# A Survey on Open-Vocabulary Detection

Jiewei Lyu
Sun Yat-sen University
Guangzhou, China
lvjw7@mail2.sysu.edu.cn

## Abstract

*In the field of visual scene understanding, deep neural networks have made impressive advancements in various core tasks like segmentation, tracking, and detection. However, most approaches operate on the close-set assumption, meaning that the model can only identify pre-defined categories that are present in the training set. Recently, open vocabulary settings were proposed due to the rapid progress of vision language pre-training. These new approaches seek to locate and recognize categories beyond the annotated label space. The open vocabulary approach is more general, practical, and effective than weakly supervised and zero-shot settings. This paper thoroughly reviews open vocabulary learning, summarizing and analyzing recent developments in the field. In particular, I begin by juxtaposing open vocabulary learning with analogous concepts such as zero-shot learning, open-set detection, and open world detection. Subsequently, I categorize the existing methods into three main categories: introduction of large-scale datasets, integration of multimodal large models, and generative open vocabulary object detection. Among these, the integration of multimodal large models can be further subdivided into knowledge distillation, pseudo-label, and transfer learning. An in-depth analysis is conducted on the main design principles, key challenges, development routes, and the strengths and weaknesses of the methodologies, accompanied by a performance comparison. Finally, several promising directions are provided and discussed to stimulate future research.*

## 1. Introduction

Object detection is a core task in computer vision for scene perception, serving as the foundation for many real-world applications such as autonomous driving and intelligent robotics. As a fundamental task of scene understanding, object detection has made significant progress in the era of deep learning[17]. However, traditional visual perception models struggle to correctly identify unfamiliar categories in open scenarios; even state-of-the-art supervised learning models find it challenging to generalize beyond a closed set of categories. Additionally, existing datasets are often limited in scale, with the largest, such as the LVIS dataset[7], annotating only 1,203 categories. These limitations of closed sets significantly hinder the application of current detectors in real-world scenarios. To address these limitations and avoid the high cost of manual labeling, zero-shot learning and open vocabulary learning have been proposed.

However, zero-shot learning has its limitations. The model relies solely on predefined word embeddings when inferring unseen categories, which restricts its ability to explore the relationships between visual information and unseen classes, leading to poor recognition performance for new categories. Open vocabulary learning aims to enable the model to use a broader vocabulary during training, allowing it to recognize more unseen categories during inference.

Open vocabulary detection(OVD) combines images with natural language descriptions to form an expandable set of labels, allowing models to continuously update and recognize new objects and scenes in real-world applications. In this way, OVD overcomes the limitations of closed sets, achieving broader generalization capabilities and performance improvements.

## 2. Background

### 2.1. Large Vision-Language Models (VLMs)

Inspired by advances in natural language processing, a new deep learning paradigm called "Vision-Language Model Pre-training" has recently garnered increasing attention. In this paradigm, Vision-Language Models (VLMs) are pre-trained on large-scale image-text pairs available abundantly on the internet. The goal is to learn image-text correlations to enable effective zero-shot predictions in visual recognition tasks.During the pre-training process, VLMs first use text encoders and image encoders to extract features from images and texts, respectively, and then learn visual-

language correlations according to specific pre-training objectives. By matching the embeddings of any given image and text, VLMs can perform zero-shot evaluations on unseen data. Benefiting from large-scale pre-training, large VLMs demonstrate exceptional zero-shot transfer capabilities, forming the basis for many studies in the Open Vocabulary Detection (OVD) field.

Image-text contrastive pre-training aims to learn visual-language correlations by contrasting image-text pairs, bringing the embeddings of paired images and texts closer while pushing apart the embeddings of other images and texts. The CLIP[16] pre-training objective aligns positive image-caption pairs within a batch, enabling efficient and scalable learning of transferable representations. CLIP employs a symmetric image-text infoNCE loss, where a multi-head self-attention pooling layer aggregates patch embeddings into a holistic representation for image embeddings. Both text and image embeddings are L2 normalized to calculate their pairwise cosine similarities. In this manner, text embeddings are treated as frozen classifiers. By performing contrastive pre-training on 400 million image-caption pairs, CLIP achieved a significant breakthrough. Inspired by CLIP's tremendous success, numerous studies have improved symmetric image-text infoNCE loss from various angles. ALIGN[8] expanded VLM pre-training scale using a massive but noisy dataset of image-text pairs through noise-robust contrastive learning.

Generative VLM pre-training learns semantic knowledge by generating images or texts through Masked Image Modeling, Masked Language Modeling, Masked Cross-Modal Modeling, and image-to-text generation. This pre-training objective involves learning image context formation by masking and reconstructing images. In OVD, image-to-text generation is commonly used to train VLMs to predict tokenized texts, generating descriptive texts for given images to capture fine-grained visual-language correlations. It first encodes the input image into intermediate embeddings, which are then decoded into descriptive texts. This approach can generate caption pseudo-labels for images.

### 2.2. Related Research Domains

In recent years, the field of Open Vocabulary Detection (OVD) has made significant advancements. However, its research trajectory does not exist in isolation and is closely intertwined with related research domains. I briefly compare these concepts in Fig.1, including zero-shot, open-set, open world, and open vocabulary.The following is a introduction to related work.

**Zero-Shot Detection (ZSD)**[2] is a precursor to OVD, focusing on recognizing new classes without annotated data. ZSD primarily relies on semantic embeddings (e.g., Word2Vec, GloVe, and BERT) to achieve cross-modal knowledge transfer. However, due to the lack of alignment with the visual modality, these semantic embeddings exhibit issues such as high noise and insufficient accuracy in practical applications.

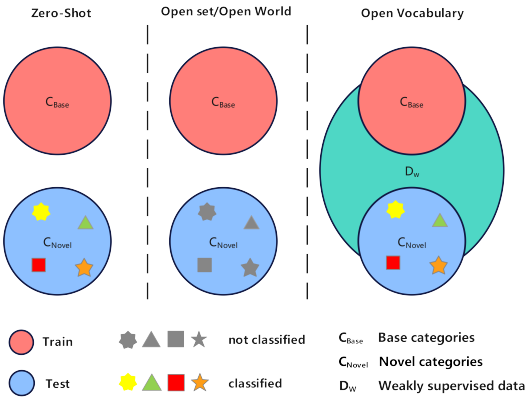



Figure 1. Concepts comparison between zero-shot, open-set/open world and open vocabulary.

**Open Set Detection** [4][15]originates from open set recognition[18][5]. It requires classifying known categories and identifying a single "unknown" category without further categorization of specific classes. The main objective is to reject unknown categories that unexpectedly appear and potentially compromise the robustness of the recognition system.

**Open World Detection**[9] further extends the concept of open set detection. It not only requires the model to detect new categories during testing but also to learn and incorporate new category knowledge into the known categories. Like open set detection, it does not require further categorization of specific classes. Open world detection necessitates the design of models capable of continuously updating and expanding in dynamic environments while ensuring stability and accuracy.

## 3. SURVEY

In this section, I will explain the research motivations of the relevant papers and the hierarchical logical relationships of their technical points. In Fig.2, I summarize the timeline of some papers on open vocabulary detection. From a macro perspective, these papers can be categorized into three types: 1) Introduction of large-scale datasets, 2) Integration of multimodal large models, and 3) Generative open vocabulary object detection.

### 3.1. Introduction of Large-Scale Datasets

The core idea of this type of method is to incorporate image-text pairs into the detection training phase.In Fig.3a,

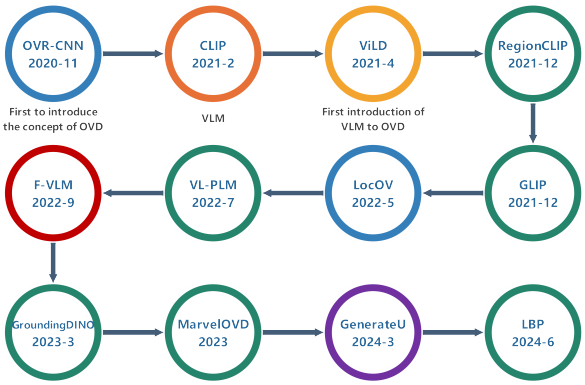



Figure 2. Timeline of Open Vocabulary Detection. Orange represents VLM, blue represents the use of image-caption pairs, yellow represents the use of knowledge distillation, green represents the use of pseudo-labels, red represents the use of Transfer Learning, and purple represents the use of generative methods.

I present the framework. The extensive vocabulary in the titles includes both Base Class and Novel Class. Therefore, aligning proposals containing Novel Classes with words containing Novel Classes can improve detection of Novel Classes.

**OVR-CNN**[20]: To address the limitation of closed category sets faced by traditional object detection systems, OVR-CNN was introduced, aiming to leverage large image-caption datasets to enhance the model's ability to recognize unknown objects. This pioneering work first proposed the concept of open vocabulary detection, allowing object detection to go beyond a limited number of annotated categories and to transfer open vocabulary visual-semantic knowledge learned from captions to downstream detection tasks, thus achieving more generalized object detection. The core of this work lies in using image-caption pair data for pre-training the visual encoder, followed by fine-tuning on a bounding box dataset. Since captions contain rich descriptions of fine-grained features of image regions, they cover more object categories. Therefore, through large-scale image-caption pre-training, the multimodal encoder (ResNet50 and V2L fully connected layers and BERT) can learn more generalized visual-semantic mappings. Ultimately, this well-pretrained multimodal encoder is integrated into the Faster R-CNN framework, replacing the original encoder, significantly enhancing the model's performance in zero-shot detection tasks.

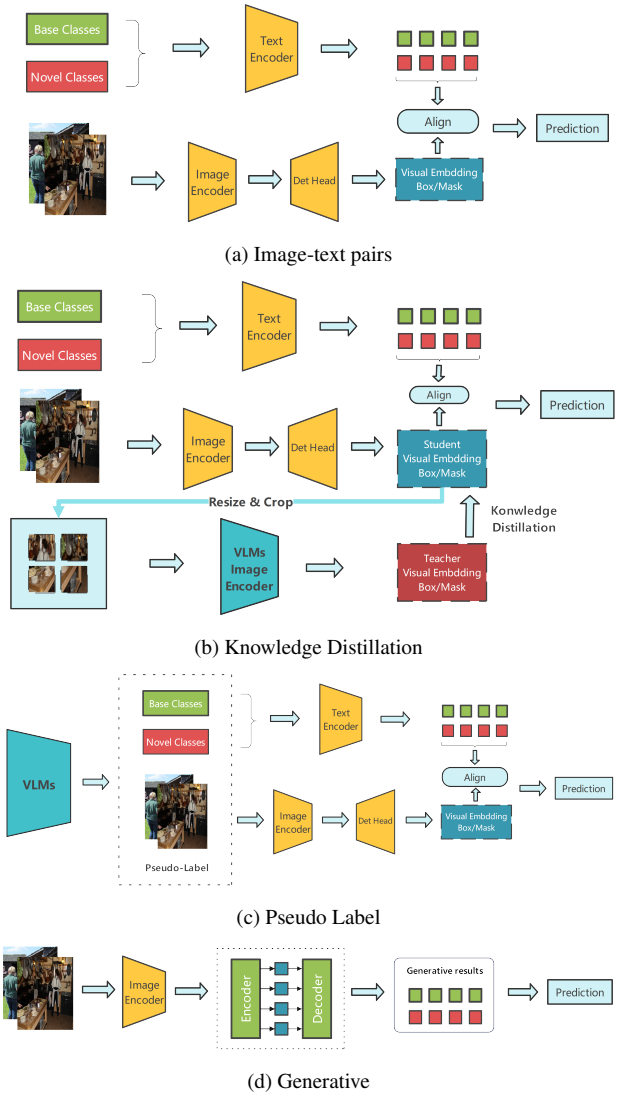
**LocOV**[3]: Building on this foundation, to further explore the potential of image-caption data, LocOV introduces consistency regularization techniques to better utilize the information from cross-modal image-caption pairs. It also introduces two stages: local semantic matching (LSM) and specific task tuning (STT), using classification suggestions from the RPN to train Faster R-CNN by matching region features and word embeddings in images and captions, respectively. Through more refined image region processing and more effective text feature matching, the model's ability to recognize novel categories is enhanced. Although GLIP constructs the object detection task as a phrase grounding task and introduces a grounding dataset, I still prefer to classify it into the next category.

## 3.2. Integration of Multimodal Large Models

This category of methods introduces large pretrained Vision-Language Models (VLMs) to enhance OVD performance. It can be further subdivided into three types: 1) Knowledge distillation methods, which utilize the powerful image-text alignment capabilities of pretrained VLMs (as teacher models) to guide the object detection models (students) in better matching visual features with semantic labels at the region level(see Fig.3c). 2) Pseudo-labeling methods, similar to those in "Introduction of Large-Scale Datasets" in Sec.3.1, which use rich image-text pairs. Additionally, these methods employ large pretrained VLMs (such as RegionCLIP) or self-training (such as GLIP) to generate pseudo-labels, enabling the model to learn about Novel Classes.(see Fig.3b) 3) Transfer learning methods, which directly use pretrained multimodal VLMs as the visual encoders (backbones) for detection models. Since VLMs have already learned rich visual and semantic features from large-scale cross-modal data, it is only necessary to add or fine-tune specialized detection heads on top to adapt to the specific needs of object detection tasks.

### 3.2.1 Knowledge Distillation

**ViLD**[6]:The background of ViLD is the significant potential shown by pretrained models in learning joint image-text representations. Against this backdrop, ViLD first introduced the pretrained multimodal model CLIP to enhance OVD performance. The core of ViLD is the application of the knowledge distillation method. First, it uses a pretrained open vocabulary image classification model (such as CLIP or ALIGN) as a teacher model to encode image regions and textual descriptions, achieving image-text alignment. Then, it trains a two-stage detector (such as Mask R-CNN). To address the limitations of the CLIP model in region-level image recognition, ViLD uses RPN scores to assist CLIP in region-level predictions, ensuring that region embeddings align with the image and text embeddings produced by the teacher model. Specifically, ViLD consists of two branches: the ViLD-text branch and the ViLD-image branch. In ViLD-text, base category texts are input into the CLIP text encoder to obtain text embeddings, which are then used to classify target regions. In ViLD-image, corresponding proposals are input into the CLIP image en-




(a) Image-text pairs

(b) Knowledge Distillation

(c) Pseudo Label

(d) Generative

Figure 3. Framework

coder to obtain image embeddings, and knowledge distillation is performed on region embeddings after ROI alignment. Compared to ViLD-text, ViLD-image distills information from both base and novel classes, as the proposals from the proposal network may contain novel classes, whereas ViLD-text only uses text information from base classes.

### 3.2.2 Pseudo-Label

**RegionCLIP**[22]: The starting point for RegionCLIP is the observation that the multimodal large model CLIP, introduced for OVD tasks, has low recognition rates at the region level and is insensitive to bounding box locations. This is because CLIP was trained to match images as a whole with text descriptions, without capturing fine-grained alignment between image regions and text spans. Therefore, RegionCLIP begins by using the CLIP model to perform initial image-text pair pretraining, obtaining visual and language encoders that efficiently encode images and text. Based on captions, it extracts key concepts and converts traditional labels into a prompt format, using the pretrained language encoder to extract text features. Simultaneously, it uses an RPN network to generate bounding boxes, inputs them into the CLIP visual encoder, and uses contrastive learning to match and label the extracted image regions with pseudo-labels. Finally, these finer-grained pseudo region-word pairs are used for CLIP-style pretraining at the region level, achieving fine-grained alignment between image regions and words, thereby improving region-level detection capabilities.

**GLIP**[12]: Another effort aiming to achieve a finer-grained understanding of images based on CLIP is GLIP. GLIP constructs object detection as a phrase grounding task, unifying phrase grounding and object detection tasks. Since grounding datasets contain very rich visual object names and concepts, training the model with both types of datasets can greatly enhance OVD performance. GLIP first uses existing grounding data and labeled detection data for supervised training, extracting feature embeddings through the Text Encoder and Visual Encoder, and aligning the cross-modal embeddings via the Deep Fusion module to obtain a teacher model. Then, based on the teacher model, it generates "region-text pairs" pseudo-labels from "image-text pairs". These pseudo-labels are combined with the original labeled data to train the student model. Through this self-training approach, GLIP can expand its training data, further improving the model's generalization capability for new concepts. Like CLIP, GLIP can adapt to different downstream tasks through prompt tuning or efficient task adapters, reducing the training costs for downstream tasks.

**Grounding DINO**[14]: While GLIP is based on the traditional single-stage detector dynamic head design, which may limit its performance ceiling when handling unseen categories, Grounding DINO upgrades GLIP's detector to a Transformer-based detector on the foundation of DINO. It uses Swin-Transformer and BERT to extract feature embeddings for text and images, respectively. These features are then sent to a Feature Enhancer for finer-grained multimodal feature fusion. After obtaining cross-modal text and image features, the Language-guided query selection module selects text features more closely aligned with the image features to initialize Query tokens. Finally, these cross-modal queries are sent to the Cross-Modality Decoder, which decodes the queries using multimodal features to predict object boxes and extract corresponding phrases.

**LBP**[11]: Researchers of LBP observed that previous studies transferring VLMs (e.g., CLIP) knowledge to object de-

tection tasks through knowledge distillation tend to represent background proposals with a single "background class," ignoring the diversity of background categories. This can cause the trained model to fail to capture various implicit knowledge within background proposals. Moreover, there may be conceptual overlap between the estimated background categories and novel categories, which could hinder accurate probability scoring for novel categories, resulting in ambiguity during inference. Hence, the LBP model was proposed. The LBP framework includes three modules: BCP, which discovers and represents latent background categories estimated from background proposals; BOD, which generates pseudo-labels by clustering background proposals in images using k-means, helping to identify and distinguish hidden objects within the background; and IPR, which corrects the probability scores for novel categories, addressing the issue of conceptual overlap between estimated background categories and novel categories during inference, thus enabling the model to accurately compute probabilities for novel categories.

#### 3.2.3 Improving Pseudo-Label Quality

**VL-PLM**[21]: While works like GLIP and RegionCLIP utilize VLMs to generate pseudo-labels for model pretraining, researchers of VL-PLM identified that pseudo-labels constructed by visual-language models (e.g., CLIP) for unlabeled images can suffer from poor localization accuracy despite having high CLIP scores. To address this, VL-PLM combines RPN scores, which correlate positively with predicted region IoU scores, with CLIP classification scores. By repeatedly feeding the predicted regions into the RoI Head for adjustment, this approach effectively reduces redundant proposals and enhances the localization accuracy of pseudo-labels.

**MarvelOVD**[1]: Researchers of MarvelOVD observed that the pseudo-labels generated by VLMs often contain noise due to the domain gap between VLMs pretraining and object detection tasks. This noise stems from VLMs' limited ability to comprehend the context of local image region proposals. To mitigate this, MarvelOVD integrates the capabilities of both the detector and VLMs in an online manner, using the rich contextual information provided by the detector to alleviate domain shift issues in VLM's local region predictions, thereby reducing noisy labels during the initial training phase and progressively improving pseudo-label quality. Additionally, MarvelOVD introduces an adaptive weighting mechanism to suppress the impact of poorly aligned training boxes and employs a hierarchical label assignment method to resolve conflicts between pseudo-labels and base class annotations, preventing negative effects on base class detection performance.

#### 3.2.4 Transfer Learning

**F-VLM**[10]: Prior to F-VLM, many works involved further modifications of CLIP, such as RegionCLIP and GLIP. However, researchers of F-VLM discovered that the original CLIP features already contain rich semantic and local perceptual information, enabling strong region classification even without parameter fine-tuning. Thus, F-VLM abandons knowledge distillation, task-specific pretraining, or weakly supervised learning methods, and instead, trains the detector head on a frozen VLMs backbone, combining the outputs of the detector and VLMs during inference to obtain the final detection results. This approach eliminates the computational demands of knowledge distillation, pretraining, or weakly supervised learning, achieving better performance compared to ViLD while saving up to 200 times in training computation.

### 3.3. Generative Open Vocabulary Object Detection

**GenerateU**[13]: Researchers of GenerateU observed that existing open vocabulary object detection techniques significantly expand object categories by leveraging weak supervision (e.g., image-text pairs) or large pretrained visual-language models (e.g., CLIP). Despite the open-set nature, these tasks still require predefined object categories during the inference stage. To enable object detection without predefined object categories during inference, they proposed a generative approach to open vocabulary object detection, formulating object detection as a generative problem(see Fig.3d) and introducing a simple framework called GenerateU. The model consists of two components: a visual object detector (Deformable DETR) to localize image regions, and a pretrained multimodal large language model (MLLM) responsible for converting visual regions into object names to generate pseudo-labels. GenerateU optimizes these two components through end-to-end training and achieves results comparable to GLIP on the LVIS dataset.

## 4. Summary of Existing Work

From the narrative in Sec. 3, we can clearly observe the developmental trajectory and key technological advancements in the field of Open Vocabulary Object Detection. Their performance can be found in Table 1. The evolution primarily focuses on addressing the limitations of closed category sets, leveraging multimodal data to enhance model generalization, and optimizing pseudo-label quality. The following is a summary categorized by technological development logic:

OVR-CNN marks the inception of open vocabulary detection by pretraining a visual encoder on a large-scale image-caption dataset, thereby learning generalized visual-semantic mappings. This approach enhances the object detection model's ability to recognize unknown categories,

Table 1. Representative works summarization and comparison in Sec. 3.

| Method | Dataset | Image Backbone | Detector | Text Encoder | $AP_{50}^{N}$ | $AP_{50}^{B}$ | $AP_{50}$ |
|---|---|---|---|---|---|---|---|
| **Image-caption Pair** | | | | | | | |
| OVR-CNN [20] | COCO | R50 | FRCNN | BERT | 22.8 | 46.0 | 39.9 |
| LocOV [3] | COCO | R50 | FRCNN | BERT | 28.6 | 51.3 | 45.7 |
| **Pseudo-Labeling** | | | | | | | |
| RegionCLIP [22] | COCO | R50 | FRCNN | CLIP | 31.4 | 57.1 | 50.4 |
| VL-PLM [21] | COCO | R50 | FRCNN | CLIP | 34.4 | 60.2 | 53.5 |
| MarvelOVD [1] | COCO | R50 | MRCNN | CLIP | 37.8 | 57.4 | 52.0 |
| GLIP [12] | COCO | Swin-L | DyHead | BERT | - | - | 60.8 |
| GLIP [12] | LVIS | Swin-L | DyHead | BERT | - | - | 26.8 |
| Grounding DINO [14] | COCO | R50 | DINO | BERT | - | - | 65.8 |
| LBP [11] | COCO | - | FRCNN | CLIP | 35.9 | 60.8 | 54.3 |
| **Knowledge Distillation** | | | | | | | |
| ViLD [6] | OV-COCO | R50 | MRCNN | CLIP | 27.6 | 59.5 | 51.3 |
| **Transfer Learning** | | | | | | | |
| F-VLM [10] | COCO | R50 | MRCNN | CLIP | 28.0 | - | 39.6 |
| **Generative** | | | | | | | |
| GenerateU [13] | LVIS | Swin-L | - | CLIP | - | - | 27.9 |

successfully transferring open vocabulary knowledge to the object detection task and overcoming traditional category limitations. LocOV further refines the utilization of image-caption data by introducing consistency regularization and local semantic matching (LSM), improving the alignment of image regions with textual features and increasing the accuracy of novel category recognition.

Subsequently, CLIP was proposed and open-sourced. Following this, ViLD leveraged the pretrained CLIP model for knowledge distillation, significantly boosting OVD performance. By employing a dual-branch structure and RPN score assistance, it optimized image-text alignment at the regional level, showcasing the immense potential of pretrained multimodal models. RegionCLIP addressed CLIP's limitations in regional recognition through contrastive learning and pseudo-label strategies, achieving fine-grained alignment of image regions with text, thus enhancing regional detection accuracy. GLIP unified object detection with phrase grounding, exploiting the richness of grounding data and deep feature fusion to markedly improve the model's understanding and generalization to new concepts. GroundingDINO upgraded the detector to a Transformer-based architecture, combining more efficient multimodal feature fusion and language-guided query selection to enhance the model's capability in handling unseen categories. LBP tackled the limitations in background proposal processing by discovering potential background categories, distinguishing background objects, and correcting probability scores, thereby improving the model's ability to differentiate between background and novel categories.

To better utilize pseudo-labels, VL-PLM and MarvelOVD respectively addressed the issues of low pseudo-label quality and domain shift. VL-PLM combined RPN scores with multi-stage adjustments, while MarvelOVD introduced an adaptive weighting mechanism and hierarchical label assignment to enhance noise suppression capabilities.

F-VLM revealed that a frozen CLIP model inherently possesses strong regional classification capabilities. By directly training the detector head on the frozen VLMs, it avoids knowledge distillation and pretraining, achieving efficient and high-performance OVD while significantly reducing training computational costs.

GenerateU transformed OVD into a generative problem, utilizing Deformable DETR for image region localization and a multimodal large language model for pseudo-label generation. This approach enabled object detection without predefined categories, demonstrating the potential of generative methods.

## 5. Challenges And Outlook

### 5.1. Challenges

Open vocabulary object detection (OVD) hinges on the semantic alignment between regional features and category vocabularies, necessitating finer granularity in training data and model predictions. The primary technical challenges include:

**Overfitting to Base Categories Weakening Generalization to New Categories**: A major challenge in OVD is ensuring that the model can generalize to detect categories not seen during training. This requires the model to learn representations that can be effectively applied to new categories. Due to the lack of annotations for new categories, the model may overfit to the base categories, making it difficult to recognize new ones effectively.

**Utilization of Weak Supervision**: Open vocabulary learning methods often rely on weak supervision signals, such as image-text pairs or pretrained vision-language models, involving information from both image and text modalities. Effectively performing cross-modal learning and integrating these weak signals into the detection pipeline to improve the detection of new categories is a significant challenge.

**Handling Large Vocabularies**: OVD systems need to manage and classify a vast, potentially infinite, number of object categories. In an open vocabulary setting, there can be significant class imbalance between new and base categories, which can impair the model's performance in recognizing new categories. Thus, the model needs to handle large-scale datasets and distinguish numerous categories based on semantic similarity.

**Evaluation Metrics and Datasets**: Evaluating OVD models is complex as it involves assessing the performance on both base and new categories. The overlapping concepts between categories (e.g., watermelon and fruit, person and child) necessitate the design of new metrics to better measure open vocabulary methods. Additionally, current datasets are still relatively small. More extensive datasets, such as OVDEval[19], are needed to achieve a true open vocabulary setting.

### 5.2. Future Work

**Combining with Large Language Models**. Compared with VLMs, most LLMs contain more text concepts, which naturally have a broader scope than various dataset taxonomies. Thus, how to better align the LLMs knowledge with visual detectors or segmenters to achieve stronger zero-shot results still needs exploration.

**Pseudo-Label Generation and Utilization**.In the context of open vocabulary learning, effectively utilizing weak supervision signals, especially through pseudo-label strategies, is crucial for expanding the model's vocabulary. Current pseudo-label strategies often rely on numerous but potentially noisy image-text pairs, requiring researchers to explore methods for generating higher quality pseudo-labels. Additionally, investigating how to dynamically adjust and utilize these pseudo-labels during training to maximize their positive impact while minimizing potential misguidance is also a key focus in this field.

**3D Open Vocabulary Scene Understanding**. Given the high cost of annotating point cloud data, particularly in dense prediction tasks, open vocabulary understanding in 3D scenes becomes especially important and urgent. Future research may delve deeper into mapping and extending the knowledge from 2D models to 3D spaces. This includes developing new geometric and semantic alignment strategies and exploring how to efficiently integrate cross-modal knowledge directly on point cloud data. Moreover, designing open vocabulary learning architectures tailored to the characteristics of 3D data and leveraging the inherent structure of point cloud data to enhance the model's ability to recognize unseen categories will also be key research areas in this field.

## 6. Conclusion

In this survey, I have examined the development of Open Vocabulary Detection (OVD). Initially, the paper introduces the definition of OVD, its related fields and tasks, and the background knowledge of large Visual Language Models (VLMs), setting the foundation for subsequent discussions. Next, I conduct a detailed analysis of twelve OVD methods, delving into their research motivations and key technologies, and establishing the hierarchical logical relationships between these methods. By categorizing the technical characteristics of these methods, I classify them into three main categories, elaborating on the unique aspects and applicable scenarios for each. In the experimental section, I provide a comprehensive description of the experimental setups and fairly compare the performance of these methods. Finally, I summarize the major challenges facing open vocabulary learning and highlight several promising future research directions.

## References


[1] Marvelovd: Marrying object recognition and vision-language models for robust open-vocabulary object detection. 5, 6

[2] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *European Conference on Computer Vision*, 2018. 2

[3] María Alejandra Bravo, Sudhanshu Mittal, and Thomas Brox. Localized vision-language matching for open-vocabulary object detection. In *German Conference on Pattern Recognition*, 2022. 3, 6

[4] Akshay Raj Dhamija, Manuel Günther, Jonathan Ventura, and Terrance E. Boult. The overlooked elephant of object detection: Open set. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1010–1019, 2020. 2

[5] Chuanxing Geng, Sheng-Jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (10):3614–3631, 2021. 2

[6] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2021. 3, 6

[7] Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5351–5359, 2019. 1

[8] Chao Jia, 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*, pages 4904–4916. PMLR, 2021. 2

[9] K. J. Joseph, Salman Hameed Khan, Fahad Shahbaz Khan, and Vineeth N. Balasubramanian. Towards open world object detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5826–5836, 2021. 2

[10] Weicheng Kuo, Yin Cui, Xiuye Gu, A. J. Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *ArXiv*, abs/2209.15639, 2022. 5, 6

[11] Jiaming Li, Jiacheng Zhang, Jichang Li, Ge Li, Si Liu, Liang Lin, and Guanbin Li. Learning background prompts to discover implicit knowledge for open vocabulary object detection. 2024. 4, 6

[12] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10955–10965, 2021. 4, 6

[13] Chuang Lin, Yi-Xin Jiang, Lizhen Qu, Zehuan Yuan, and Jianfei Cai. Generative region-language pretraining for open-ended object detection. *ArXiv*, abs/2403.10191, 2024. 5, 6

[14] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ArXiv*, abs/2303.05499, 2023. 4, 6

[15] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7, 2017. 2

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 2

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 1

[18] Walter J. Scheirer, Anderson Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35: 1757–1772, 2013. 2

[19] Yi Yao, 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. *ArXiv*, abs/2308.13177, 2023. 7

[20] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14388–14397, 2020. 3, 6

[21] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, Vijay Kumar B.G, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N. Metaxas. Exploiting unlabeled data with vision and language models for object detection. *ArXiv*, abs/2207.08954, 2022. 5, 6

[22] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel C. F. Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16772–16782, 2021. 4, 6