling the detector to identify categories that do not appear in the training set. The initial research was almost entirely dependent on large-scale visual-language pre-trained models, such as CLIP [13], ALIGN [23] or Florence [9]. These models, with their vast semantic space of cross-modal alignment, enable the detection task to infer visual concepts from the language level for the first time. But models need to have a huge scale of parameters, complex reasoning structures and a high dependence on computing resources. The truth is that models can understand, but the cost of running it is high, it's hard for them to be available at any time in the real system. Figure 3 shows the methods of OVD. Table 1 lists some methods and core mechanisms.

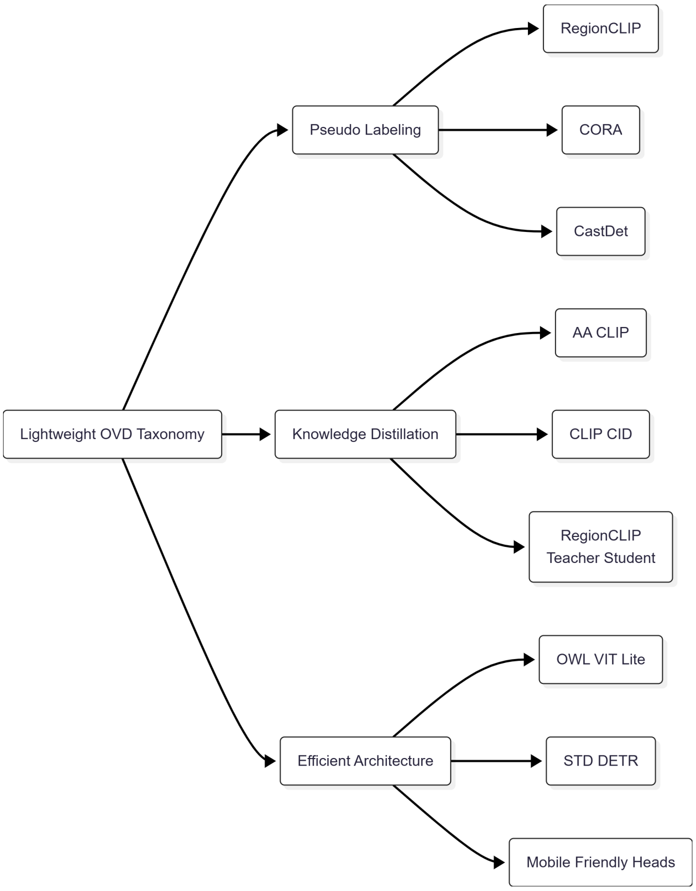

Fig. 3. Lightwight OVD taxonomy

Table 1
Representative lightweight open-vocabulary detection methods and core mechanisms

Method	Paradigm	Core Mechanism	Datasets	Key Strengths	Limitations
RegionCLIP [14]	Pseudo-label + KD	Aligns regional features with text embeddings using CLIP; distillation head for compact transfer	COCO, LVIS	Reduced annotation; strong generalization	Weak spatial grounding
CORA [15]	Pseudo-label	Region prompting and anchor pre-matching for better localization	COCO	Improved region–text consistency	Transformer cost
CastDet [22]	Pseudo-label	Dynamic label queues for aerial imagery	VisDrone	Cross-view robustness	Small-object bias
AA-CLIP [17]	Knowledge Distillation	Attribute-aware loss for rare class recognition	LVIS	Semantic transfer	Requires large teacher
CLIP-CID [16]	Knowledge Distillation	Cluster–instance discrimination for structured features	COCO	Compact representation	Complex training
OWL-ViT [21]	Architecture	Unified vision–language transformer for zero-shot inference	COCO	End-to-end generalization	High memory use
STD-DETR [18]	Architecture	Sparse-token distillation with lightweight decoder	COCO, VisDrone	Fast inference	Requires specialized training

Can the detector be made lighter, faster and more deployable without sacrificing semantic openness? Therefore, research on lightweight open vocabulary detection was born. It is not only making the model smaller; more importantly, it is about finding a sustainable balance between "semantic generalization" and "computational efficiency".

The first systematic exploration direction is pseudo-label learning, researchers first attempted to start from the data layer. The idea is simple and straightforward: Since CLIP understands images from text descriptions, why not let it generate training labels for the detector in return? This approach enables the model to expand the category based on zero manual annotation, to achieve "unsupervised open vocabulary learning". RegionCLIP enables the detector to automatically label candidate regions through image-text matching, thereby learning to recognize unseen categories. Since then, researchers have begun to focus on the quality of pseudo-labels. For instance, CORA [15] enhances positioning accuracy through regional prompts, while CastDet [22] utilizes dynamic label queues to mitigate viewing angle shifts. The pseudo-label paradigm significantly reduces costs, but it reveals an inherent deficiency of the semantic capabilities of the CLIP region: it can recognize "cats", but it may not accurately indicate "where the cat is". Figure 4 shows the structure of pseudo-label.

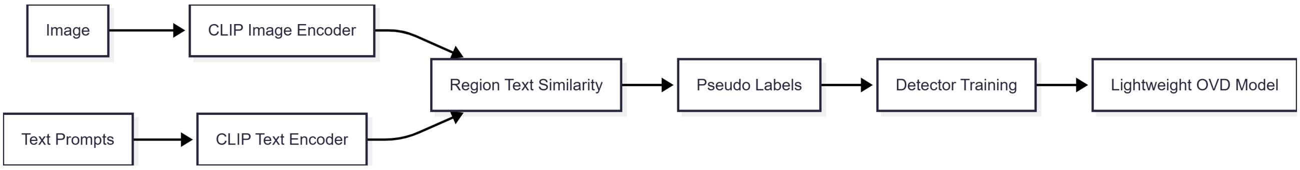

Fig. 4. Pseudo-label learning pipeline for open-vocabulary detection

This contradiction drove the arrival of knowledge distillation. The core idea of knowledge distillation is transferring the knowledge learned by a large, complex “teacher” model into a smaller, simpler “student” model without significantly sacrificing performance. Researchers no longer rely on large models for direct reasoning, but instead let them act as "teachers", transferring their semantic understanding to smaller and more efficient "student" detectors. A typical representative is OVD -st: it fixes the CLIP as a teacher and enables students to learn semantic distribution through contrastive loss. AA-CLIP [17] adds attribute distillation on this basis, enabling the student model to capture rare class features. The advantage of distillation lies in enabling small models to acquire the "soul" of large models, but it also brings new limitations —students are always constrained by the teacher's perspective. To decrease this static dependence, subsequent research has proposed adaptive distillation and feedback distillation mechanisms, enabling students to actively evaluate the reliability of teachers' information during the training process, thereby reducing redundant transfer while retaining semantic capabilities. Figure 5 shows the structure of knowledge distillation. Table 2 summarizes key distillation designs and outcomes.

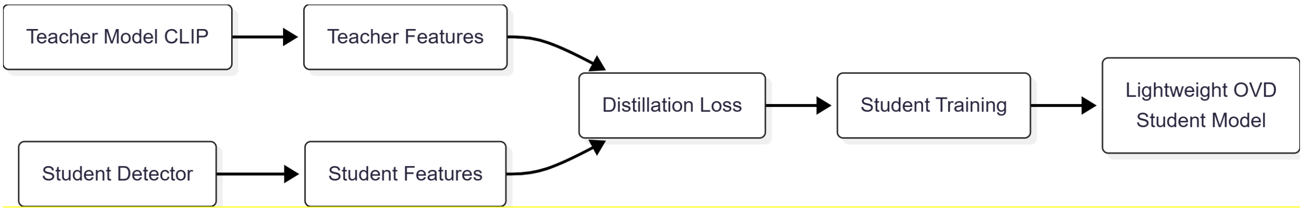

Fig. 5. Knowledge distillation framework for lightweight OVD

Table 2
Representative knowledge-distillation methods for lightweight OVD

Method	Teacher Model	Distillation Objective	Efficiency Gain	Strengths	Limitations
RegionCLIP [14]	CLIP	Feature-alignment and contrastive loss	≈ 30% smaller	Balanced accuracy/size	Partial local grounding
AA-CLIP [17]	CLIP	Attribute-aware semantic loss	≈ 25% smaller	Rare-class robustness	Higher teacher cost
CLIP-CID [16]	CLIP	Cluster–instance discrimination	≈ 40% smaller	Compact features	Complex training schedule

Pseudo-labels and distillation significantly enhance the semantic generalization ability of the model. However, they still cannot completely address the inherent weakness of open vocabulary detection in spatial positioning. The traditional detection methods are different from it. Traditional methods continuously enhance the spatial representation ability. They employ mechanisms such as multi-scale feature fusion (such as FPN [24] and spatial pyramid pooling [25]), keypoint modeling (such as CentripetalNet [26]), and unsupervised pre-training (such as UP-DETR [27]). However, early vision-language models typically performed poorly in terms of region alignment, geometric structure, and small target modeling. Therefore, when the detection task shifts to open vocabulary scenarios, this "strong semantics but weak space" feature will cause obvious performance bottlenecks. It affects dense scenarios, small targets and precise positioning tasks. This limitation has prompted a gradual change in the research direction.Research is no longer merely about improving the semantic side; it has shifted to the systematic optimization of structural and spatial representations.

Thus, lightweight structural design emerged, including lightweight DETR models [28] and modular low-cost OVD frameworks [29]. Methods such as STD-DETR [18], OWL-ViT [21] pull open vocabulary detection from the cloud to real-time applications, through sparse attention, lightweight decoders, or structure-heavy parameterization. These models achieve zero-shot detection on mobile and edge devices, bringing "open vocabulary" from theory to practice, Figure 6 shows the structure of lightweight OVD. However, the lighter the structure, the smaller the semantic capacity. How to maintain the reasoning speed, and keep the cross-modal understanding ability becomes a new bottleneck.

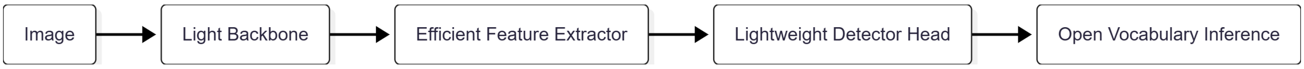
Image
Light Backbone
Efficient Feature Extractor
Lightweight Detector Head
Open Vocabulary Inference

Fig. 6. Efficient architecture design for lightweight OVD

To alleviate the problem of semantic degradation, semantic enhancement and prompt learning have begun to come into view. DetPro [30] and PromptDet [31] enable the model to flexibly adjust the semantic space based on the visual context through learnable or dynamically generated text prompts. AA-CLIP [17] goes a step further by combining attribute modeling with distillation, achieving a more fine-grained semantic expression. These models collectively reveal a trend: future OVDs should not only "understand language", but also "use language" —that is, make language an active mechanism for the model to adapt to different tasks, rather than a static accessory condition.

In the past two years, hybrid and adaptive methods have become the mainstream direction in this field. OVD-SAM [32] integrates the segmentation model with the detector and enhances the positioning accuracy through semantic masking. Multi-stage distillation simultaneously utilizes pseudo-labels and distillation signals to dynamically optimize multi-stage training. Adaptive Token Routing allocates computing resources based on the input complexity to achieve "on-demand reasoning". The series of studies shows that the development of lightweight OVD is shifting from single-point improvement to system integration. Models no longer passively follow preset paradigms but self-adjust according to task and resource conditions.

Early OVD mainly focused on "how to understand more", while lightweight OVD addresses "how to understand more intelligently". Looking back on this evolution, three main threads can be identified:

1. data self-generation —the model gradually learns to create training signals without human intervention;
2. knowledge self-transfer —semantic capabilities no longer rely on large models but can be compressed and transplanted.
3. structural adaptability —computing resources become part of the model design rather than an external constraint.

These three main lines together constitute the core logic of lightweight open vocabulary detection: from external dependence to internal autonomy.

Through trends, we can see that on the one hand, the addition of large language models (LLMS) [33] will make pseudo-label generation and semantic prompts more intelligent and context-aware; on the other hand, hardware-aware model optimization will enable OVD to truly have cross-platform deployment capabilities. The long-term goal is to establish a unified evaluation framework, enabling semantic openness, computational efficiency and domain adaptability to be quantitatively compared within the same coordinate system. As these directions gradually converge, a "scalable, transferable and deployable" visual inspection system may no longer be the future but the standard.

4. Performance Comparison and Trend Analysis

With the continuous development of technology, Lightweight Open-Vocabulary Detection (OVD) has moved from the algorithm design stage to the stage of system performance competition. And it closely revolves around the issue of how to maintain the ability to recognize unknown categories within a limited computing budget. Based on the results of mainstream public database benchmarks (COCO, LVIS, VisDrone) and typical methods (RegionCLIP [14], CORA [15], ViLD [20], OWL-ViT [21], CastDet [22], etc.), this chapter compares and summarizes the performance of lightweight OVD, and focuses on the underlying patterns, trade-offs, trends, and technological evolution directions behind the performance.

For overall accuracy, pseudo-label learning methods are the most stable, as they have the most stable generalization ability. In the open category tasks of COCO and LVIS, RegionCLIP and CORA have consistently led. They all use visual-language models to generate false labels and distillation modules to optimize feature representations. RegionCLIP uses a smaller model and balances recognized classes and unrecognized classes. CORA has incorporated a regional prompt mechanism and achieved the highest AP on unseen classes. CORA has incorporated a regional prompt mechanism and achieved the highest AP on unseen classes, as it enables the model to learn more comprehensive semantic and regional correspondence relationships, and effectively solves the problem of the solidification of traditional detector categories. Compared with them, transformer-based frameworks such as ViLD and OWL-ViT have stronger zero-shot reasoning capabilities. However, their cross-domain generalization is not good enough. When the visual scenes differ greatly from the training data, their performance will decline significantly. From an overall trend perspective, the refinement of semantic alignment remains the key to the successful transfer of open vocabulary detection.

For reasoning speed, the lightweight structure brings significant benefits. RegionCLIP operates on the RN50-Mobile backbone network. Its frame rate is more than twice that of ViLD, and its model size is only half that of ViLD. RegionCLIP operates on the RN50-Mobile backbone network. Its frame rate is more than twice that of ViLD, and its model size is only half that of ViLD, and they keep the detection delay at a real-time level. However, such acceleration usually reduces semantic capacity, as evidenced by the experimental results of OWL-ViT. Transformer models have an advantage in open semantic expression, but they are significantly slow in running speed. This reveals a fundamental contradiction of lightweight OVD: the lighter the model, the more compact its semantic representation, and the more the model relies on the integrity of pre-trained semantic transfer. Therefore, how to design an "elastic balance" between maintaining semantics and compressing structures has become a new focus of current research.

Domain generalization and small object detection tasks make this trade-off even more severe. On the VisDrone aerial photography dataset, CastDet incorporated a dynamic tag caching mechanism, achieving the highest mAP. This proves that the false label strategy can be transferred in specific domains. RegionCLIP is not designed for aerial photography scenes, but it still maintains good cross-domain performance, which indicates that its visual-language alignment has a certain degree of universality. However, when the task shifted to small object detection, the performance of all models dropped significantly. The key reason lies in the spatial resolution limitations of multimodal features, there is an inherent conflict between semantic embedding and local localization. This is also a common drawback of the current lightweight OVDs: after the fusion of visual features in coarse-grained spaces, the expression of the target boundary is often weakened. This kind of problem also occurs in high-precision scenarios such as medical imaging. Although lightweight models such as Pterygium-Net [34] have demonstrated the potential of compact architectures. However, in fine-grained tasks, semantic loss still cannot be ignored.

We can conclude that there are three stable rules for the performance of lightweight OVD.

i. The pseudo-labeling method has a natural advantage in "open semantics". It is particularly suitable for tasks with little data or across domains.
ii. The knowledge distillation method has the greatest potential in "efficiency compression". It can reduce the model size without significantly lowering the accuracy.
iii. The architecture optimization method has made a breakthrough in "real-time reasoning". It lays the foundation for embedded and mobile deployments.

Hybrid strategies such as RegionCLIP combine the advantages of all three. It is currently a representative solution for balancing performance and efficiency. This trend indicates that future open vocabulary detection will not rely on just one route. It is moving towards a system optimization direction that integrates multiple strategies.

However, the numerical comparisons across papers are not entirely fair. Different jobs vary in backbone structure, data partitioning, prompt templates, and even the granularity of training labels. For instance, some models use weak label completion on LVIS, while others directly apply COCO weight transfer. This inconsistency in the experimental setup will cause performance differences on the surface. Therefore, a major issue in current OVD research is the lack of a unified evaluation standard. This can result in situations where, under lightweight settings, the same level of semantic accuracy yields completely different latency and energy consumption across different hardware platforms or batch-processing scales. Therefore, we must establish a standardized evaluation protocol to quantify the three-dimensional relationship of "semantic openness —computational cost —deployment feasibility".

In conclusion, the research on lightweight OVD has shifted from merely competing in performance indicators to optimizing the overall system structure. The core breakthrough in the future lies in the coordinated development of distillation, pseudo-label generation and hardware adaptive architecture. The core breakthrough in the future lies in the coordinated development of distillation, pseudo-label generation and hardware adaptive architecture.

5. Limitation and Future Direction

Progress has been made in lightweight open vocabulary object detection (OVD). It has made significant progress in algorithm performance and deployment feasibility. However, there are still some key issues before it becomes a true universal perception system. These limitations mainly focus on three aspects: semantic bias, structural issues and evaluation systems [35,36]. They jointly determine the highest level of performance for lightweight OVD.

At the semantic level, the current mainstream vision-language models (such as CLIP) have inherent biases. This deviation usually stems from the imbalance of the training data [13]. It leads to the model's excellent recognition of high-frequency and obvious categories. However, its ability to recognize abstract, fine-grained or rare categories has significantly declined. On MS-COCO, the detection accuracy of abstract classes such as "anomaly" is only 18.4% [37]. Semantic bias will be further magnified during the process of false labeling and distillation. This limits the generalization ability of lightweight models in open scenarios. Future research can start from dynamic prompts (prompts based on LLM [33]) and semantic causal modeling.Researchers can utilize large language models to generate context-related prompt words, this can improve semantic alignment and long-tail class understanding.

At the model level, the contradiction between lightweighting and spatial accuracy remains significant. The recall rate of traditional regional proposal networks on small targets is lower than 60% [38]. This directly led to a decline in the detection rate of the new category. Over-compressing the model structure can increase speed, but it impairs the fidelity of semantics. Recently, the idea of no-candidate box detection and adaptive distillation has been proposed, which provides a direction for breaking through this bottleneck. The former improves the recall efficiency through full convolutional feature fusion. The latter enhances the adaptability of the student model at different semantic levels through dynamic weighted distillation. The future trend will be a dynamic model framework with a bidirectional coupling of "structural lightweighting" and "semantic preservation".

The future trend will be a dynamic model framework with a bidirectional coupling of "structural lightweighting" and "semantic preservation". Different studies vary greatly in the selection of backbone networks, prompt design and training strategies [39]. This makes the performance comparison across papers lack reproducibility. Establishing cross-domain and continuous learning benchmarks has become a consensus direction in this field [40]. A truly sustainable open detection system should be able to continuously expand the category vocabulary without changing the calculation budget, while avoiding catastrophic forgetting. To achieve this goal, it is necessary to remember efficient fine-tuning mechanisms and adaptive evaluation protocols, with the aim of maintaining consistent performance in a multi-domain environment.

The ultimate goal of lightweight OVD is to deploy it at low power consumption in real-world environments, including mobile devices, drones, autonomous driving computers and edge devices. Relevant studies have verified this demand in different practical scenarios. For instance: mobile augmented reality and edge computing tasks [41,42], unmanned aerial vehicles and 3D perception scenes [43], as well as autonomous driving and traffic perception systems [44]. However, most current OVD methods still rely on high-bandwidth video memory. They require a fixed input resolution and a large-scale Transformer structure. This is significantly different from the limitations of actual hardware in terms of computing power, latency and energy consumption [45,46]. Future work must develop hardware-aware compression methods. It also requires a cross-platform integration framework. This enables the open vocabulary model to be migrated between different devices and maintain stable performance. As semantic openness, computational efficiency and multi-platform adaptability all enter the unified evaluation system, a truly scalable, transferable and deployable OVD system will transform from a research idea into an industry standard.

6. Conclusion

This paper establishes an analytical framework covering three layers: data, representation and system. By classifying and comparing the mainstream methods, we have summarized three main development lines: Pseudo-label learning enhances data efficiency and cross-domain generalization, knowledge distillation promotes lightweight semantic transfer of large models, and architectural optimization drives real-time deployment. Hybrid strategies are becoming the core direction for balancing accuracy and efficiency.

Furthermore, there is still a structural trade-off between open semantics and computational efficiency in current lightweight OVDs. Semantic bias and insufficient spatial accuracy are the main bottlenecks, and the unified evaluation standard and continuous learning mechanism still need to be improved. Future research should focus on establishing a dynamic balance among semantic understanding, structural lightweighting and system scalability, enabling detection systems to continuously learn and maintain generalization capabilities in resource-constrained environments.

The development of lightweight open vocabulary detection is not only a technological innovation in the field of object detection, but also an important step for visual understanding to move from closed recognition to open cognition. With the in-depth integration of multimodal models and adaptive computing, a scalable, deployable and interpretable open perception system is taking shape.

Acknowledgement
The authors would like to thank Universiti Kebangsaan Malaysia for providing financial support und er the “FTM1 - Peruntukan Dana Fakulti Teknologi dan Sains Maklumat, UKM”.

References

[1] Girshick, Ross. "Fast r-cnn." In *Proceedings of the IEEE international conference on computer vision*, pp. 1440-1448. 2015. https://doi.org/10.1109/ICCV.2015.169

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

[3] Duan, Kaiwen, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. "Centernet: Keypoint triplets for object detection." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6569-6578. 2019. https://doi.org/10.1109/ICCV.2019.00667

[4] Zhao, Zhong-Qiu, Peng Zheng, Shou-tao Xu, and Xindong Wu. "Object detection with deep learning: A review." *IEEE transactions on neural networks and learning systems* 30, no. 11 (2019): 3212-3232. https://doi.org/10.1109/TNNLS.2018.2876865

[5] Bachute, Mrinal R., and Javed M. Subhedar. "Autonomous driving architectures: insights of machine learning and deep learning algorithms." *Machine Learning with Applications* 6 (2021): 100164. https://doi.org/10.1016/j.mlwa.2021.100164

[6] Iqbal, Muhammad Javed, Muhammad Munwar Iqbal, Iftikhar Ahmad, Madini O. Alassafi, Ahmed S. Alfakeeh, and Ahmed Alhomoud. "Real-time surveillance using deep learning." *Security and Communication Networks* 2021, no. 1 (2021): 6184756. https://doi.org/10.1155/2021/6184756

[7] Karim, Tajbia, Zainal Rasyid Mahayuddin, and Mohammad Kamrul Hasan. "Singular and multimodal techniques of 3D object detection: Constraints, advancements and research direction." *Applied Sciences* 13, no. 24 (2023): 13267. https://doi.org/10.3390/app132413267

[8] Saif, FM Saifuddin, and Zainal Rasyid Mahayuddin. "Vision based 3D object detection using deep learning: Methods with challenges and applications towards future directions." *International Journal of Advanced Computer Science and Applications* 13, no. 11 (2022): 203-214. https://doi.org/10.14569/IJACSA.2022.0131123

[9] Yuan, Lu, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu et al. "Florence: A new foundation model for computer vision." *arXiv preprint arXiv:2111.11432* (2021).

[10] Gan, Zhe, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. "Vision-language pre-training: Basics, recent advances, and future trends." *Foundations and Trends® in Computer Graphics and Vision* 14, no. 3–4 (2022): 163-352. https://doi.org/10.1561/0600000105

[11] Gan, Zhe, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. "Vision-language pre-training: Basics, recent advances, and future trends." *Foundations and Trends® in Computer Graphics and Vision* 14, no. 3–4 (2022): 163-352. https://doi.org/10.1561/0600000105

[12] Su, Zhihao, Afzan Adam, Mohammad Faidzul Nasrudin, and Anton Satria Prabuwono. "Proposal-free fully convolutional network: object detection based on a box map." *Sensors* 24, no. 11 (2024): 3529. https://doi.org/10.3390/s24113529

[13] Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. "Learning transferable visual models from natural language supervision." In *International conference on machine learning*, pp. 8748-8763. PmLR, 2021.

[14] Zhong, Yiwu, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou et al. "Regionclip: Region-based language-image pretraining." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16793-16803. 2022. https://doi.org/10.1109/CVPR52688.2022.01629

[15] Wu, Xiaoshi, Feng Zhu, Rui Zhao, and Hongsheng Li. "Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7031-7040. 2023. https://doi.org/10.1109/CVPR52729.2023.00679

[16] Yang, Kaicheng, Tiancheng Gu, Xiang An, Haiqiang Jiang, Xiangzi Dai, Ziyong Feng, Weidong Cai, and Jiankang Deng. "Clip-cid: Efficient clip distillation via cluster-instance discrimination." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 20, pp. 21974-21982. 2025. https://doi.org/10.1609/aaai.v39i20.35505

[17] Ma, Wenxin, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S. Kevin Zhou. "Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip." In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4744-4754. 2025. https://doi.org/10.1109/CVPR52734.2025.00447

[18] YIN, ZY, B. Yang, and JL CHEN. "Lightweight small object detection algorithm based on STD-DETR." *Laser & Optoelectronics Progress* 62, no. 8 (2025): 146-156. https://doi.org/10.3788/LOP241849

[19] Nafea, Mohammed Mansoor, Siok Yee Tan, Mohammed Ahmed Jubair, and Tareq Abd Mustafa. "A review of lightweight object detection algorithms for mobile augmented reality." *International Journal of Advanced Computer Science and Applications* 13, no. 11 (2022). https://doi.org/10.14569/IJACSA.2022.0131162

[20] Minderer, Matthias, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran et al. "Simple open-vocabulary object detection." In *European conference on computer vision*, pp. 728-755. Cham: Springer Nature Switzerland, 2022. https://doi.org/10.1007/978-3-031-20080-9_42

[21] Minderer, Matthias, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran et al. "Simple open-vocabulary object detection." In *European conference on computer vision*, pp. 728-755. Cham: Springer Nature Switzerland, 2022. https://doi.org/10.1007/978-3-031-20080-9_42

[22] Kondo, Yuki, Norimichi Ukita, Riku Kanayama, Yuki Yoshida, Takayuki Yamaguchi, Xiang Yu, Guang Liang et al. "Mva 2025 small multi-object tracking for spotting birds challenge: Dataset, methods, and results." In *2025 19th International Conference on Machine Vision and Applications (MVA)*, pp. 1-15. IEEE, 2025. https://doi.org/10.23919/MVA65244.2025.11175061

[23] Jia, Chao, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. "Scaling up visual and vision-language representation learning with noisy text supervision." In *International conference on machine learning*, pp. 4904-4916. PMLR, 2021.

[24] Lin, Tsung-Yi, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. "Feature pyramid networks for object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117-2125. 2017. https://doi.org/10.1109/CVPR.2017.106

[25] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Spatial pyramid pooling in deep convolutional networks for visual recognition." *IEEE transactions on pattern analysis and machine intelligence* 37, no. 9 (2015): 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824

[26] Dong, Zhiwei, Guoxuan Li, Yue Liao, Fei Wang, Pengju Ren, and Chen Qian. "Centripetalnet: Pursuing high-quality keypoint pairs for object detection." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10519-10528. 2020. https://doi.org/10.1109/CVPR42600.2020.01053

[27] Dai, Zhigang, Bolun Cai, Yugeng Lin, and Junying Chen. "Up-detr: Unsupervised pre-training for object detection with transformers." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1601-1610. 2021. https://doi.org/10.1109/CVPR46437.2021.00165

[28] Wang, Yu, Xiangbo Su, Qiang Chen, Xinyu Zhang, Teng Xi, Kun Yao, Errui Ding, Gang Zhang, and Jingdong Wang. "OVLW-DETR: Open-Vocabulary Light-Weighted Detection Transformer." *arXiv preprint arXiv:2407.10655* (2024).

[29] Faye, Bilal, Binta Sow, Hanane Azzag, and Mustapha Lebbah. "A Lightweight Modular Framework for Low-Cost Open-Vocabulary Object Detection Training." *arXiv preprint arXiv:2408.10787* (2024).

[30] Du, Yu, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. "Learning to prompt for open-vocabulary object detection with vision-language model." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14084-14093. 2022. https://doi.org/10.1109/CVPR52688.2022.01369

[31] Feng, Chengjian, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. "Promptdet: Towards open-vocabulary detection using uncurated images." In *European conference on computer vision*, pp. 701-717. Cham: Springer Nature Switzerland, 2022. https://doi.org/10.1007/978-3-031-20077-9_41

[32] Yuan, Haobo, Xiangtai Li, Chong Zhou, Yining Li, Kai Chen, and Chen Change Loy. "Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively." In *European Conference on Computer Vision*, pp. 419-437. Cham: Springer Nature Switzerland, 2024. https://doi.org/10.1007/978-3-031-72775-7_24

[33] Wu, Aoqi, Yifan Yang, Xufang Luo, Yuqing Yang, Chunyu Wang, Liang Hu, Xiyang Dai, Dongdong Chen, Chong Luo, and Lili Qiu. "LLM2CLIP: Powerful Language Model Unlock Richer Visual Representation." In *NeurIPS 2024 Workshop: Self-Supervised Learning-Theory and Practice*.

[34] Zulkifley, Mohd Asyraf, Siti Raihanah Abdani, and Nuraisyah Hani Zulkifley. "Pterygium-Net: a deep learning approach to pterygium detection and localization." *Multimedia Tools and Applications* 78, no. 24 (2019): 34563-34584. https://doi.org/10.1007/s11042-019-08130-x

[35] Zhu, Chaoyang, and Long Chen. "A survey on open-vocabulary detection and segmentation: Past, present, and future." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, no. 12 (2024): 8954-8975. https://doi.org/10.1109/TPAMI.2024.3413013

[36] Bianchi, Lorenzo, Fabio Carrara, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. "The devil is in the fine-grained details: Evaluating open-vocabulary object detectors for fine-grained understanding." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22520-22529. 2024. https://doi.org/10.1109/CVPR52733.2024.02125

[37] Shao, Jie-Jing, Jiang-Xin Shi, Xiao-Wen Yang, Lan-Zhe Guo, and Yu-Feng Li. "Investigating the limitation of clip models: The worst-performing categories." *arXiv preprint arXiv:2310.03324* (2023).

[38] Chen, Chenyi, Ming-Yu Liu, Oncel Tuzel, and Jianxiong Xiao. "R-CNN for small object detection." In *Asian conference on computer vision*, pp. 214-230. Cham: Springer International Publishing, 2016. https://doi.org/10.1007/978-3-319-54193-8_14

[39] Lin, Feng, Wenze Hu, Yaowei Wang, Yonghong Tian, Guangming Lu, Fanglin Chen, Yong Xu, and Xiaoyu Wang. "Universal object detection with large vision model." *International Journal of Computer Vision* 132, no. 4 (2024): 1258-1276. https://doi.org/10.1007/s11263-023-01929-0

[40] Yao, Yiyang, Peng Liu, Tiancheng Zhao, Qianqian Zhang, Jiajia Liao, Chunxin Fang, Kyusong Lee, and Qing Wang. "How to evaluate the generalization of detection? a benchmark for comprehensive open-vocabulary detection." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, pp. 6630-6638. 2024. https://doi.org/10.1609/aaai.v38i7.28485

[41] Chen, Chen, Bin Liu, Shaohua Wan, Peng Qiao, and Qingqi Pei. "An edge traffic flow detection scheme based on deep learning in an intelligent transportation system." *IEEE transactions on intelligent transportation systems* 22, no. 3 (2020): 1840-1852. https://doi.org/10.1109/TITS.2020.3025687

[42] Ghasemi, Yalda, Heejin Jeong, Sung Ho Choi, Kyeong-Beom Park, and Jae Yeol Lee. "Deep learning-based object detection in augmented reality: A systematic review." *Computers in Industry* 139 (2022): 103661. https://doi.org/10.1016/j.compind.2022.103661

[43] Karim, Tajbia, Zainal Rasyid Mahayuddin, and Mohammad Kamrul Hasan. "Singular and multimodal techniques of 3D object detection: Constraints, advancements and research direction." *Applied Sciences* 13, no. 24 (2023): 13267. https://doi.org/10.3390/app132413267

[44] Chen, Long, Shaobo Lin, Xiankai Lu, Dongpu Cao, Hangbin Wu, Chi Guo, Chun Liu, and Fei-Yue Wang. "Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey." *IEEE Transactions on Intelligent Transportation Systems* 22, no. 6 (2021): 3234-3246. https://doi.org/10.1109/TITS.2020.2993926

[45] Chen, Chen, Bin Liu, Shaohua Wan, Peng Qiao, and Qingqi Pei. "An edge traffic flow detection scheme based on deep learning in an intelligent transportation system." *IEEE transactions on intelligent transportation systems* 22, no. 3 (2020): 1840-1852. https://doi.org/10.1109/TITS.2020.3025687

[46] Liu, Youlin, Zainal Rasyid Mahayuddin, and Mohammad Faidzul Nasrudin. "Text-Guided Spatio-Temporal 2D and 3D Data Fusion for Multi-Object Tracking with RegionCLIP." *Applied Sciences* 15, no. 18 (2025): 10112. https://doi.org/10.3390/app151810112