A Survey on Open-Vocabulary Detection and Segmentation: Past, Present, and Future

Chaoyang Zhu, Long Chen

Abstract—As the most fundamental scene understanding tasks, object detection and segmentation have made tremendous progress in deep learning era. Due to the expensive manual labeling cost, the annotated categories in existing datasets are often small-scale and pre-defined, *i.e.*, state-of-the-art fully-supervised detectors and segmentors fail to generalize beyond the closed vocabulary. To resolve this limitation, in the last few years, the community has witnessed an increasing attention toward **Open-Vocabulary Detection (OVD)** and **Segmentation (OVS)**. By "open-vocabulary", we mean that the models can classify objects beyond pre-defined categories. In this survey, we provide a comprehensive review on recent developments of OVD and OVS. A taxonomy is first developed to organize different tasks and methodologies. We find that the permission and usage of weak supervision signals can well discriminate different methodologies, including: *visual-semantic space mapping, novel visual feature synthesis, region-aware training, pseudo-labeling, knowledge distillation, and transfer learning*. The proposed taxonomy is universal across different tasks, covering object detection, semantic/instance/panoptic segmentation, 3D and video understanding. The main design principles, key challenges, development routes, methodology strengths, and weaknesses are thoroughly analyzed. In addition, we benchmark each task along with the vital components of each method in appendix and updated online at awesome-ovd-ovs. Finally, several promising directions are provided and discussed to stimulate future research.

Index Terms—Open-Vocabulary, Zero-Shot Learning, Object Detection, Image Segmentation, Future Directions

1 Introduction

Object detection and segmentation are core high-level scene perception tasks in computer vision. They are cornerstones of numerous real-world applications such as autonomous driving [1], [2], intelligent robotics [3], [4], to name a few. Given an image, video, or a set of point clouds, object detection [5], [6], [7] predicts tightly-enclosed bounding boxes along with class labels, while segmentation task groups pixels or points into a semantically coherent area or volume (semantic segmentation) [8], [9], an instance with a distinctive ID (instance segmentation) [10], or a combination of both (panoptic segmentation) [11].

The past decade has witnessed a steady and tremendous progress in object detection and segmentation tasks brought by CNN-based and Transformer-based models [5], [9], [10], [12], [13], [14]. However, existing detectors/segmentors can only localize pre-defined semantic concepts (labels) within a specific dataset. They are typically at small-scale, *e.g.*, 20 classes in Pascal VOC [15], 80 classes in COCO [16], even the largest LVIS [17] dataset merely annotates 1,203 categories. On the contrary, our human perception system can associate arbitrary visual concepts with open-ended class names or natural language descriptions. The closed-set limitation greatly hinders the utilization of current detectors/segmentors in real-world applications.

To resolve the closed-vocabulary constraint for scene perception tasks, research endeavors have been devoted to **zero-shot** or **open-vocabulary** detection and segmentation:

Zero-Shot. In the early stage, zero-shot detection (ZSD) [18], [19], [20] and segmentation (ZSS) [21], [22] are first proposed as an attempt *w/o allowing the access to any unannotated unseen visual object*. Typical methods always replace the learnable classifier (a fully-connected layer) with fixed *semantic embeddings*, *e.g.*, Word2Vec [23], GloVe [24], or BERT [25]. Semantic embeddigs can transfer knowledge from seen (base) categories to unseen (novel) ones. However, due to the unsupervised training on text corpus only, semantic embeddings lack the alignment with visual modality, *i.e.*, they are noisy to serve as anchors for calibrating visual space and impede performance improvement [26].

Open-Vocabulary. Open-vocabulary detection [26], [27] and segmentation [28], [29], [30] *allow the model to train on images with unannotated novel objects*. The closed-set limitation is addressed via weak supervision signals, *i.e.*, image-text pairs (image-caption pairs and image-level labels), or large pretrained Vision-Language Models (VLMs), such as CLIP [31]. VLMs text encoder (VLMs-TE) encodes template prompts [31] filled with class names into *text embeddings*[1], which are deemed as the frozen classifier. They well align with image modality due to the image-text contrastive pretraining [31]. Therefore, OVD and OVS achieve a huge performance leap compared to ZSD and ZSS.

Given the great application value, a plethora of methods have been proposed in recent years, making it hard for researchers to keep pace with them. However, to the best of our knowledge, only a few related surveys are available,

• *Chaoyang Zhu and Long Chen are with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Kowloon, Hong Kong.*
E-mail: sean.zhuh@gmail.com, longchen@ust.hk.
The corresponding author is Long Chen.

Manuscript received July 18, 2023; revised April 15, 2024.

1. Both text and semantic embeddings are label embeddings.

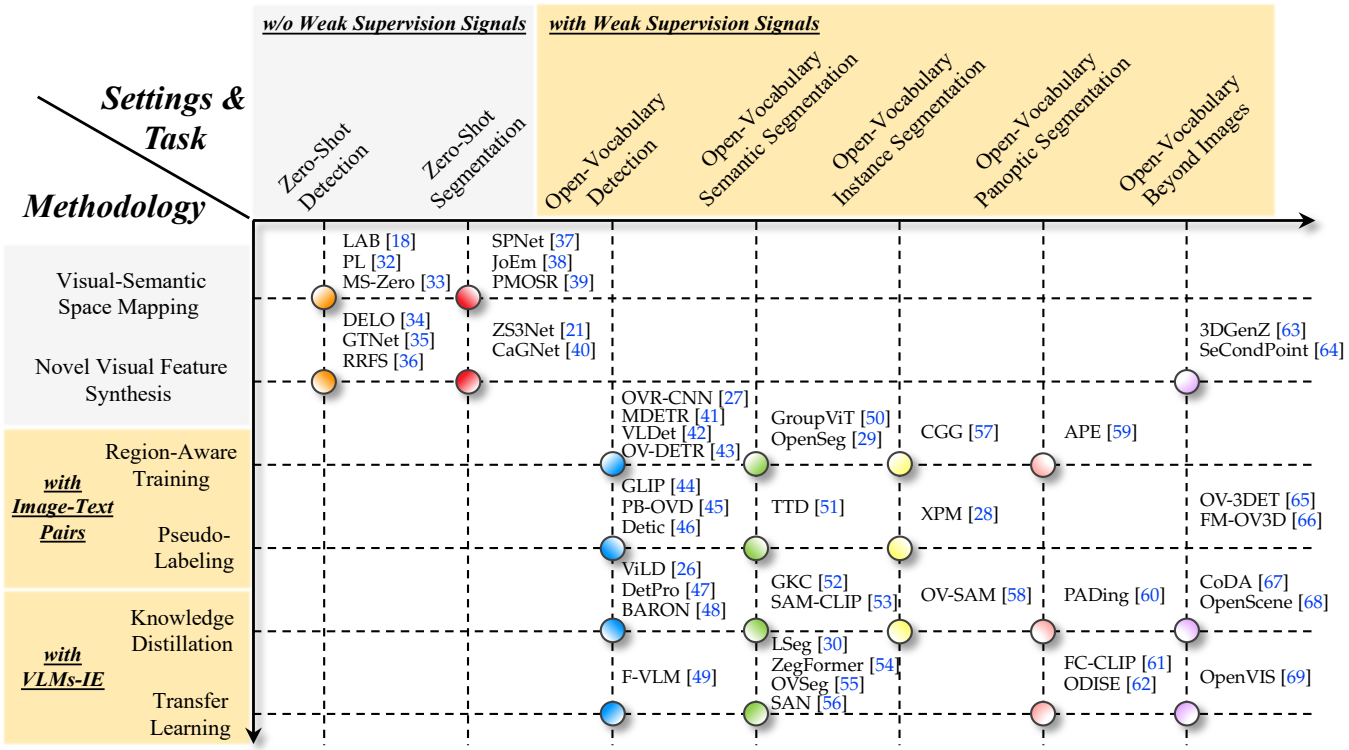

Fig. 1: The proposed taxonomy. Typical models are shown in each category. VLMs-IE denote the image encoder of VLMs.

focusing on limited tasks and settings[2]. We include zero-shot setting as a complement to open-vocabulary setting for two reasons: 1) both settings aim for resolving the closed-vocabulary constraint; 2) methodologies under the two settings are interchangeable, *e.g.*, the novel visual feature synthesis methodology (discussed later) under zero-shot setting can be transferred to open-vocabulary setting with negligible effort. The motivation for separating detection and segmentation task is that their definitions, training losses, architectures, evaluation metrics, and datasets are different. These tasks are also advanced separately over the past decade in literature, we discuss on unifying open-vocabulary detection and segmentation in Sec. 8.

In this paper, we provide a comprehensive review on different scene understanding tasks and settings including zero-shot/open-vocabulary detection, zero-shot/open-vocabulary semantic/instance/panoptic segmentation, as well as 3D scene and video understanding. To organize methods from these diverse tasks and settings, we need to answer the question: *How to build a taxonomy that differentiates zero-shot and open-vocabulary settings while in the meantime abstracts universal methodologies across tasks?* We find that, *whether or not to permit access to weak supervision signals, and if permitted, which one of them to utilize is key to categorization*. As shown in Fig. 1, zero-shot and open-vocabulary settings are differentiated by the permission of weak supervision signals, and different methodologies differ on which weak supervision signal to use during training. Under each setting, different tasks can share the same taxonomy.

Concretely, **ZSD** and **ZSS** are not allowed to access weak supervision signals. To generalize beyond seen objects, besides substituting the learnable classifier in closed-set detectors/segmentors with fixed semantic embeddings, the class-specific localizer is also switched to a class-agnostic one, *i.e.*, the output dimension of last regression layer is four ($[x_1, y_1, x_2, y_2]$ or $[x, y, w, h]$) instead of four times the number of test classes (see Figs. 2a and 2b). Methodologies under zero-shot setting can be grouped into:

Visual-Semantic Space Mapping. Though the visual and semantic space may bear discriminative capabilities in one modality, there is no direct cross-modality training mechanisms mining mutual relationships between both spaces. Thus, learning a mapping from visual to semantic space, semantic to visual space, or a joint mapping of visual-semantic space via tailored losses is crucial to enable such a reliable cross-space similarity measurement. However, due to the lack of unseen annotations, the prediction confidence is always biased toward seen classes.

Novel Visual Feature Synthesis. This methodology utilizes an additional generative model [71], [72], [73] to synthesize fake unseen visual features conditioned on semantic embeddings and random noises, which transfer the problem into a "fully-supervised" setting. The generation loss in Fig. 2c is to approximate the underlying distribution of real visual features. Then, the classifier embedded in the detector head is retrained on both pristine real seen and generated unseen visual features. Since non-seen regions are typically classified as background, this methodology alleviates both the confusion between novel and background concepts, and the bias issue in previous methodology. A more detailed pipeline is given in Fig. 3.

Once allowed to access the weak supervision signals, **OVD** and **OVS** methodologies can be mainly categorized into four types: **Region-Aware Training** mainly lever-

2. In this survey, "zero-shot" and "open-vocabulary" are regarded as two different settings following prior work.

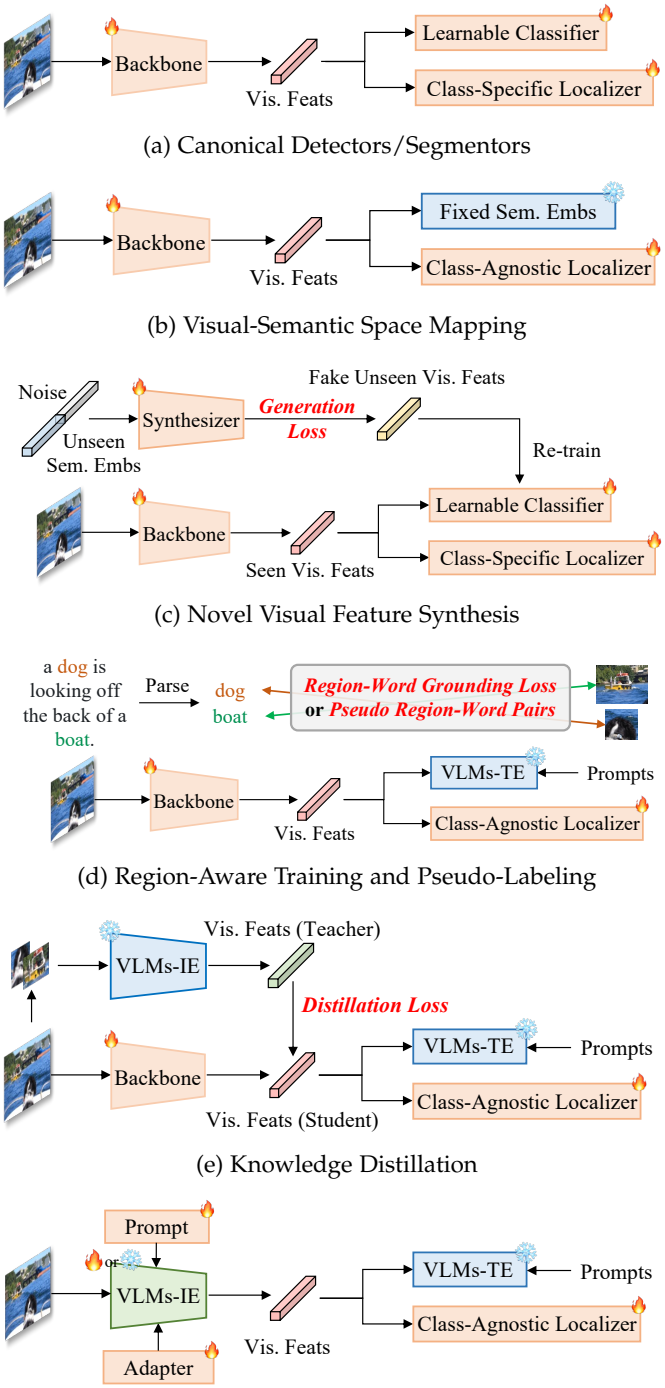

Fig. 2: A general comparison of each methodology. "Vis. Feats" and "Sem. Embs" are visual features and semantic embeddings [23], [24], [25], [70], respectively.

ages image-text pairs *w/o* VLMs image encoder (VLMs-IE); **Pseudo-Labeling** leverages both image-text pairs and VLMs-IE (optional); **Knowledge Distillation** and **Transfer Learning** mainly utilize VLMs-IE and seldomly trained on image-text pairs. Note that open-vocabulary setting also adopt a class-agnostic localization branch, and all four methodologies use VLMs-TE to encode class names into frozen text embeddings as the classifier. We now detail the four methodologies as the following.

Region-Aware Training. As Fig. 2d shows, regions and words are "*implicitly*" aligned on the cheap and abundant image-text pairs besides ground-truth annotations. Its main characteristic is the imposed bi-directional weakly-supervised grounding or contrastive loss [27] (see Eqs. (1) and (2)). The loss pulls regions and words within the same image-text pair close while pushing away other negatives inside a batch. However, the exact one-to-one correspondence between regions and words remains unknown due to its weakly supervised nature. Nonetheless, during training the model encounters a larger vocabulary containing novel classes, hence improving the base-to-novel generalization ability. Models leveraging bipartite matching [13] and region-level grounding annotations [74], [75], [76] are also grouped into this methodology.

Pseudo-Labeling. In contrast, pseudo-labeling knows the exact correspondence between words and regions. It "*explicitly*" constructs either pseudo region-word or region-caption pairs for novel classes in a teacher-student framework. It can be seen as a "**hard**" alignment, *i.e.*, one region can only correspond to one word, and vice versa. In contrast, region-aware training is "**soft**" alignment, where one word/region may correspond to multiple regions/words weighted by *softmax*. Pesudo-labeling can adopt VLMs-IE as teachers to produce pseudo labels, it can also be deemed as self-training w/o VLMs-IE (the teacher is the model itself). One defect of pseudo-labeling is that it requires knowing novel class names in advance during training which is impractical and breaks open-vocabulary setting.

Knowledge Distillation. VLMs (*e.g.*, CLIP) trained via contrastive learning yield superior zero-shot recognition ability across various downstream tasks. Shown in Fig. 2e, this methodology mainly distills region embeddings from VLMs-IE into a detector using only downstream detection/segmentation data. Since region-of-interests (RoIs) may contain novel objects, mimicking region embeddings of the teacher enables the detector to approach CLIP visual-semantic space more effectively. A more detailed framework is given in Fig. 4.

Transfer Learning. We term the usage of transferring a pretrained VLMs-IE to downstream perception tasks as transfer learning. In knowledge distillation, both the detector backbone and VLMs-IE are employed during training (increasing computational overhead and memory footprint) and the VLMs-IE is typically discarded in inference. In contrast, transfer learning utilizes only VLMs-IE in both training and test stages, adding negligible computation cost. Shown in Fig. 2f and Fig. 5, these methods can be further categorized into: 1) frozen VLMs-IE as feature extractor; 2) fine-tuning VLMs-IE on downstream data; 3) learning visual prompts [77] or a lightweight adapter [56], [78], [79] to the frozen VLMs-IE in a parameter-efficient way.

The above taxonomy can also be seamlessly applied to both open-vocabulary 3D scene and video understanding. Due to the lack of large-scale point clounds-text or video-text pairs, large VLMs in the image domain are exploited as an intermediate bridge between point clouds and texts or between videos and texts. For instance, pseudo-labels generated by OVD detectors in the 2D domain are associated with semantic labels via VLMs and they are back projected to 3D space as supervision.

TABLE 1: Differences of ZSD/ZSS and OVD/OVS. G and Non-G denote generalized and non-generalized evaluation.

External Information	ZSD/ZSS				OVD/OVS
	Inductive		Transductive		
	G	Non-G	G	Non-G	
Occurrence of Unlabeled & Novel Objects	✗	✗	✓	✓	✓
Image-Text Pairs	✗	✗	✗	✗	✓
Large VLMs	✗	✗	✗	✗	✓
Test Classes	$\mathcal{C}_B \cup \mathcal{C}_N$	$\mathcal{C}_N$	$\mathcal{C}_B \cup \mathcal{C}_N$	$\mathcal{C}_N$	$\mathcal{C}_B \cup \mathcal{C}_N$

The remainder of the paper is organized as follows: Sec. 2 introduces preliminary background. Then we review ZSD and ZSS in Sec. 3 and Sec. 4, OVD and OVS in Sec. 5 and Sec. 6, respectively. Open-vocabulary 3D scene and video understanding are covered in Sec. 7. We point out challenges and outlook in Sec. 8 and conclude in Sec. 9. Additional benchmarks with vital components of each method are listed in appendix.

2 Preliminaries

2.1 Problem Definition

The goal of OVD/OVS is to detect or segment unseen or novel classes that occupy semantically coherent regions or volumes within an image, video, or a set of point clouds. During its early stage of development, **inductive** ZSD/ZSS is first formulated to achieve this goal. ZSD/ZSS imposes a constraint that training images do not contain any unseen instance even if it is not annotated. To resolve the limitation, **transductive** ZSD [80] considers unannotated and unseen samples during training. However, inductive ZSD/ZSS has been the mainstream and more actively studied than transductive ZSD/ZSS. OVD/OVS can be deemed as a variant of transductive ZSD/ZSS that further allows the usage of weak supervision signals. Nonetheless, both zero-shot and open-vocabulary settings avoid annotated novel objects appearing in the training set. OVD/OVS splits the labeled set $\mathcal{C}$ of annotations into two disjoint subsets of base and novel categories. We denote them by $\mathcal{C}_B$ and $\mathcal{C}_N$, respectively. Note that $\mathcal{C}_B \cap \mathcal{C}_N = \varnothing$ and $\mathcal{C} = \mathcal{C}_B \cup \mathcal{C}_N$. Thus the labeled set for training and test is $\mathcal{C}_{train} = \mathcal{C}_B$ and $\mathcal{C}_{test} = \mathcal{C}_B \cup \mathcal{C}_N$. With this definition, the difference with closed-set tasks is clear, where $\mathcal{C}_{test} = \mathcal{C}_{train} = \mathcal{C}$. A complete comparison between ZSD/ZSS and OVD/OVS is listed in Table 1.

2.2 Related Domains and Tasks

In this subsection, we describe several highly related domains with OVD/OVS and summarize their differences.

Visual Grounding. It grounds semantic concepts to image regions [75], [76], [81], [82], [83]. Specifically, the task can be divided into: 1) phrase localization [82] that grounds all nouns in the sentence; 2) referring expression comprehension [75], [76], [81] that only grounds the referent in the sentence. In the latter, the referent is labeled not with a class name but with freeform natural language describing instance attributes, positions, and relationships with other objects or backgrounds. This task greatly expands the concepts that the model can ground but the vocabulary is still closed-set. An ideal way to achieve OVD and OVS is to scale the small-scale grounding datasets to web-scale datasets. However, the laborious labeling cost is non-negligible.

Weakly-Supervised Detection and Segmentation. Without any bounding box or dense mask annotation, training labels of weakly-supervised setting [84] only comprise image-level class names. The image-level labels indicate object classes that appear in the image. Many weakly-supervised learning techniques like multiple instance learning [85] and weakly-supervised grounding loss have been introduced into OVD/OVS. However, weakly-supervised object detection and segmentation work under the closed-set setting.

Open-Set Detection and Segmentation. Open-set detection [86], [87] and segmentation [88], [89] stem from open-set recognition [90], [91]. It requires classifying known classes and identifying a single "*unknown*" class without further classifying exact classes. It is equivalent to setting $\mathcal{C}_{train} = \mathcal{C}_B$ and $\mathcal{C}_{test} = \mathcal{C}_B \cup \{u\}$ where u represents the single "*unknown*" class. The main target is to reject unknown classes that emerge unexpectedly and may hamper the robustness of the recognition system.

Open-World Detection and Segmentation. Open-world detection [92], [93] and segmentation [94] take a step further to open-set detection and segmentation. At time step t, objects are classified as $\mathcal{C}_B^t = \{c_1, c_2, \ldots, c_k\} \cup \{u\}$. Then, unknown instances are labeled as newly known classes $\{c_{k+1}, \ldots, c_{k+m}\}$ by an oracle and added back to $\mathcal{C}_B^t$. At time t+1, detector is required to detect $\mathcal{C}_B^{t+1} = \{c_1, \ldots, c_k, \ldots, c_{k+m}\} \cup \{u\}$. After each time step, the number of unknown classes belonging to "*unknown*" will decrease. This continual learning cycle repeats over the lifetime of the detector. The task aims at incrementally learning new classes without forgetting previously learned classes in a dynamic world. Note that the detector is not re-trained from scratch across time steps.

Out-of-Distribution Detection. In out-of-distribution detection [95], [96], test samples are *not* assumed to be drawn from the same distribution of training data (*i.e.*, *i.i.d.*). Specifically, distribution shift can be divided into: 1) **Semantic shift**, where out-of-distribution samples are from different classes. 2) **Covariate shift**, where out-of-distribution samples are from different domains but with same classes, such as sketches or adversarial examples. Out-of-distribution detection primarily focuses on the former which resembles open-set recognition. Nonetheless, it still does not require sub-classifying the single "*out-of-distribution*" class which is the same as the "*unknown*" class in open-set/world detection and segmentation.

2.3 Canonical Closed-Set Detectors and Segmentors

Object Detection. Faster R-CNN [5] is a representative **two-stage** detector. Based on anchor boxes, the region proposal network (RPN) first hypothesizes potential object regions to separate foreground and background proposals by measuring their objectness scores. Then, an RCNN-style [97] detection head predicts per-class probability and refines the locations of positive proposals. Meanwhile, **one-stage** detectors directly refine the positions of anchors without proposal

generation stage. FCOS [98] regards each feature map grid within the ground-truth box as a positive anchor point and regresses its distances to the four edges of the target box. The centerness suppresses low-quality predictions of anchor points that are near the boundary of ground-truths. With the development of Transformers in NLP, Transformer-based detectors have dominated the literature recently. DETR [13] reformulates object detection as a set matching problem with a Transformer encoder-decoder architecture. The learnable object queries attend to encoder output via cross-attention and specialize in detecting objects with different positions and sizes. Deformable DETR [99] designs a multi-scale deformable attention mechanism that sparsely attends sampled points around queries to accelerate convergence.

Segmentation. DeepLab [100], [101] enhances FCN [8] with dilated convolution, conditional random field, and atrous spatial pyramid pooling. Mask R-CNN [10] adds a parallel mask branch to Faster R-CNN and proposes RoI Align for instance segmentation. Following DETR, MaskFormer [9] obtains mask embeddings from object queries, and performs dot-product with up-sampled pixel embeddings to produce segmentation maps. It transforms the per-pixel classification paradigm into a mask region classification framework. Mask2Former [14] follows the same meta-architecture of MaskFormer and introduces a masked cross-attention module that only attends to predicted mask regions.

2.4 Large Vision-Language Models (VLMs)

Large VLMs have demonstrated a superior zero-shot transfer capability benefited from large-scale pretraining, and found various applications in OVD/OVS. In this subsection, we review large VLMs and related fine-tuning techniques.

Image-Text Contrastive Pretraining. CLIP [31] makes the first breakthrough via contrastively pretraining on 400M image-caption pairs. The pretraining objective simply aligns positive image-caption pairs within a batch, making it efficient and scalable to learn transferable representations. During inference, template prompts filled with class names (*e.g.*, "a photo of a [CLASS]") are fed into the CLIP text encoder, and the output of a special token [EOS] is regarded as the text embedding. For image embedding, a multi-head self-attention pooling layer aggregates patch embeddings into a holistic representation. Both text and image embeddings are l_2 normalized to compute their pair-wise cosine similarity. By this means, text embeddings are regarded as a frozen classifier. Later, ALIGN [102] leverages one billion noisy image alt-text pairs for pretraining with the same architecture and contrastive loss in CLIP.

Self-Supervised Learning with Local Semantics. The work of DINO [103], [104] self-distills its knowledge on unlabeled images. Two different random transformations of the same input image are passed to the student and teacher network separately. The student receives both local and global views, while the teacher only receives global views of the image. By matching the predicted distributions of teacher and student, DINO encourages a local-to-global correspondence establishment. Different heads in the last self-attention block of DINO clearly separates different object boundaries, enabling unsupervised object localization and pseudo-labeling in OVD/OVS. Meanwhile, MAE [105] designs an asymmetric encoder-decoder framework, the encoder only perceives non-masked image patches and the decoder regresses the pixel value of masked patches. Since CLIP behaves like bag-of-words [106], *i.e.*, a bag of local objects matches a bag of semantic concepts without differentiating distinct local regions, MAE and DINO are mainly utilized to enhance local image feature representations for OVD/OVS.

Text-to-Image (T2I) Diffusion Models. T2I diffusion models [107], [108] are also trained on internet-scale data. The step-by-step de-noising process gradually evolves pure noise tensors into realistic images conditioned on languages. Clustering on its internal feature representations clearly separates different objects, which is much less noisy than CLIP [31], [109], [110]. Since to generate distinct objects the model has to differentiate semantic concepts, text-guided generation objective naturally imposes precise region-word correspondence learning for the model during pretraining. In contrast, CLIP may cheat by only performing bag-of-words classification [106].

Parameter-Efficient Fine-Tuning (PEFT). Given the computational cost and memory footprint of these large VLMs, current endeavors only fine-tune a small set of additional parameters, such as prompt tuning [77], [111], adapters [56], [78], [79]. Compared to fully fine-tuning the whole model on downstream data, PEFT balances the trade-off between overfitting on downstream data and preserving the prior knowledge of VLMs.

3 Zero-Shot Detection (ZSD)

Removing training images containing unannotated unseen objects induces the main challenge for ZSD. Hence, models resort to transfer knowledge of semantic embeddings [23], [24], [25], [70] that are unsupervisedly trained on text corpus alone from seen to unseen classes. Methods in this section can be discriminative or generative: 1) visual-semantic space mapping is discriminative that seeks to maximize separation between the decision boundaries of ambiguous classes in visual, semantic, or common embedding space, especially the background and unseen classes; 2) novel visual feature synthesis is generative that uses a synthesizer for generating unseen visual features to bridge the data scarcity gap between seen and unseen classes.

3.1 Visual-Semantic Space Mapping

3.1.1 Learning a Mapping from Visual to Semantic Space

This mapping assumes the intrinsic structure of the semantic space is discriminative and it can well reflect the inter-class relationships. Besides the two modifications made to canonical detectors in Fig. 2b, a linear layer (mapping function) is additionally added to the backbone to make the dimension of visual features the same as semantic embeddings.

Two-Stage Methods. At the time when ZSD is proposed, two-stage detectors [5] dominate closed-set object detection task. ZSD is first proposed by Bansal *et al.* [18]. RoI features in Faster-RCNN are linearly projected into semantic space driven by a max-margin loss, then classified by GloVe embeddings [24]. Bansal *et al.* propose two techniques to remedy the ambiguity between background and unseen concepts. One is a fixed label vector ([1,0,...,0]) for modeling background, which is hard to cope with the high background visual variances. The other is dynamically assigning

multiple classes from WordNet [112] belonging to neither seen nor unseen classes to the background. A contemporaneous work [19], [113] proposes a meta-class clustering loss besides the max-margin separation loss. It groups similar concepts to improve the separation between semantically-dissimilar concepts and reduce the noise in word vectors. Luo *et al.* [114] provide external relationship knowledge graph as pairwise potentials besides unary potentials in the conditional random field to achieve context awareness. ZS-DTD [115] leverages textual descriptions to guide the mapping process. The textual descriptions are a general source for improving ZSD due to its rich and diverse context compared to a single word vector. Following polarity loss [32], BLC [116] develops a cascade architecture to progressively learn the mapping with an external vocabulary and a background learnable RPN to model background appropriately. Rahman *et al.* [80] explore transductive generalized ZSD via fixed and dynamic pseudo-labeling strategies to promote training in unseen samples. Recently SSB [117] establishes a simple but strong baseline. It carefully ablates model characteristics, learning dynamics, and inference procedures from a myriad of design options.

One-Stage Methods. Besides two-stage methods, applying one-stage detectors [12], [118], [119] to ZSD is also explored. To improve the low recall rate of novel objects, ZS-YOLO [20] conditions the objectness branch of YOLO [119] on the combination of semantic attributes, visual features, and localization output instead of visual features alone. HRE [120] constructs two parallel visual-to-semantic mapping branches for classification. One is a convex combination of semantic embeddings, while the other maps grid features associated with positive anchors into the semantic space. Later Rahman *et al.* [32] design a polarity loss that explicitly maximizes the margin between predictions of target and negative classes based on focal loss [12]. A vocabulary metric learning approach is also proposed to provide a richer and more complete semantic space for learning the mapping. Similar to the hybrid branches in HRE [120], Li *et al.* [121] perform the prediction of super-classes and fine-grained classes in parallel.

3.1.2 Learning a Joint Mapping of Visual-Semantic Space

Learning a mapping from visual to semantic space neglects the discriminative structure of visual space itself. MS-Zero [33] demonstrates that classes can have poor separation in semantic space but are well separated in visual space, and vice versa. Hence it exploits this complementary information via two unidirectional mapping functions. Similarity metrics are calculated in both spaces which are then averaged as the final prediction. Similar to previous works [120], [121], DPIF [122] proposes a dual-path inference fusion module. It integrates empirical analysis of unseen classes by analogy with seen classes (past knowledge) into the basic knowledge transfer branch. The association predictor learns unseen concepts using training data from a group of associative seen classes as their pseudo instances. ContrastZSD [123] proposes to contrast seen region features to make the visual space more discriminative. It contrasts seen region features with both seen and unseen semantic embeddings under the guidance of semantic relation matrix.

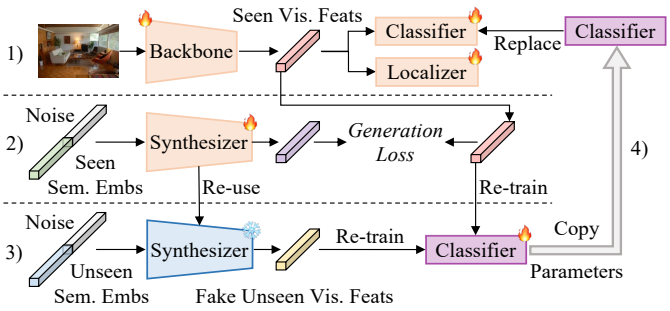

Fig. 3: Flowchart of novel visual feature synthesis.

3.1.3 Learning a Mapping from Semantic to Visual Space

Zhang *et al.* [124] argue that learning a mapping from visual to semantic space or a joint space will shrink the variance of projected visual features and thus aggravates the hubness problem [125], *i.e.*, the high-dimensional visual features are likely to be embedded into a low dimensional area of incorrect labels. Hence they embed semantic embeddings to the visual space via a least square loss.

3.2 Novel Visual Feature Synthesis

Overview. To enable recognition of novel concepts, novel visual feature synthesis produces *fake* unseen visual features as training samples for a new classifier. This methodology follows a multi-stage pipeline (as shown in Fig. 3): 1) Train the base model with seen classes annotations in a fully-supervised manner. 2) Train the feature synthesizer $G : \mathcal{W} \times \mathcal{Z} \mapsto \tilde{\mathcal{F}}$ on seen semantic embeddings $\mathbf{w} \in \mathcal{W}_s \in \mathbb{R}^d$ and real seen visual features $\mathbf{f}_s \in \mathcal{F}_s \in \mathbb{R}^c$ extracted from the base model to learn the underlying distribution of visual features. 3) Conditioned on the unseen semantic embeddings $\mathbf{w} \in \mathcal{W}_u$ and a random noise vector $\mathbf{z} \sim \mathcal{N}(0, 1)$, the synthesizer generates novel unseen visual features. A new classifier is retrained on fake unseen and real seen visual features, while the remaining parts of the base model are kept frozen. 4) Finally, the new classifier is plugged back into the base model. Note that the noise vector perturbs the synthesizer to produce various visually diverging features given the semantic embeddings.

DELO [34] leverages conditional variational auto-encoder [73] with three consistency losses forcing the generated visual features to be coherent with the original real ones on the predicted objectness score, category, and class semantic. Then, it [34] retrains the objectness branch to assign high confidence scores on both seen and unseen objects. This is to mitigate the low recall rate of unseen objects. Later on, Hayat *et al.* [126] use the same class consistency loss but they adopt the mode seeking regularization [127] which maximizes the distances of generated data points *w.r.t* their noise vectors. At the same time with DELO [34], GTNet [35] proposes an IoU-aware synthesizer based on wasserstein generative adversarial network [128]. Since DELO is only trained on ground-truths that encloses the object boundary tightly, so its synthesizer can not generate unseen RoI features with diverse spatial context clues. That is, the retrained classifier can not correctly classify RoIs that loosely encloses the object boundary. To mitigate this context gap between unseen RoI features from RPN and those synthesized by

the generator, GTNet [35] randomly samples foreground and background RoIs as the additional generation target, thus making the new classifier robust to various degrees of context. RRFS [36] proposes an intra-class semantic diverging loss and an inter-class structure preserving loss. The former pulls positive synthesized features lying within the hypersphere of the corresponding noise vector close while pushing away those generated from distinct noise vectors. The latter constructs a hybrid feature pool of real and fake features to avoid mixing up the inter-class relationship.

4 Zero-Shot Segmentation (ZSS)

ZSS takes a step further than ZSD at a finer pixel-level granularity. We cover zero-shot semantic and instance segmentation in this section as a complement to ZSD.

4.1 Zero-Shot Semantic Segmentation (ZSSS)

4.1.1 Visual-Semantic Space Mapping

Learning a Mapping from Visual to Semantic Space. SPNet [37] is the first work that proposes the zero-shot semantic segmentation task. It directly maps pixel features into semantic space optimized by the canonical *cross-entropy* loss. During inference, SPNet calibrates seen predictions by subtracting a factor tuned on a held-out validation set.

Learning a Joint Mapping of Visual-Semantic Space. Hu *et al.* [129] argue that the noisy and irrelevant samples in seen classes have negative effects on learning visual-semantic correspondence. An uncertainty-aware loss is proposed to adaptively strengthen representative samples while an attenuating loss for uncertain samples with high variance estimation. JoEm [38] learns a joint embedding space via the proposed boundary-aware regression loss and semantic consistency loss. At the test stage, semantic embeddings are transformed into semantic prototypes acting as a nearest-neighbor classifier (without the classifier retraining stage in Sec. 4.1.2). The apollonius calibration inference technique is further proposed to alleviate the bias problem.

Learning a Mapping from Semantic to Visual Space. Kato *et al.* [130] propose a variational mapping from semantic space to visual space via sampling the conditions (mimicking the support images in few-shot semantic segmentation [131]) from the predicted distribution. PMOSR [39] abstracts a set of seen visual prototypes, then trains a projection network mapping seen semantic embeddings to these prototypes. Similar to JoEm [38], since one can simply project unseen semantic embeddings to unseen prototypes for classification, new unseen classes can be flexibly added in inference without classifier retraining. An open-set rejection module is further proposed to prevent unseen classes from directly competing with seen classes.

4.1.2 Novel Visual Feature Synthesis

Concurrent with SPNet [37], ZS3Net [21] conditions the synthesizer [71] on adjacency graph encoding structural object arrangement to capture contextual cues for the generation process. CSRL [132] transfers the relational structure constraint in the semantic space including point-wise, pair-wise, and list-wise granularities to the visual feature generation process. However, in both methods, the **mode collapse problem**, *i.e.*, the generator often ignores the random noise vectors appended to the semantic embeddings and produces limited visual diversity, hindering the effectiveness of generative models. CaGNet [40] addresses this problem by replacing the simple noise with contextual latent code, which captures pixel-wise contextual information via dilated convolution and adaptive weighting between different dilation rates. Following CaGNet [40], SIGN [133] also substitutes the noise vector but with a spatial latent code incorporating the relative positional encoding. While previous ZS3Net [21] and CaGNet [40] simply discard pseudo-labels whose confidence scores are below a threshold and weight the importance of the remaining pseudo-labels equally, SIGN [133] utilizes all pseudo annotations but assigns different loss weights according to the confidence scores of pseudo-labels.

4.2 Zero-Shot Instance Segmentation (ZSIS)

Zheng *et al.* [22] are the first to propose the task of zero-shot instance segmentation. They establish a simple mapping from visual features to semantic space then classify them using fixed semantic embeddings. The mapping is optimized by a mean-squared error reconstruction loss. Zheng *et al.* also argue that disambiguation between background and unseen classes is crucial [18], [116], they design a background-aware RPN and a synchronized background strategy to adaptively represent background.

5 Open-Vocabulary Detection (OVD)

OVD removes the stringent restriction of inductive ZSD on the absence of unannotated novel objects as in Sec. 3. From this section on, we discuss methodologies resorting to weak supervision signals, *i.e.*, the open-vocabulary setting in Sec. 1 and Sec. 2.

5.1 Region-Aware Training

This methodology incorporates image-text pairs [134], [135], [136] into detection training phase. The vast vocabulary $\mathcal{C}_T$ in captions encompasses both $\mathcal{C}_B$ and $\mathcal{C}_N$, thus aligning proposals containing novel proposals with words containing novel classes improves classification on $\mathcal{C}_N$.

Weakly-Supervised Grounding or Contrastive Loss. This line of work establishes a coarse and soft correspondence between regions and words via the average of the following two symmetrical losses:

$$\mathcal{L}_{T\to I} = -\log \frac{\exp(sim(I,T))}{\sum_{I'\in\mathcal{B}} \exp(sim(I',T))}, \tag{1}$$

$$\mathcal{L}_{I\to T} = -\log \frac{\exp(sim(I,T))}{\sum_{T'\in\mathcal{B}} \exp(sim(I,T'))}, \tag{2}$$

where $\mathcal{B}$ represents the image-text batch. The similarity $sim(I,T)$ between image I and caption T is given by:

$$sim(I,T) = \frac{1}{N_T}\sum_{i=1}^{N_T}\sum_{j=1}^{N_I} \alpha_{i,j}\langle e_i^T \cdot e_j^I\rangle, \tag{3}$$

where N_T and N_I are the number of nouns in the caption and the number of proposals in the image, respectively. e_i^T

and e_j^I are the text and region embedding typically encoded by the CLIP text encoder and detection head. The weight $\alpha_{i,j}$ is calculated by:

$$\alpha_{i,j} = \frac{\exp\langle e_i^T \cdot e_j^I \rangle}{\sum_{j'=1}^{N_I} \langle e_i^T \cdot e_{j'}^I \rangle}. \quad (4)$$

By minimizing the distance of matched image-caption pairs, novel proposals and novel classes are aligned in a weakly-supervised manner (*w/o* knowing the correspondence between proposals and words). OVR-CNN [27] first formulates the OVD task. Previous ZSD methods only train the vision-to-language projection layer from scratch on base classes, which is prone to overfitting. OVR-CNN learns the projection layer during pretraining on image-caption pairs using Eq. (1) and Eq. (2) to align image grids and words. Following OVR-CNN, LocOv [137] introduces a consistency loss that regularizes image-caption similarities over a batch to be the same before and after the multi-modal fusion transformer. LocOv utilizes both region and grid features for measuring Eq. (3) while OVR-CNN only adopts the latter. However, the multi-modal masked language modeling [138], [139], [140], [141] objective in OVR-CNN and LocOv tends to attend to similar global proposals covering many concepts for distinct masked words. To force the multi-modal fusion transformer focus more on exclusive proposals containing only one concept for different masked words, MMC-Det [142] drives the attention map over proposals be divergent for different masked words via the proposed divergence loss. DetCLIP [143] adds per-category definition from WordNet [112] for CLIP text encoder to encode richer semantics. The category names and definitions are individually and parallelly encoded to avoid unnecessary interactions between category names. DetCLIPv2 [144] selects a single region that best fits the current word via *argmax* instead of aggregating all region features via softmax in Eq. (4). It excludes the $\mathcal{L}_{I\rightarrow T}$ in its bi-directional loss due to the partial labeling problem, *i.e.*, the caption usually only describes salient objects in the image, hence most proposals can not find their matching words in the caption. WSOVOD [145] recalibrates RoI features via input-conditional coefficients over dataset attribute prototypes to de-bias different distributions in different datasets. It employs multiple instance learning [85] to address the lack of box annotations.

Previous work perform weakly-supervised grounding only on relatively small-scale image-text pairs, another series of work pretrains the model on web-scale image-text datasets. RO-ViT [146] is pre-trained from scratch on the same dataset of ALIGN [102] using focal loss [12] instead of cross-entropy loss to mine the hard negative examples. The positional embeddings are randomly cropped and resized to the whole-image resolution during pretraining, causing the model to regard pretrained images not as full images but as region crops from some unknown larger images. This matches the usage of proposals in the detection finetuning stage as they are cropped from a holistic image. To improve local feature representation for localization, CFM-ViT [147] adds the masked autoencoder objective [105] besides the bi-directional contrastive loss during pretraining on ALIGN [102] dataset. It randomly masks the positional embeddings during pretraining to achieve a similar effect of RO-ViT [146]. RO-ViT and CFM-ViT only pretrain the image encoder, while detection-specific components such as FPN [6] and RoI head [5] are randomly initialized and trained only on detection data. To bridge the architecture gap between pretraining and detection finetuning, DITO [148] pretrains FPN [6] and RoI head [5] along with image encoder. Embeddings of randomly sampled regions are max pooled and image-text contrastive loss is applied on each FPN level separately.

Bipartite Matching. VLDet [42] formulates region-word alignment as a set-matching problem between regions and nouns. The matching is automatically learned on image-caption pairs via the off-the-shelf Hungarian algorithm [13]. Following VLDet, GOAT [149] mitigates the biased objectness score by comparing region features with an open corpus of external object concepts as another assessment of objectness. GOAT reconstructs the base classifier by taking a weighted mean of top-k similar external concept embeddings *w.r.t.* a base concept to enhance generalization. OV-DETR [43] conditions object queries on concept embeddings. They are constructed either from text embeddings or image embeddings by feeding base ground-truth boxes and novel proposals to the CLIP image encoder. It reformulates the set matching problem into conditional binary matching, which measures the matchability between detection outputs and the conditional object queries. However, the conditioned object queries are class-specific, *i.e.*, the number is linearly proportional to the number of classes. Prompt-OVD [150] addresses this slow inference speed of OV-DETR by prepending class prompts instead of repeatedly adding to object queries and changing the binary matching objective to a multi-label classification cost. It further proposes RoI-based masked attention and RoI pruning to extract region embeddings in just one forward pass of CLIP. CORA [151] learns region prompts following the addition variant of VPT [77] to adapt CLIP into region-level classification. The proposed anchor pre-matching makes object queries class-aware and can avoid repetitive per-class inference in OV-DETR [43]. Based on OV-DETR [43], EdaDet [152] preserves fine-grained and generalizable local image semantics for attaining better base-to-novel generalization.

Leveraging Visual Grounding Datasets. Only resorting to image-level captions brings noise and misalignment in region-word correspondence learning. This line of work leverages ground-truth region-level texts to help mitigate this problem. Each region-level description may contain one object (phrase grounding [74], [82]) or multiple objects with a subject (referring expression comprehension and segmentation [75], [76]). These datasets are combined into one dataset termed GoldG in MDETR [41]. It proposes soft token prediction to predict the span of tokens in these texts and performs contrastive loss at the region-word level in the latent feature space. Note that GLIP [44] in Sec. 5.2 also uses the GoldG dataset but its main purpose is to generate reliable pseudo labels for subsequent self-training. MAVL [153] improves MDETR by multi-scale deformable attention [99] and late fusion between image and text modality. MQ-Det [154] augments language queries in GLIP [44] with fine-grained vision exemplars in a gated residual-like manner. It takes vision queries as keys and values to the class-specific

cross-attention layer. The vision-conditioned masked language prediction forces the model to align with vision cues to reduce the learning inertia problem [154]. YOLO-World [155] follows the formulation of GLIP and mainly focuses on equipping the YOLO series with open-vocabulary detection capability. SGDN [156] also leverages grounding datasets [74], [82], but it exploits additional object relations in a scene graph to facilitate discovering, classifying, and localizing novel objects.

5.2 Pseudo-Labeling

Models advocating pseudo-labeling also leverage abundant image-text pairs as in Sec. 5.1, but additionally, they adopt large pretrained VLMs or themselves (via self-training) to generate pseudo labels. They need to know the exact novel categories $\mathcal{C}_N$ in the training stage. Detectors are then trained on the unification of base seen class annotations and pseudo labels. According to the type and granularity of pseudo labels, methods can be grouped into: pseudo region-word pairs, region-caption pairs, and pseudo captions.

Pseudo Region-Word Pairs. The bi-directional grounding or contrastive loss in Sec. 5.1 allows one region/word to correspond to multiple words/regions weighted by softmax. However, this line of work explicitly allows only one region/word to correspond to one word/region (*i.e.*, hard alignment). RegionCLIP [157] leverages CLIP to create pseudo region-word pairs to pretrain the image encoder of the detector. *However, proposals with the highest CLIP scores yield low localization performance*. Targeting at this problem, VL-PLM [158] fuses CLIP scores with objectness scores and repeatedly applies the RoI head to remove redundant proposals. GLIP [44] reformulates object detection into phrase grounding and trains the model on the union of detection and grounding data. It enables the teacher to utilize language context for grounding novel concepts, while previous pseudo-labeling methods only train the teacher on detection data. GLIPv2 [159] further reformulates visual question answering [160], [161] and image captioning [162], [163] into a grounded VL understanding task. Following GLIP [44], Grounding DINO [164] upgrades the detector into a Transformer-based one, enhancing the capacity of the teacher model. Instead of generating pseudo labels once, PromptDet [165] iteratively learns region prompts and sources uncurated web images in two rounds, leading to more accurate pseudo boxes in the second round. Similar to semi-supervised detection [166], [167], SAS-Det [168] also proposes to refine the quality of pseudo boxes online. The student is optimized on pseudo boxes and base ground-truths. Then, the teacher updates its parameters through the exponential moving average of the student. Thereby the quality of pseudo boxes is gradually improved during training. Previous approaches adopt a simple thresholding [44], [159], [164] or a top-scoring heuristic rule [157], [165], [169], [170] to filter pseudo labels, which is vulnerable and lacking explainability. PB-OVD [45] employs GradCAM [171] to compute the activation map in the cross-attention layer of ALBEF [141] *w.r.t.* an object of interest in the caption. Then, all proposals that overlap the most with the activation map are regarded as pseudo ground-truths. CLIM [172] combines multiple images into a mosaicked image with each image treated as a pseudo region, and the paired caption is deemed as the region label. CLIM can be applied to many open-vocabulary detectors [27], [46], [48]. To generate more accurate pseudo labels, VTP-OVD [173] introduces an adapting stage to enhance the alignment between pixels and categories via learnable visual [77] and textual prompts. ProxyDet [174] argues that, pseudo-labeling can not improve novel classes other than those defined in $\mathcal{C}_N$. It finds that many novel classes reside in the convex hull constructed by base classes in the CLIP visual-semantic embedding space. These novel classes can be approximated via a linear mixup between a pair of base classes. Hence, ProxyDet synthesizes region and text embeddings of proxy (fake) novel classes and aligns them to achieve a broader generalization ability beyond those pseudo-labeled novel classes. CoDet [175] builds a concept group comprising all image-text pairs that mention the concept in their captions. The most common object appearing in the group is assumed to match the concept. Hence it reduces the problem of modeling cross-modality correspondence (region-word) to in-modality correspondence (region-region). The new formulation requires finding the most co-occurring objects by utilizing the fact that VLMs-IE exhibits feature consistency for visually similar regions across images.

Pseudo Region-Caption Pairs. Models in this category establish pseudo correspondence between the whole caption and a single image region, which is easier and less noisy compared to the more fine-grained region-word correspondence. Contrary to weakly-supervised detection that propagates image-level labels to corresponding proposals, Detic [46] side-steps this error-prone label assignment process. It simply trains the max-size proposal to predict all image-level labels, similar to multiple instance learning [85] and multi-label classification. The max-size proposal is assumed to be big enough to cover all image-level labels. Thus, the classifier encounters various novel classes during training on ImageNet21K [176] and it can be generalized to novel objects at inference. Building on top of Detic, Kaul *et al.* [177] employ a large language model, *i.e.*, GPT-3 [178], to generate rich descriptions for text embeddings. Besides, vision exemplars are encoded and merged with text embeddings to incorporate in-modality classification clues. 3Ways [169] regards the top-scoring bounding box per image as correspondence to the whole caption. It also augments text embeddings to avoid overfitting and includes trainable gated shortcuts to stabilize training. PLAC [179] learns a region-to-text mapping module that pulls close image regions with the corresponding caption. Then, novel proposals are matched to these region embeddings containing novel semantics.

Pseudo Captions. PCL [180] proposes to generate another type of pseudo label, *i.e.*, pseudo captions describing objects in natural language instead of generating bounding boxes. It leverages an image captioning model to generate captions for each object, which are then fed into the CLIP text encoder, encoding the class attributes and relationships with the surrounding environment. It can be regarded as better prompting compared to the template prompts in CLIP.

5.3 Knowledge Distillation

Knowledge distillation methodology (*c.f.* Fig. 4) employs a teacher-student paradigm where the student learns to

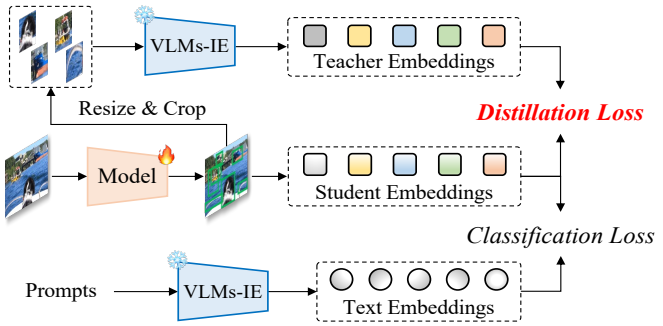

Fig. 4: A basic pipeline of knowledge distillation methodology. Distillation loss is typically a $\mathcal{L}_1$ loss. We omit the localization branch for brevity.

mimic visual knowledge from the teacher image encoder, thus approaching the well-aligned visual-semantic space of the teacher more effectively. Current methods mainly vary on the granularity of the distillation region and type of distillation objective.

Distillation at Single-Region Level. Depending on the IoUs with base ground-truths and objectness scores, proposals can be categorized into base, novel, and background proposals. Two pioneering works ViLD [26] and ZSD-YOLO [181] distill CLIP visual embeddings into Faster R-CNN and YOLOv5 detectors, respectively. ViLD first verifies that RPN well localizes novel proposals, regardless of trained either only on base annotations or the union of base and novel annotations. Hence, mimicking teacher visual embeddings of both base and novel proposals instead of only base ones improves the alignment between novel objects and novel classes. ZSD-YOLO also confirms the importance of applying distillation on these generalizable proposals instead of only on base ground truths. Though student visual embeddings resemble the teacher to an extent after distillation, there is still a gap between them. ViLD further ensembles the predictions of teacher and student via a weighted geometric mean during inference. The ensemble trusts teacher over student on novel predictions and vice versa on base predictions. Therefore it alleviates the overfitting of students to base classes, and it is inherited in many subsequent works. LP-OVOD [182] employs a sigmoid focal loss [12] (vs. softmax in ViLD) and constructs top relevant proposals of novel classes for training a novel classifier. In contrast to ViLD, EZSD [183] argues that RPN is biased toward base classes. It regards predefined anchor boxes tiled over the image and then these boxes are filtered by CLIP score over base and novel classes as distillation regions. EZSD finetunes normalization layers of CLIP on detection data, then distills from this adapted CLIP. SIC-CADS [184] designs a multi-label recognition ranking loss for text alignment instead of the cross-entropy loss in previous work. The image-level scores are then used to weight instance-level scores to infer which objects may exist in the current image.

Distillation at Bag-of-Regions Level. Distilling single region consisting of at most one object into student discards the knowledge of co-occurrence of visual concepts and their compositional structure implicitly captured in the CLIP image encoder. To harness this knowledge, BARON [48] samples nearby regions to form multiple groups of bag-of-regions. These bag-of-regions embeddings are encoded by CLIP text encoder then contrasted with the image crops enclosing the bag-of-regions via InfoNCE loss [185].

Distillation at Image-Region Level. Beyond region level distillation, OADP [186] introduces a distillation pyramid that distills block-wise patches and a box covering the whole image. Compared to only using a single region of downstream images, this encourages the student to recover the full visual-semantic space of CLIP. GridCLIP [187] also aligns global image embeddings between two CLIP image encoders with one trainable and the other frozen.

Distillation Objective Beyond $\mathcal{L}_1$ Loss. Rasheed *et al.* [170] propose an inter-embedding relationship matching loss that forces student embeddings to share the same inter-embedding similarity structure as teacher embeddings. DK-DETR [188] argues that the $\mathcal{L}_1$ loss constrains each element of student and teacher visual embeddings to be the same which is too strict and poses training difficulty. It instead maximizes the similarity between student and teacher visual embeddings of the same object and pushes them far away for different objects via a binary cross-entropy loss. Since the caption consists of novel nouns, and the image consists of novel objects, HierKD [189] contrasts the caption representations from the teacher with the ones from the student to improve base-to-new generalization. Another representative work DetPro [47] focuses on learning continuous context prompts [111] following ViLD via a cross-entropy loss. Concretely, DetPro forces all novel and background proposals equally unlike any base class, thus learned context prompts avoid classifying novel proposals into base classes. In addition, DetPro divides foreground proposals into disjoint groups according to their IoUs *w.r.t.* ground-truths. Prompt representations for each group are learned separately to describe different levels of contexts accurately, *e.g.*, the learned prompts may be "A partial photo of [CLASS]." or "A complete photo of [CLASS]". CLIP-Self [110] discovers that, in terms of classification accuracy, for ViT-based CLIP image encoders, using RoIAlign [10] to crop the ground-truths directly on the dense feature maps performs much inferior than forwarding the image crops to CLIP image encoder. In fact, F-VLM [49] only applied to ResNet-based CLIP image encoders. CLIPSelf rectifies this issue for ViT-based CLIP via a cosine self-distillation loss between the holistic image crop representation and patch-level dense presentations.

5.4 Transfer Learning

Knowledge distillation separates VLMs-IE and detector backbone during training, while transfer learning-based models discard the detector backbone and only utilize VLMs-IE. For example, fine-tuning VLMs-IE on detection data, or extracting visual features via the frozen VLMs-IE. OWL-ViT [190] removes the final token pooling layer of the CLIP image encoder and attaches a lightweight detection head to each Transformer output token. Then it fine-tunes the whole model on detection data through a bunch of dedicated finetuning techniques in an end-to-end manner. UniDetector [191] also trains the whole model, the image backbone is initialized with RegionCLIP [157]. During training, UniDetector leverages images of multiple

sources and heterogeneous label spaces. During inference, it calibrates the base and novel predictions via a class-specific prior probability recording how the network biases towards that category. Instead of fine-tuning the whole model, F-VLM [49] leverages the frozen CLIP image encoder to extract features and only trains the detection head. It ensembles predictions of the detector and CLIP via the same geometric mean (dual-path inference) as ViLD [26]. Another line of work only employs the CLIP text encoder for open-vocabulary classification and discards the CLIP image encoder. ScaleDet [192] unifies the multi-dataset label spaces by relating labels with sematic similarities across datasets. OpenSeeD [193] unifies OVD and OVS in one network. It proposes decoupled foreground-background decoding and conditioned mask decoding to compensate for task and data discrepancies, respectively. CRR [194] analyzes whether the classification in RoI head network hampers the generalization ability of RPN and proposes to decouple them, *i.e.*, RPN and RoI head do not share the backbone and they are separately trained. Sambor [195] introduces a ladder side adapter which assimilates localization and semantic prior from SAM and CLIP simultaneously. The automatic mask generation in SAM [196] is employed to enhance the robustness of class-agnostic proposal generation in RPN.

6 Open-Vocabulary Segmentation (OVS)

In this section, we review semantic, instance, and panoptic segmentation tasks using the same taxonomy in Sec. 5.

6.1 Open-Vocabulary Semantic Segmentation (OVSS)

6.1.1 Region-Aware Training

Weakly-Supervised Grounding or Contrastive Loss. Using the same Eq. (2) and Eq. (1), OpenSeg [29] randomly drops each word to prevent overfitting. SLIC [197] incorporates local-to-global self-supervised learning in DINO [103], [104] to improve the quality of local features for dense prediction.
Learning from Natural Language Supervision Only. Methods fall into this category aim to learn transferrable segmentation models purely on image-text pairs without densely annotated masks. GroupViT [50] progressively groups segment tokens into arbitrary-shaped and semantically-coherent segments given their assigned group index via gumbel-softmax [198]. Segment tokens in the last layer are average pooled and contrasted with captions like CLIP [31]. Besides the image-text contrastive loss, ViL-Seg [199] incorporates the same local-to-global correspondence learning in DINO [103] via a multi-crop strategy [200] to capture fine-grained semantics. An online clustering head trained by mutual information maximization [201] groups pixel embeddings at the end of ViT. Following GroupViT, SegCLIP [202] enhances local feature representation with additional MAE [105] objective and a superpixel-based Kullback-Leibler loss. OVSegmentor [203] adopts slot attention [204] to bind patch tokens into groups. To ensure visual invariance across images for the same object, OVSegmentor proposes a cross-image mask consistency loss. CLIP ViT-based image encoder performs worse on patch-level classification. To remedy this issue, PACL [205] adds an additional patch-text contrastive loss between patch embeddings and the caption. TCL [206] designs a region-level text grounder that produces text-grounded masks containing objects only described in the caption without irrelevant areas. Then the matching is performed on the grounded image region and text instead of the whole image and text. SimSeg [207] points out that CLIP heavily relies on contextual pixels and contextual words instead of entity words. Hence, instead of densely aligning all image patches with all words, SimSeg sparsely samples a portion of patches and words used for the bi-directional contrastive loss in Eq. (2) and Eq. (1).

6.1.2 Pseudo-Labeling

Zabari *et al.* [51] leverage an interpretability method [208] to generate a coarse relevance map for each category. The relevance map is refined by test-time-augmentation techniques (*e.g.*, horizontal flip, contrast change, and crops). The synthetic supervision is generated from the refined relevance maps using stochastic pixel sampling.

6.1.3 Knowledge Distillation

GKC [52] enriches the template prompts with synonyms from WordNet [112] instead of relying only on a single category name to guess what the object looks like. The text-guided knowledge distillation transfers the inter-class distance relationships in semantic space into visual space, which is similar to [170]. SAM-CLIP [53] merges the image encoders of SAM [196] and CLIP [31] into one via cosine distillation loss in a memory replay and rehearsal way on a subset of their pretraining images. It combines the superior localization ability of SAM and semantic understanding ability of CLIP. ZeroSeg [209] builds on top of MAE [105] and GroupViT [50]. Unlike GroupViT which requires text supervision, ZeroSeg distills the multi-scale CLIP image features into learnable segment tokens purely on unlabled images. The image is divided into multiple views thus capturing both local and global semantics.

6.1.4 Transfer Learning

This methodology aims to transfer the VLMs-TE and VLMs-IE to segmentation tasks. The transfer strategy is explored in the following aspects: 1) only adopting VLMs-TE for open-vocabulary classification; 2) leveraging frozen VLMs-IE as a feature extractor; 3) directly fine-tuning VLMs-IE on segmentation datasets; 4) employing visual prompts or attaching a lightweight adapter to frozen VLMs-IE for feature adaptation. A detailed comparison is given in Fig. 5.
VLMs-TE as Classifier. LSeg [30] simply replaces the learnable classifier in segmentor [210] with text embeddings from the CLIP text encoder. SAZS [211] focuses on improving boundary segmentation supervised by the output of edge detection on ground-truth masks. During inference, SAZS fuses the predictions with eigen segments obtained through spectral analysis on DINO [103] to promote shape-awareness. Son *et al.* [212] align object queries and text embeddings by bipartite matching [13]. They force queries not matched to any base or novel class to predict a uniform distribution over base and novel classes, thus avoiding the use of "background" embedding. A multi-label ranking loss is employed to encourage the similarity of any positive label to be higher than that of any negative label.

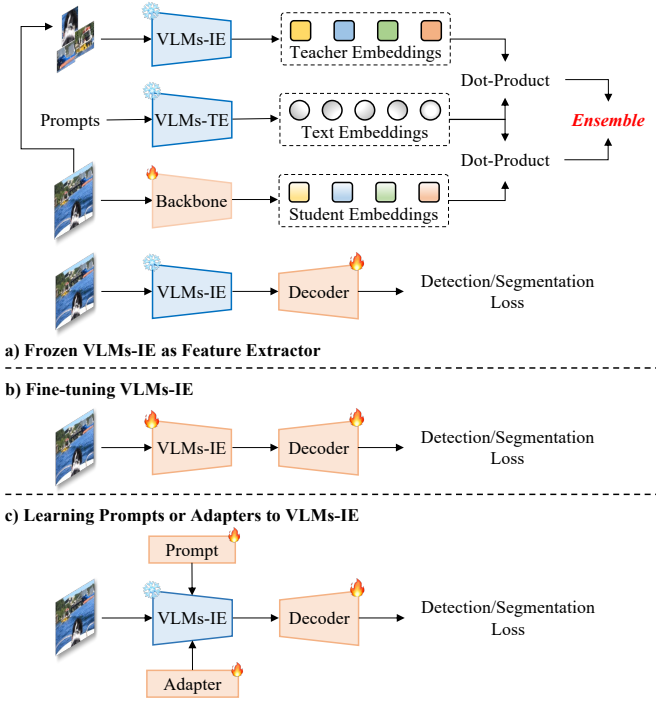

Fig. 5: Framework for transfer learning-based models.

Frozen VLMs-IE as Feature Extractor. As shown in Fig. 5, there are two variants of leveraging frozen VLMs-IE as feature extractor. The first is adding a parallel branch to the detector backbone, the predictions of two branches are later ensembled using geometric mean [26], [49]. The second discards the trainable detector backbone and utilize the frozen VLMs-IE as the only feature extractor.

For the first variant, ZegFormer [54] forwards masked image crops to frozen CLIP image encoder and ensembles its predicted scores via geometric mean [26], [49] with CLIP during inference. ReCo [213] first retrieves an archive of exemplar images for each class from unlabelled images. Then, it leverages MoCo [214] to extract seed pixels across exemplar images to construct reference image embeddings as 1×1 convolution classifier. The prediction is ensembled with DenseCLIP [215]. SCAN [216] injects semantic prior of CLIP into proposal embeddings to avoid biasing toward base classes. To mitigate the domain shift between natural images and masked crops, SCAN replaces background token embeddings in the feature map with CLIP [CLS] token.

For the second variant, ZSSeg [217] adopts the same architecture as ZegFormer except that ZSSeg adopts learnable prompts similar to CoOp [111] without ensembling with CLIP. MaskCLIP [218] removes the query and key projection in the final attention pooling layer of CLIP image encoder to enhance local semantic consistency and suppress noisy context information. This trick has been utilized in many subsequent works. CLIP-DINOiser [219] performs a concept-aware linear combination of patch features guided by DINO [103] to suppress the noisy features [109], [218] in CLIP. It also leverages an unsupervised segmentation method [220] to produce a background map that rectifies false background assignment. MVP-SEG [221] proposes multi-view prompt learning optimized by orthogonal contrastive loss to focus on different object parts. Peekaboo [222] explores how off-the-shelf Stable Diffusion (SD) [108] can perform grouping pixels with the proposed dream loss. OVDiff [223] is also entirely training-free, the same as Peekaboo [222], it relies only on SD [108]. OVDiff removes the cross-modality similarity measurement, it directly compares against image features with part-, instance-, and class-level prototypes (support set) sampled from SD. FOSSIL [224] also transforms cross-modal comparison into uni-modal comparison by constructing visual prototypes from DINOv2 [104]. POMP [225] condenses semantic concepts over twenty-thousand classes into the learned prompts. It introduces local contrast and local correction strategy to reduce the memory consumption of CoOp [111]. AttrSeg [226] decomposes each category name into attributes, *e.g.*, color, shape, texture, and parts, then hierarchically aggregates attribute tokens into the class embedding. In this way, the lexical ambiguity, neologism, and unnameability caused by the limited expressivity of a single class name are solved by providing any number of attributes describing the object. PnP-OVSS [227] pursues a training-free paradigm where GradCAM [171] plus cross-attention saliency map are combined to locate the object. An iterative saliency dropout is proposed to trade-off between over-segment (patches unrelated to the class are segmented in cross-attention map) and under-segment (GradCAM often narrowly focuses on the most discriminative patches while ignoring other useful patches). TagAlign [228] introduces a multi-tag classification loss besides the image-text contrastive loss. Self-Seg [229] leverages the feature consistency of BLIP [230], [231] to cluster image patches into distinct clusters then uses the captioner to generate open-ended descriptions. These captions are parsed into nouns and sent to X-Decoder [232] for segmentation. The nouns in generated captions may not exist in downstream mask labels. To resolve this issue, Self-Seg employs LLaMA2 [233] to map non-existent parsed nouns into the most relevant predefined categories in the dataset.

Fine-tuning VLMs-IE. This group finetunes the CLIP image encoder shown in Fig. 5 to adapt its feature representations to segmentation tasks. DenseCLIP [215] learns pixel-text matching directly on the feature map using densely annotated masks. OVSeg [55] fine-tunes CLIP image encoder on the constructed mask-category pairs to address the domain gap between masked image crops with blank areas and natural images used to pretrain CLIP. CAT-Seg [234] devises a cost aggregation module including spatial and class aggregation to produce the segmentation map. It only fine-tunes the attention layers as finetuning all parameters of CLIP harms its open-vocabulary capabilities. Since ViT-based CLIP image encoders exihibit weak local region classification ability [49], [110], SED [235] also utilizes the cost map as CAT-Seg with a ConvNext [236] backbone. MAFT [237] fine-tunes CLIP image encoder by mimicking the IoU between proposals and ground-truths to make classfication mask-aware.

Learning Prompts or Adapters for VLMs-IE. Compared to fine-tuning CLIP image encoder, learning prompts or adapters can better preserve the generalization ability in novel classes. ZegCLIP [238] and TagCLIP [239] both adopt deep prompt tuning [77]. Additionally, TagCLIP designs a learnable trusty token generating trusty maps used to mea-

sure the reliability of the raw segmentation map by CLIP. CLIPSeg [240] attaches a lightweight decoder to CLIP image encoder with U-Net [241] skip connections conditioned on text embeddings. SAN [56] attaches a lightweight vision transformer to the frozen CLIP image encoder. It requires only a single forward pass of CLIP. SAN decouples the mask proposal and classification stage by predicting attention biases applied to deeper layers of CLIP for recognition. CLIP Surgery [109] discovers that CLIP has opposite visualizations similar to the findings of SimSeg [207] and noisy activations. The proposed architecture surgery replaces the query-key self-attention with a value-value self-attention to causes the opposite visualization problem. The feature surgery identifies and removes redundant features to reduce noisy activations. Instead of learning visual prompts, CaR [242] puts a red circle on the proposal as a prompt to guide the attention of CLIP toward the region [243]. It designs a training-free and RNN-like framework where text queries are deemed as hidden states and are gradually refined to remove categories not present in the image. However, repeatedly forwarding the whole image multiple times incurs a slow inference speed.

6.2 Open-Vocabulary Instance Segmentation (OVIS)

Region-Aware Training. CGG [57] achieves the region-text alignment via a grounding loss, but not with the whole caption as in OVR-CNN [27]. CGG extracts object nouns so that object-unrelated words do not interfere with the matching process. In addition, CGG proposes caption generation to reproduce the caption paired with the image. D^2Zero [244] proposes an unseen-constrained feature extractor and an input-conditional classifier to address the bias issue. It proposes image-adaptive background representations to discriminate novel and background classes more effectively.

Pseudo-Labeling. XPM [28] first trains a teacher model on base annotations, then self-trains a student model. The pseudo regions are selected as the most compatible region *w.r.t* object nouns in the caption. However, pseudo masks contain noises that degrade performance. XPM assumes that each pixel in pseudo masks is corrupted by a Gaussian noise, and the student is trained to predict the noise level to down weight incorrect teacher predictions. Mask-free OVIS [245] performs iterative masking using ALBEF [141] and GradCAM [171] to generate pseudo-instances both for base and novel categories. To alleviate the overfitting issue, it avoids training base categories using strong supervision and novel categories using weak supervision. MosaicFusion [246] runs the T2I diffusion model on a mosaic image canvas to generate multiple pseudo instances simultaneously. Pseudo masks are obtained by aggregating cross-attention maps across heads, layers, and time steps.

Knowledge Distillation. OV-SAM [58] proposes to combine CLIP and SAM into one unified architecture. Adapters are inserted at the end of the CLIP image encoder and mimic SAM features via a mean-squared error loss. CLIP features are also fed into the SAM mask decoder to enhance open-vocabulary recognition ability.

6.3 Open-Vocabulary Panoptic Segmentation (OVPS)

Region-Aware Training. Uni-OVSeg [247] employs bipartite matching on image-caption pairs similar to VLDet [42] except that Uni-OVSeg designs a multi-scale cost matrix to achieve more accurate matching. X-Decoder [232] proposes an image-text decoupled framework for not only open-vocabulary panoptic segmentation but also other vision-language tasks. It defines latent and text queries responsible for pixel-level segmentation and semantic-level classification, respectively. APE [59] jointly trains on multiple detection and grounding datasets as well as SA-1B [196]. In contrast to GLIP [44], it reformulates visual grounding as object detection and independently encodes concepts instead of concatenating concepts and encoding them all. It treats stuff class as multiple disconnected standalone regions, removing separate head networks for things and stuff as in OpenSeeD [193]. Moreover, SA-1B is semantic-unaware, *i.e.*, thing and stuff are indistinguishable. APE can utilize SA-1B for training with only one head without differentiating things and stuff.

Knowledge Distillation. Previous synthesizers [72], [73], [128] in Sec. 4.1.2 with several linear layers do not consider the feature granularity gap between image and text modality. PADing [60] proposes learnable primitives to reflect the rich and fine-grained attributes of visual features, which are then synthesized via weighted assemblies from these abundant primitives. In addition, PADing [60] decouples visual features into semantic-related and semantic-unrelated parts and it only aligns the semantic-related parts to the inter-class relationship structure in the semantic space.

Transfer Learning. FC-CLIP [61] resembles closely with F-VLM [49] in that both use a frozen CNN-based CLIP image encoder and take a geometric ensemble for base and novel classes separately. FreeSeg [248] learns prompts separately for semantic/instance/panoptic tasks and different classes during the training stage. In inference, FreeSeg optimizes class prompts following test-time adaption [249], [250] by minimizing the entropy. PosSAM [251] employs the frozen CLIP and SAM image encoder and fuses their output visual features via cross-attention. MasQCLIP [252] follows MaskCLIP [218] except that the query projection of the mask class token is learnable. This is to reduce the shift of CLIP from extracting image-level features to classifying a masked region in an image. OMG-Seg [253] explores the paradigm of unifying semantic/instance/panoptic segmentation and their counterparts in videos under both closed and open vocabulary settings. Semantic-SAM [254] consolidates multiple datasets across granularities and trains on decoupled object and part classification, achieving semantic-awareness and granularity-abundance through a multiple-choice learning paradigm. ODISE [62] resorts to T2I diffusion models [108] as the mask feature extractor. It also proposes an implicit captioner via the CLIP image encoder to map images into pseudo words. The training is driven by a bi-directional grounding loss in Sec. 5.1. Same as ODISE [62], HIPIE [255] ensembles classification logits with CLIP. It can hierarchically segment things, stuff, and object parts. It employs two separate decoders for things and stuff to deal with different losses. Zheng *et al.* [256] design mask class tokens to extract dense image features corresponding to each mask area via the proposed relative mask attention. OPSNet [257] attaches spatial adapter on top of CLIP image encoder. Since mask embeddings are reliable at predicting base classes, and CLIP image embeddings preserve novel

class recognition ability, thus OPSNet modulates the two to learn from each other.

7 Open-Vocabulary Beyond Images

7.1 Open-Vocabulary 3D Scene Understanding

Open-vocabulary 3D scene understanding is relatively under-explored and suffers a more severe data scarcity issue, even pairing point clouds with text descriptions is not feasible. Hence, the point-cloud and text modality are typically bridged via the intermediate image modality, where VLMs (*e.g.*, CLIP) step in to guide the association. Typically, the dataset annotates the projection matrix that transforms 3D point clouds into 2D boxes and vice versa.

7.1.1 Open-Vocabulary 3D Object Detection (OV3D)

OV-3DET [65] leverages pseudo 2D boxes *w/o* class labels generated from a pretrained 2D open-vocabulary detector [46]. These pseudo 2D boxes are then back projected to 3D space, which is deemed as the supervision signal for localization learning. To classify the predicted 3D RoI features, they are first back projected to image crops which are then encoded by CLIP image encoder. These image crop features are assigned a semantic label by CLIP, thus the 3D RoI feature can connect to its labeled text embedding via paired relationship between 3D and 2D space. A triplet contrastive loss is applied to drive the 3D RoI feature to approach both the projected image feature and its associated text embedding. FM-OV3D [66] follows a similar pipeline by back projecting 2D boxes of Grounded-SAM [258]. For open-vocabulary 3D recognition, FM-OV3D aligns 3D RoI features with CLIP representations of both text and visual prompts from GPT-3 [178] and Stable Diffusion [108]. OpenSight [259] increases temporal awareness that correlates the predicted 3D boxes across consecutive timestamps and recalibrates missed or inaccurate boxes. In contrast to relying on a 2D open-vocabulary detector, CoDA [67] discovers novel 3D objects in an online and progressive manner, even though the detector is trained only on a few 3D base annotations. The projected 2D box features are used to distill CLIP knowledge into 3D object features. Though the image modality bridges the intermediate representation between point clouds and the text modality, it requires extra alignment between point clouds and images which may limit the performance. L3Det [260] proposes not to leverage images but large-scale large-vocabulary 3D object datasets. It inserts 3D objects covering both base and novel classes into the scene in a physically reasonable way and generates grounded descriptions for them. Therefore, L3Det bypasses the image modality and directly learns the alignment between 3D objects and texts through contrastive learning.

7.1.2 Open-Vocabulary 3D Segmentation

Open-Vocabulary 3D Semantic Segmentation (OV3SS). SeCondPoint [64] and 3DGenZ [63] basically follow the pipeline of novel visual feature synthesis in Sec. 3.2. OpenScene [68] embeds point features into CLIP latent space, minimizing the differences with the aggregated pixel features via a distillation loss. Thus, by aligning point features with pixel features which in turn aligned with text embeddings, point features are aligned with text embeddings. PLA [261] first calibrates predictions to avoid over-confident scores on base classes. Then, hierarchical coarse-to-fine point-caption pairs, *i.e.*, scene-, view-, and entity-level point-caption association via a pretrained captioning model [262] are constructed, effectively facilitating learning from language supervisions. However, the pseudo-captions at the view level only cover sparse and salient objects in a scene, failing to provide fine-grained descriptions. To enable dense regional point-language associations, RegionPLC [263] captions image patches in a sliding window fashion together with object proposals. A point-discriminative contrastive learning objective is further proposed that makes the gradient of each point unique. Note that both PLA and RegionPLC are capable of instance segmentation.

Open-Vocabulary 3D Instance Segmentation (OV3IS). OpenMask3D [264] aggregates per-mask feature via multi-view fusion of CLIP image embeddings. The projected 2D segmentation masks are refined by SAM [196] to remove outliers. MaskClustering [265] proposes multi-view consensus rate to assess whether 2D masks from off-the-shelf detectors belong to the same 3D instance (*i.e.*, they should be merged). An iterative graph clustering is designed to better distinguish distinct 3D instances in a class-agnostic manner. OpenIns3D [266] requires no image inputs as a bridge between point clouds and texts. It synthesizes scene-level images and leverages OVD detector in 2D domain to detect objects the associates them with semantic labels. Open3DIS [267] addresses the problem of proposing high-quality small-scale and geometrically ambiguous instances by aggregating them across frames.

7.2 Open-Vocabulary Video Understanding (OVVU)

OV2Seg [268] first proposes open-vocabulary video instance segmentation task that simultaneously detects, segments, and tracks arbitrary instances regardless of their presence in the training set. It collects a large vocabulary video instance segmentation dataset covering 1,212 categories for benchmarking. OV2Seg leverages CLIP text encoder to classify queries from its memory-induced tracking module. A concurrent work OpenVIS [69] first proposes instances in a frame exhaustively. Then in the second stage, it designs square crop that avoids distorting the aspect ratio of instances to better conform to the pre-processing of the CLIP image encoder. However, previous work [69], [268] align the same instance in different frames separately with text embeddings without considering the correlation across frames. BriVIS [269] links instance features across frames as a Brownian bridge and aligns the bridge center with text embeddings. DVIS++ [270] presents a unified framework capable of various video segmentation simultaneously.

8 Challenges and Outlook

8.1 Challenges

Overfitting on Base Classes. Due to the lack of novel annotations, the overfitting or bias issue is severe and manifested in three folds: 1) Base proposals are of higher quality and quantity than novel proposals. 2) Novel proposals are prone to be misclassified into base classes. 3) Recognition confidence on base classes is much higher than novel classes. For

the first aspect, many endeavors seek to 1) adopt a standalone and frozen proposal generator to avoid classification of base classes in detector head affecting gradients of proposal generator [194]; 2) employ pure localization quality-based objectness score [271] *w/o* foreground-background binary classification, or design complementary objectness measures utilizing a large corpus of concepts [149]. Leveraging unsupervised localization methods [272] built on DINO [103] or SAM [196] for proposal generation can also potentially mitigate the bias. For the second aspect, recalibration in the inference stage is used in many works [26], [49] by separating and ensembling the predictions of base and novel classes between the detector and CLIP.

Confusion on Novel and Background Concepts. Since only base and background text embeddings are used as classifier weights during training, both novel and background proposals are classified as background. This drawback will cause novel proposals misclassified into the background in inference. Besides, background text embedding is typically encoded by passing the template prompt "A photo of the [background]." into a CLIP text encoder or an all-zero vector. This simple representation is not sufficient and representative to cope with diverse contexts.

Correct Region-Word Correspondence. Though image-text pairs are cheap and abundant, the region-word correspondence is weak, noisy, and explicitly unknown. The bi-directional grounding loss in Sec. 5.1 may cheat on establishing correct region-word correspondence by only aligning bag-of-regions to bag-of-words. Besides, the object nouns in the caption may only cover salient objects, and they are far less than the number of proposals, *i.e.*, many objects may not find the matching words. Pseudo labels impose the constraint of one region connecting to one word and vice versa. However, they are generated once and done, iteratively refining the quality of pseudo labels in online training is less explored.

Large VLMs Adaptation. *There is a distinct discrepancy and domain gap in terms of image resolution, context, and task statistics between the pretraining and detection tuning phases*. During the pertaining phase, CLIP receives low-resolution images with full contexts including object occurrences, relationships, spatial layout, *etc*. However, in the detection tuning phase, CLIP either receives high-resolution images or low-resolution masked image crops containing a single object without any context. The masked image crops are of non-square sizes or extreme aspect ratios, the pre-processing step of CLIP resizes the shorter edge and center crop, adds more distortion and aggravates this gap. The prediction of CLIP is also not sensitive to localization quality, *i.e.*, given an image crop with only a small portion of objects of interest, CLIP still makes predictions with high confidence. Besides, fully finetuning the whole VLMs for adaptation always leads to catastrophic forgetting of prior knowledge on open-vocabulary tasks. In light of this, lightweight adapters or prompt tuning plays a crucial role in large VLMs adaptation.

Inference Speed and Evaluation Metrics. Current OVD and OVS methods mainly build on top of mainstream object detectors, such as Faster R-CNN [5], DETR [13], and Mask2Former [14], which are slow when deployed on edge devices. However, lightweight detectors like YOLO [119] may aggravate the above challenges. Real-time detectors result in lower recall rate of novel objects. Besides, distilling the knowledge of large VLMs into these small-scale models remains questionable due to their limited learning capacity. Meanwhile, the evaluation is also problematic [273], for example, suppose the predicted category and label are synonyms, current metrics will not deem the prediction as a true positive. This might be too strict given the fact that in an open-world, many words are interchangeable.

8.2 Future Directions

Enabling Open-Vocabulary on Other Scene Understanding Tasks. Currently, other tasks including open-vocabulary 3D scene understanding, video analysis [274], action recognition, object tracking, and human-object interaction [275], *etc*, are underexplored. In these problems, either the weak supervision signals are absent or the large VLMs yield pool open-vocabulary classification ability. Enabling open-vocabulary beyond detection and segmentation has became a mainstream trend.

Unifying OVD and OVS. Unification is an inevitable trend for computer vision. Though there are several works addressing different segmentation tasks simultaneously [62], [248], [255], [276] or training on multiple detection datasets [191], [192]. A universal foundational model for all tasks and datasets [193] remains barely untouched, or even further, accomplishing 2D and 3D open-vocabulary perception simultaneously can be more challenging.

Multimodal Large Language Models (LLMs) for Perception. Multimodal large language models [277], [278], [279], [280] typically comprise three parts: 1) a vision encoder; 2) a mapper that maps the visual features to the input space of LLM.2) an LLM for decoding desired outputs. Bounding boxes are represented as two corner integer points [281] and similarly for segmentation masks via sampling points on the contour [83], [282] of the mask. The reasoning capability of user intentions and interactive detection within a language context endow multi-modal LLMs for detection and segmentation in the wild.

Combining Large Foundation Models. Different foundation models have different capabilities. SAM [196] excels in localizing objects but in a class-agnostic manner. CLIP [31] is superior at image-text alignment but behaves like bag-of-words and lacks spatial-awareness for dense prediction tasks. DINO [103], [104] exhibits a superior cross-image correspondence for objects or parts of the same class but is mainly used in unsupervised localization tasks. T2I Diffusion models generate astonishing images but their usage in discriminative dense prediction tasks remains underexplored. In a nutshell, how to benefit from these emerging large foundation models and combine them are key questions for future research.

Real-Time OVD and OVS. Current models possess a heavy backbone and neck architecture, which are unsuitable for real-time applications. To fully unleash the productive potential of OVD and OVS, exploring real-time detectors [155], [283] with open-vocabulary recognition ability is a promising research direction.

9 Conclusion

We covered a broad and concrete development of OVD and OVS in this survey. First, the background section con-

sistig of definitions, related domains and tasks, canonical closed-set detectors and segmentors, and large VLMs were introduced. Then, we detailed near two hundred OVD and OVS methods. At the task level, both 2D detection and different segmentation tasks are discussed, along with 3D scene and video understanding. At the methodology level, we pivoted on the permission and usage of weak supervision signals and grouped most of the existing methods into six categories, which are universal across tasks. In the end, challenges and promising directions are discussed to facilitate future research. In addition, we benchmarked the performance of state-of-the-art methods along with their vital components for each task in the appendix.

REFERENCES

[1] H. Caesar, V. Bankiti *et al.*, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.
[2] Z. Li, W. Wang *et al.*, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *ECCV*, 2022.
[3] F. Zhu, Y. Zhu *et al.*, "Deep learning for embodied vision navigation: A survey," *arXiv*, 2021.
[4] P. Anderson, Q. Wu *et al.*, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *CVPR*, 2018.
[5] S. Ren, K. He *et al.*, "Faster r-cnn: Towards real-time object detection with region proposal networks," *NeurIPS*, 2015.
[6] T.-Y. Lin, P. Dollár *et al.*, "Feature pyramid networks for object detection," in *CVPR*, 2017.
[7] I. Misra, R. Girdhar *et al.*, "An end-to-end transformer model for 3d object detection," in *ICCV*, 2021.
[8] J. Long, E. Shelhamer *et al.*, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.
[9] B. Cheng, A. Schwing *et al.*, "Per-pixel classification is not all you need for semantic segmentation," *NeurIPS*, 2021.
[10] K. He, G. Gkioxari *et al.*, "Mask r-cnn," in *ICCV*, 2017.
[11] A. Kirillov, K. He *et al.*, "Panoptic segmentation," in *CVPR*, 2019.
[12] T.-Y. Lin, P. Goyal *et al.*, "Focal loss for dense object detection," in *ICCV*, 2017.
[13] N. Carion, F. Massa *et al.*, "End-to-end object detection with transformers," in *ECCV*, 2020.
[14] B. Cheng, I. Misra *et al.*, "Masked-attention mask transformer for universal image segmentation," in *CVPR*, 2022.
[15] M. Everingham, S. A. Eslami *et al.*, "The pascal visual object classes challenge: A retrospective," *IJCV*, 2015.
[16] T.-Y. Lin, M. Maire *et al.*, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
[17] A. Gupta, P. Dollar *et al.*, "Lvis: A dataset for large vocabulary instance segmentation," in *CVPR*, 2019.
[18] A. Bansal, K. Sikka *et al.*, "Zero-shot object detection," in *ECCV*, 2018.
[19] S. Rahman, S. Khan *et al.*, "Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts," in *ACCV*, 2019.
[20] P. Zhu, H. Wang *et al.*, "Zero shot detection," *TCSVT*, 2019.
[21] M. Bucher, T.-H. Vu *et al.*, "Zero-shot semantic segmentation," *NeurIPS*, 2019.
[22] Y. Zheng, J. Wu *et al.*, "Zero-shot instance segmentation," in *CVPR*, 2021.
[23] T. Mikolov, I. Sutskever *et al.*, "Distributed representations of words and phrases and their compositionality," *NeurIPS*, 2013.
[24] J. Pennington, R. Socher *et al.*, "Glove: Global vectors for word representation," in *EMNLP*, 2014.
[25] J. Devlin, M.-W. Chang *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv*, 2018.
[26] X. Gu, T.-Y. Lin *et al.*, "Open-vocabulary object detection via vision and language knowledge distillation," *arXiv*, 2021.
[27] A. Zareian, K. D. Rosa *et al.*, "Open-vocabulary object detection using captions," in *CVPR*, 2021.
[28] D. Huynh, J. Kuen *et al.*, "Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling," in *CVPR*, 2022.
[29] G. Ghiasi, X. Gu *et al.*, "Scaling open-vocabulary image segmentation with image-level labels," in *ECCV*, 2022.
[30] B. Li, K. Q. Weinberger *et al.*, "Language-driven semantic segmentation," *arXiv*, 2022.
[31] A. Radford, J. W. Kim *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.
[32] S. Rahman, S. Khan *et al.*, "Improved visual-semantic alignment for zero-shot object detection," in *AAAI*, 2020.
[33] D. Gupta, A. Anantharaman *et al.*, "A multi-space approach to zero-shot object detection," in *WACV*, 2020.
[34] P. Zhu, H. Wang *et al.*, "Don't even look once: Synthesizing features for zero-shot detection," in *CVPR*, 2020.
[35] S. Zhao, C. Gao *et al.*, "Gtnet: Generative transfer network for zero-shot object detection," in *AAAI*, 2020.
[36] P. Huang, J. Han *et al.*, "Robust region feature synthesizer for zero-shot object detection," in *CVPR*, 2022.
[37] Y. Xian, S. Choudhury *et al.*, "Semantic projection network for zero-and few-label semantic segmentation," in *CVPR*, 2019.
[38] D. Baek, Y. Oh *et al.*, "Exploiting a joint embedding space for generalized zero-shot semantic segmentation," in *ICCV*, 2021.
[39] H. Zhang and H. Ding, "Prototypical matching and open set rejection for zero-shot semantic segmentation," in *ICCV*, 2021.
[40] Z. Gu, S. Zhou *et al.*, "Context-aware feature generation for zero-shot semantic segmentation," in *ACM MM*, 2020.
[41] A. Kamath, M. Singh *et al.*, "Mdetr - modulated detection for end-to-end multi-modal understanding," in *ICCV*, 2021.
[42] C. Lin, P. Sun *et al.*, "Learning object-language alignments for open-vocabulary object detection," *arXiv*, 2022.
[43] Y. Zang, W. Li *et al.*, "Open-vocabulary detr with conditional matching," in *ECCV*, 2022.
[44] L. H. Li, P. Zhang *et al.*, "Grounded language-image pre-training," in *CVPR*, 2022.
[45] M. Gao, C. Xing *et al.*, "Open vocabulary object detection with pseudo bounding-box labels," in *ECCV*, 2022.
[46] X. Zhou, R. Girdhar *et al.*, "Detecting twenty-thousand classes using image-level supervision," in *ECCV*, 2022.
[47] Y. Du, F. Wei *et al.*, "Learning to prompt for open-vocabulary object detection with vision-language model," in *CVPR*, 2022.
[48] S. Wu, W. Zhang *et al.*, "Aligning bag of regions for open-vocabulary object detection," in *CVPR*, 2023.
[49] W. Kuo, Y. Cui *et al.*, "F-vlm: Open-vocabulary object detection upon frozen vision and language models," *arXiv*, 2022.
[50] J. Xu, S. De Mello *et al.*, "Groupvit: Semantic segmentation emerges from text supervision," in *CVPR*, 2022.
[51] N. Zabari and Y. Hoshen, "Open-vocabulary semantic segmentation using test-time distillation," in *ECCV*, 2022.
[52] K. Han, Y. Liu *et al.*, "Global knowledge calibration for fast open-vocabulary segmentation," *arXiv*, 2023.
[53] H. Wang, P. K. A. Vasu *et al.*, "Sam-clip: Merging vision foundation models towards semantic and spatial understanding," *arXiv*, 2023.
[54] J. Ding, N. Xue *et al.*, "Decoupling zero-shot semantic segmentation," in *CVPR*, 2022.
[55] F. Liang, B. Wu *et al.*, "Open-vocabulary semantic segmentation with mask-adapted clip," in *CVPR*, 2023.
[56] M. Xu, Z. Zhang *et al.*, "Side adapter network for open-vocabulary semantic segmentation," in *CVPR*, 2023.
[57] J. Wu, X. Li *et al.*, "Betrayed by captions: Joint caption grounding and generation for open vocabulary instance segmentation," *arXiv*, 2023.
[58] H. Yuan, X. Li *et al.*, "Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively," *arXiv*, 2024.
[59] Y. Shen, C. Fu *et al.*, "Aligning and prompting everything all at once for universal visual perception," 2024.
[60] S. He, H. Ding *et al.*, "Primitive generation and semantic-related alignment for universal zero-shot segmentation," in *CVPR*, 2023.
[61] Q. Yu, J. He *et al.*, "Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip," *NeurIPS*, 2024.
[62] J. Xu, S. Liu *et al.*, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *CVPR*, 2023.
[63] B. Michele, A. Boulch *et al.*, "Generative zero-shot learning for semantic segmentation of 3d point clouds," in *3DV*, 2021.
[64] B. Liu, S. Deng *et al.*, "Language-level semantics conditioned 3d point cloud segmentation," *arXiv*, 2021.
[65] Y. Lu, C. Xu *et al.*, "Open-vocabulary point-cloud object detection without 3d annotation," in *CVPR*, 2023.

[66] D. Zhang, C. Li *et al.*, "Fm-ov3d: Foundation model-based cross-modal knowledge blending for open-vocabulary 3d detection," *arXiv*, 2023.

[67] Y. Cao, Z. Yihan *et al.*, "Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection," *NeurIPS*, 2024.

[68] S. Peng, K. Genova *et al.*, "Openscene: 3d scene understanding with open vocabularies," in *CVPR*, 2023.

[69] P. Guo, T. Huang *et al.*, "Openvis: Open-vocabulary video instance segmentation," *arXiv*, 2023.

[70] A. Joulin, E. Grave *et al.*, "Bag of tricks for efficient text classification," *arXiv*, 2016.

[71] Y. Li, K. Swersky *et al.*, "Generative moment matching networks," in *ICML*, 2015.

[72] I. Goodfellow, J. Pouget-Abadie *et al.*, "Generative adversarial networks," *ACM Communications*, 2020.

[73] K. Sohn, H. Lee *et al.*, "Learning structured output representation using deep conditional generative models," *NeurIPS*, 2015.

[74] R. Krishna, Y. Zhu *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, 2017.

[75] L. Yu, P. Poirson *et al.*, "Modeling context in referring expressions," in *ECCV*, 2016.

[76] V. K. Nagaraja, V. I. Morariu *et al.*, "Modeling context between objects for referring expression understanding," in *ECCV*, 2016.

[77] M. Jia, L. Tang *et al.*, "Visual prompt tuning," in *ECCV*, 2022.

[78] R. Zhang, R. Fang *et al.*, "Tip-adapter: Training-free clip-adapter for better vision-language modeling," *arXiv*, 2021.

[79] Y.-L. Sung, J. Cho *et al.*, "Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks," in *CVPR*, 2022.

[80] S. Rahman, S. Khan *et al.*, "Transductive learning for zero-shot object detection," in *ICCV*, 2019.

[81] S. Kazemzadeh, V. Ordonez *et al.*, "Referitgame: Referring to objects in photographs of natural scenes," in *EMNLP*, 2014.

[82] B. A. Plummer, L. Wang *et al.*, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *IJCV*, 2015.

[83] C. Zhu, Y. Zhou *et al.*, "Seqtr: A simple yet universal network for visual grounding," in *ECCV*, 2022.

[84] D. Zhang, J. Han *et al.*, "Weakly supervised object localization and detection: A survey," *TPAMI*, 2021.

[85] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *CVPR*, 2016.

[86] A. Dhamija, M. Gunther *et al.*, "The overlooked elephant of object detection: Open set," in *WACV*, 2020.

[87] D. Miller, L. Nicholson *et al.*, "Dropout sampling for robust object detection in open-set conditions," in *ICRA*, 2018.

[88] T. Pham, T.-T. Do *et al.*, "Bayesian semantic instance segmentation in open set world," in *ECCV*, 2018.

[89] J. Hwang, S. W. Oh *et al.*, "Exemplar-based open-set panoptic segmentation network," in *CVPR*, 2021.

[90] W. J. Scheirer, A. de Rezende Rocha *et al.*, "Toward open set recognition," *TPAMI*, 2012.

[91] C. Geng, S.-j. Huang *et al.*, "Recent advances in open set recognition: A survey," *TPAMI*, 2020.

[92] K. Joseph, S. Khan *et al.*, "Towards open world object detection," in *CVPR*, 2021.

[93] A. Gupta, S. Narayan *et al.*, "Ow-detr: Open-world detection transformer," in *CVPR*, 2022.

[94] J. Cen, P. Yun *et al.*, "Deep metric learning for open world semantic segmentation," in *ICCV*, 2021.

[95] W. Liu, X. Wang *et al.*, "Energy-based out-of-distribution detection," *NeurIPS*, 2020.

[96] J. Yang, K. Zhou *et al.*, "Generalized out-of-distribution detection: A survey," *arXiv*, 2021.

[97] R. Girshick, "Fast r-cnn," in *ICCV*, 2015.

[98] Z. Tian, C. Shen *et al.*, "Fcos: Fully convolutional one-stage object detection," in *ICCV*, 2019.

[99] X. Zhu, W. Su *et al.*, "Deformable detr: Deformable transformers for end-to-end object detection," in *ICLR*, 2021.

[100] L.-C. Chen, G. Papandreou *et al.*, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *TPAMI*, 2017.

[101] L.-C. Chen, Y. Zhu *et al.*, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.

[102] C. Jia, Y. Yang *et al.*, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021.

[103] M. Caron, H. Touvron *et al.*, "Emerging properties in self-supervised vision transformers," in *ICCV*, 2021.

[104] M. Oquab, T. Darcet *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv*, 2023.

[105] K. He, X. Chen *et al.*, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022.

[106] M. Yuksekgonul, F. Bianchi *et al.*, "When and why vision-language models behave like bags-of-words, and what to do about it?" in *ICLR*, 2022.

[107] J. Ho, A. Jain *et al.*, "Denoising diffusion probabilistic models," *NeurIPS*, 2020.

[108] R. Rombach, A. Blattmann *et al.*, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[109] Y. Li, H. Wang *et al.*, "Clip surgery for better explainability with enhancement in open-vocabulary tasks," *arXiv*, 2023.

[110] S. Wu, W. Zhang *et al.*, "CLIPSelf: Vision transformer distills itself for open-vocabulary dense prediction," in *ICLR*, 2024.

[111] K. Zhou, J. Yang *et al.*, "Learning to prompt for vision-language models," *IJCV*, 2022.

[112] G. A. Miller, "Wordnet: a lexical database for english," *ACM Communications*, 1995.

[113] S. Rahman, S. H. Khan *et al.*, "Zero-shot object detection: Joint recognition and localization of novel concepts," *IJCV*, 2020.

[114] R. Luo, N. Zhang *et al.*, "Context-aware zero-shot recognition," in *AAAI*, 2020.

[115] Z. Li, L. Yao *et al.*, "Zero-shot object detection with textual descriptions," in *AAAI*, 2019.

[116] Y. Zheng, R. Huang *et al.*, "Background learnable cascade for zero-shot object detection," in *ACCV*, 2020.

[117] S. Khandelwal, A. Nambirajan *et al.*, "Frustratingly simple but effective zero-shot detection and segmentation: Analysis and a strong baseline," *arXiv*, 2023.

[118] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *CVPR*, 2017.

[119] ——, "Yolov3: An incremental improvement," *arXiv*, 2018.

[120] B. Demirel, R. G. Cinbis *et al.*, "Zero-shot object detection by hybrid region embedding," *arXiv*, 2018.

[121] Y. Li, Y. Shao *et al.*, "Context-guided super-class inference for zero-shot detection," in *CVPRW*, 2020.

[122] Y. Li, P. Li *et al.*, "Inference fusion with associative semantics for unseen object detection," in *AAAI*, 2021.

[123] C. Yan, X. Chang *et al.*, "Semantics-guided contrastive network for zero-shot object detection," *TPAMI*, 2022.

[124] L. Zhang, X. Wang *et al.*, "Zero-shot object detection via learning an embedding from semantic space to visual space," in *IJCAI*, 2020.

[125] G. Dinu, A. Lazaridou *et al.*, "Improving zero-shot learning by mitigating the hubness problem," *arXiv*, 2014.

[126] N. Hayat, M. Hayat *et al.*, "Synthesizing the unseen for zero-shot object detection," in *ACCV*, 2020.

[127] Q. Mao, H.-Y. Lee *et al.*, "Mode seeking generative adversarial networks for diverse image synthesis," in *CVPR*, 2019.

[128] M. Arjovsky, S. Chintala *et al.*, "Wasserstein generative adversarial networks," in *ICML*, 2017.

[129] P. Hu, S. Sclaroff *et al.*, "Uncertainty-aware learning for zero-shot semantic segmentation," *NeurIPS*, 2020.

[130] N. Kato, T. Yamasaki *et al.*, "Zero-shot semantic segmentation via variational mapping," in *ICCVW*, 2019.

[131] K. Wang, J. H. Liew *et al.*, "Panet: Few-shot image semantic segmentation with prototype alignment," in *ICCV*, 2019.

[132] P. Li, Y. Wei *et al.*, "Consistent structural relation learning for zero-shot segmentation," *NeurIPS*, 2020.

[133] J. Cheng, S. Nandi *et al.*, "Sign: Spatial-information incorporated generative network for generalized zero-shot semantic segmentation," in *ICCV*, 2021.

[134] P. Sharma, N. Ding *et al.*, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *ACL*, 2018.

[135] S. Changpinyo, P. Sharma *et al.*, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *CVPR*, 2021.

[136] X. Chen, H. Fang *et al.*, "Microsoft coco captions: Data collection and evaluation server," *arXiv*, 2015.

[137] M. A. Bravo, S. Mittal *et al.*, "Localized vision-language matching for open-vocabulary object detection," in *DAGM GCPR*, 2022.
[138] Z. Huang, Z. Zeng *et al.*, "Pixel-bert: Aligning image pixels with text by deep multi-modal transformers," *arXiv*, 2020.
[139] J. Lu, D. Batra *et al.*, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *NeurIPS*, 2019.
[140] W. Kim, B. Son *et al.*, "Vilt: Vision-and-language transformer without convolution or region supervision," in *ICML*, 2021.
[141] J. Li, R. Selvaraju *et al.*, "Align before fuse: Vision and language representation learning with momentum distillation," *NeurIPS*, 2021.
[142] Y. Xu, M. Zhang *et al.*, "Exploring multi-modal contextual knowledge for open-vocabulary object detection," *arXiv*, 2023.
[143] L. Yao, J. Han *et al.*, "Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection," *arXiv*, 2022.
[144] ——, "Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment," in *CVPR*, 2023.
[145] J. Lin, Y. Shen *et al.*, "Weakly supervised open-vocabulary object detection," *arXiv*, 2023.
[146] D. Kim, A. Angelova *et al.*, "Region-aware pretraining for open-vocabulary object detection with vision transformers," in *CVPR*, 2023.
[147] ——, "Contrastive feature masking open-vocabulary vision transformer," in *ICCV*, 2023.
[148] ——, "Detection-oriented image-text pretraining for open-vocabulary detection," *arXiv*, 2023.
[149] J. Wang, H. Zhang *et al.*, "Open-vocabulary object detection with an open corpus," in *ICCV*, 2023.
[150] H. Song and J. Bang, "Prompt-guided transformers for end-to-end open-vocabulary object detection," *arXiv*, 2023.
[151] X. Wu, F. Zhu *et al.*, "Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching," in *CVPR*, 2023.
[152] C. Shi and S. Yang, "Edadet: Open-vocabulary object detection using early dense alignment," in *ICCV*, 2023.
[153] M. Maaz, H. Rasheed *et al.*, "Class-agnostic object detection with multi-modal transformer," in *ECCV*, 2022.
[154] Y. Xu, M. Zhang *et al.*, "Multi-modal queried object detection in the wild," *arXiv*, 2023.
[155] T. Cheng, L. Song *et al.*, "Yolo-world: Real-time open-vocabulary object detection," in *CVPR*, 2024.
[156] H. Shi, M. Hayat *et al.*, "Open-vocabulary object detection via scene graph discovery," *arXiv*, 2023.
[157] Y. Zhong, J. Yang *et al.*, "Regionclip: Region-based language-image pretraining," in *CVPR*, 2022.
[158] S. Zhao, Z. Zhang *et al.*, "Exploiting unlabeled data with vision and language models for object detection," in *ECCV*, 2022.
[159] H. Zhang, P. Zhang *et al.*, "Glipv2: Unifying localization and vision-language understanding," *NeurIPS*, 2022.
[160] S. Antol, A. Agrawal *et al.*, "Vqa: Visual question answering," in *ICCV*, 2015.
[161] P. Anderson, X. He *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018.
[162] K. Xu, J. Ba *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.
[163] S. J. Rennie, E. Marcheret *et al.*, "Self-critical sequence training for image captioning," in *CVPR*, 2017.
[164] S. Liu, Z. Zeng *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv*, 2023.
[165] C. Feng, Y. Zhong *et al.*, "Promptdet: Towards open-vocabulary detection using uncurated images," in *ECCV*, 2022.
[166] Y.-C. Liu, C.-Y. Ma *et al.*, "Unbiased teacher for semi-supervised object detection," in *ICLR*, 2020.
[167] M. Xu, Z. Zhang *et al.*, "End-to-end semi-supervised object detection with soft teacher," in *ICCV*, 2021.
[168] S. Zhao, S. Schulter *et al.*, "Improving pseudo labels for open-vocabulary object detection," *arXiv*, 2023.
[169] R. Arandjelović, A. Andonian *et al.*, "Three ways to improve feature alignment for open vocabulary detection," *arXiv*, 2023.
[170] H. Bangalath, M. Maaz *et al.*, "Bridging the gap between object and image-level representations for open-vocabulary detection," *NeurIPS*, 2022.
[171] R. R. Selvaraju, M. Cogswell *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017.
[172] S. Wu, W. Zhang *et al.*, "Clim: Contrastive language-image mosaic for region representation," *AAAI*, 2024.
[173] Y. Long, J. Han *et al.*, "Fine-grained visual–text prompt-driven self-training for open-vocabulary object detection," *TNNLS*, 2023.
[174] J. Jeong, G. Park *et al.*, "Proxydet: Synthesizing proxy novel classes via classwise mixup for open vocabulary object detection," *arXiv*, 2023.
[175] C. Ma, Y. Jiang *et al.*, "Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection," *NeurIPS*, 2024.
[176] O. Russakovsky, J. Deng *et al.*, "Imagenet large scale visual recognition challenge," *IJCV*, 2015.
[177] P. Kaul, W. Xie *et al.*, "Multi-modal classifiers for open-vocabulary object detection," *arXiv*, 2023.
[178] T. Brown, B. Mann *et al.*, "Language models are few-shot learners," *NeurIPS*, 2020.
[179] S. Kang, J. Cha *et al.*, "Learning pseudo-labeler beyond noun concepts for open-vocabulary object detection," *arXiv*, 2023.
[180] H.-C. Cho, W. Y. Jhoo *et al.*, "Open-vocabulary object detection using pseudo caption labels," *arXiv*, 2023.
[181] e. a. Xie, Johnathan, "Zero-shot object detection through vision-language embedding alignment," in *ICDMW*, 2022.
[182] C. Pham, T. Vu *et al.*, "Lp-ovod: Open-vocabulary object detection by linear probing," in *WACV*, 2024.
[183] Z. Liu, X. Hu *et al.*, "Efficient feature distillation for zero-shot detection," *arXiv*, 2023.
[184] R. Fang, G. Pang *et al.*, "Simple image-level classification improves open-vocabulary object detection," *arXiv*, 2023.
[185] A. v. d. Oord, Y. Li *et al.*, "Representation learning with contrastive predictive coding," *arXiv*, 2018.
[186] L. Wang, Y. Liu *et al.*, "Object-aware distillation pyramid for open-vocabulary object detection," in *CVPR*, 2023.
[187] J. Lin and S. Gong, "Gridclip: One-stage object detection by grid-level clip representation learning," *arXiv*, 2023.
[188] L. Li, J. Miao *et al.*, "Distilling detr with visual-linguistic knowledge for open-vocabulary object detection," in *ICCV*, 2023.
[189] Z. Ma, G. Luo *et al.*, "Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation," in *CVPR*, 2022.
[190] M. Minderer, A. Gritsenko *et al.*, "Simple open-vocabulary object detection with vision transformers," *arXiv*, 2022.
[191] Z. Wang, Y. Li *et al.*, "Detecting everything in the open world: Towards universal object detection," in *CVPR*, 2023.
[192] Y. Chen, M. Wang *et al.*, "Scaledet: A scalable multi-dataset object detector," in *CVPR*, 2023.
[193] H. Zhang, F. Li *et al.*, "A simple framework for open-vocabulary segmentation and detection," *arXiv*, 2023.
[194] J. Li, C. Xie *et al.*, "What makes good open-vocabulary detector: A disassembling perspective," *arXiv*, 2023.
[195] X. Han, L. Wei *et al.*, "Boosting segment anything model towards open-vocabulary learning," *arXiv*, 2023.
[196] A. Kirillov, E. Mintun *et al.*, "Segment anything," *arXiv*, 2023.
[197] M. F. Naeem, Y. Xian *et al.*, "Silc: Improving vision language pretraining with self-distillation," *arXiv*, 2023.
[198] E. Jang, S. Gu *et al.*, "Categorical reparameterization with gumbel-softmax," *arXiv*, 2016.
[199] Q. Liu, Y. Wen *et al.*, "Open-world semantic segmentation via contrasting and clustering vision-language embedding," in *ECCV*, 2022.
[200] M. Caron, I. Misra *et al.*, "Unsupervised learning of visual features by contrasting cluster assignments," *NeurIPS*, 2020.
[201] M. Tschannen, J. Djolonga *et al.*, "On mutual information maximization for representation learning," *arXiv*, 2019.
[202] H. Luo, J. Bao *et al.*, "Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation," *arXiv*, 2022.
[203] J. Xu, J. Hou *et al.*, "Learning open-vocabulary semantic segmentation models from natural language supervision," in *CVPR*, 2023.
[204] F. Locatello, D. Weissenborn *et al.*, "Object-centric learning with slot attention," *NeurIPS*, 2020.
[205] J. Mukhoti, T.-Y. Lin *et al.*, "Open vocabulary semantic segmentation with patch aligned contrastive learning," in *CVPR*, 2023.
[206] J. Cha, J. Mun *et al.*, "Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs," in *CVPR*, 2023.
[207] M. Xu, Z. Zhang *et al.*, "A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model," in *ECCV*, 2022.

[208] H. Chefer, S. Gur *et al.*, "Transformer interpretability beyond attention visualization," in *CVPR*, 2021.

[209] J. Chen, D. Zhu *et al.*, "Exploring open-vocabulary semantic segmentation from clip vision encoder distillation only," in *ICCV*, 2023.

[210] R. Ranftl, A. Bochkovskiy *et al.*, "Vision transformers for dense prediction," in *ICCV*, 2021.

[211] X. Liu, B. Tian *et al.*, "Delving into shape-aware zero-shot semantic segmentation," in *CVPR*, 2023.

[212] S. D. Dao, H. Shi *et al.*, "Class enhancement losses with pseudo labels for open-vocabulary semantic segmentation," *TMM*, 2023.

[213] G. Shin, W. Xie *et al.*, "Reco: Retrieve and co-segment for zero-shot transfer," *NeurIPS*, 2022.

[214] K. He, H. Fan *et al.*, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020.

[215] Y. Rao, W. Zhao *et al.*, "Denseclip: Language-guided dense prediction with context-aware prompting," in *CVPR*, 2022.

[216] Y. Liu, S. Bai *et al.*, "Open-vocabulary segmentation with semantic-assisted calibration," *arXiv*, 2023.

[217] M. Xu, Z. Zhang *et al.*, "A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model," in *ECCV*, 2022.

[218] C. Zhou, C. C. Loy *et al.*, "Extract free dense labels from clip," in *ECCV*, 2022.

[219] M. Wysoczańska, O. Siméoni *et al.*, "Clip-dinoiser: Teaching clip a few dino tricks," *arXiv*, 2023.

[220] O. Siméoni, C. Sekkat *et al.*, "Unsupervised object localization: Observing the background to discover objects," in *CVPR*, 2023.

[221] J. Guo, Q. Wang *et al.*, "Mvp-seg: Multi-view prompt learning for open-vocabulary semantic segmentation," *arXiv*, 2023.

[222] R. Burgert, K. Ranasinghe *et al.*, "Peekaboo: Text to image diffusion models are zero-shot segmentors," *arXiv*, 2022.

[223] L. Karazija, I. Laina *et al.*, "Diffusion models for zero-shot open-vocabulary segmentation," *arXiv*, 2023.

[224] L. Barsellotti, R. Amoroso *et al.*, "Fossil: Free open-vocabulary semantic segmentation through synthetic references retrieval," in *WACV*, 2024, pp. 1464–1473.

[225] S. Ren, A. Zhang *et al.*, "Prompt pre-training with twenty-thousand classes for open-vocabulary visual recognition," *arXiv*, 2023.

[226] C. Ma, Y. Yang *et al.*, "Open-vocabulary semantic segmentation via attribute decomposition-aggregation," in *NeurIPS*, 2023.

[227] L. Jiayun, S. Khandelwal *et al.*, "Plug-and-play, dense-label-free extraction of open-vocabulary semantic segmentation from vision-language models," *arXiv*, 2023.

[228] Q. Liu, K. Zheng *et al.*, "Tagalign: Improving vision-language alignment with multi-tag classification," *arXiv*, 2023.

[229] O. Ülger, M. Kulicki *et al.*, "Self-guided open-vocabulary semantic segmentation," *arXiv*, 2023.

[230] J. Li, D. Li *et al.*, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022.

[231] ——, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, 2023.

[232] X. Zou, Z.-Y. Dou *et al.*, "Generalized decoding for pixel, image, and language," in *CVPR*, 2023.

[233] H. Touvron, L. Martin *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv*, 2023.

[234] S. Cho, H. Shin *et al.*, "Cat-seg: Cost aggregation for open-vocabulary semantic segmentation," *arXiv*, 2023.

[235] B. Xie, J. Cao *et al.*, "Sed: A simple encoder-decoder for open-vocabulary semantic segmentation," *arXiv*, 2023.

[236] Z. Liu, H. Mao *et al.*, "A convnet for the 2020s," in *CVPR*, 2022.

[237] S. Jiao, Y. Wei *et al.*, "Learning mask-aware clip representations for zero-shot segmentation," *NeurIPS*, 2023.

[238] Z. Zhou, Y. Lei *et al.*, "Zegclip: Towards adapting clip for zero-shot semantic segmentation," in *CVPR*, 2023.

[239] J. Li, P. Chen *et al.*, "Tagclip: Improving discrimination ability of open-vocabulary semantic segmentation," *arXiv*, 2023.

[240] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *CVPR*, 2022.

[241] O. Ronneberger, P. Fischer *et al.*, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[242] S. Sun, R. Li *et al.*, "Clip as rnn: Segment countless visual concepts without training endeavor," *arXiv*, 2023.

[243] A. Shtedritski, C. Rupprecht *et al.*, "What does clip know about a red circle? visual prompt engineering for vlms," in *ICCV*, 2023.

[244] S. He, H. Ding *et al.*, "Semantic-promoted debiasing and background disambiguation for zero-shot instance segmentation," in *CVPR*, 2023.

[245] V. VS, N. Yu *et al.*, "Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations," in *CVPR*, 2023.

[246] J. Xie, W. Li *et al.*, "Mosaicfusion: Diffusion models as data augmenters for large vocabulary instance segmentation," *arXiv*, 2023.

[247] Z. Wang, X. Xia *et al.*, "Open-vocabulary segmentation with unpaired mask-text supervision," *arXiv*, 2024.

[248] J. Qin, J. Wu *et al.*, "Freeseg: Unified, universal and open-vocabulary image segmentation," in *CVPR*, 2023.

[249] D. Wang, E. Shelhamer *et al.*, "Tent: Fully test-time adaptation by entropy minimization," *arXiv*, 2020.

[250] M. Shu, W. Nie *et al.*, "Test-time prompt tuning for zero-shot generalization in vision-language models," *NeruIPS*, 2022.

[251] V. VS, S. Borse *et al.*, "Possam: Panoptic open-vocabulary segment anything," *arXiv*, 2024.

[252] X. Xu, T. Xiong *et al.*, "Masqclip for open-vocabulary universal image segmentation," in *ICCV*, 2023.

[253] X. Li, H. Yuan *et al.*, "Omg-seg: Is one model good enough for all segmentation?" *arXiv*, 2024.

[254] F. Li, H. Zhang *et al.*, "Semantic-sam: Segment and recognize anything at any granularity," *arXiv*, 2023.

[255] X. Wang, S. Li *et al.*, "Hierarchical open-vocabulary universal image segmentation," *arXiv*, 2023.

[256] Z. Ding, J. Wang *et al.*, "Open-vocabulary panoptic segmentation with maskclip," *arXiv*, 2022.

[257] X. Chen, S. Li *et al.*, "Open-vocabulary panoptic segmentation with embedding modulation," *arXiv*, 2023.

[258] T. Ren, S. Liu *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv*, 2024.

[259] H. Zhang, J. Xu *et al.*, "Opensight: A simple open-vocabulary framework for lidar-based object detection," *arXiv*, 2023.

[260] C. Zhu, W. Zhang *et al.*, "Object2scene: Putting objects in context for open-vocabulary 3d detection," *arXiv*, 2023.

[261] R. Ding, J. Yang *et al.*, "Pla: Language-driven open-vocabulary 3d scene understanding," in *CVPR*, 2023.

[262] A. Radford, J. Wu *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, 2019.

[263] J. Yang, R. Ding *et al.*, "Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding," *arXiv*, 2023.

[264] A. Takmaz, E. Fedele *et al.*, "Openmask3d: Open-vocabulary 3d instance segmentation," 2023.

[265] M. Yan, J. Zhang *et al.*, "Maskclustering: View consensus based mask graph clustering for open-vocabulary 3d instance segmentation," *arXiv*, 2024.

[266] Z. Huang, X. Wu *et al.*, "Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation," *arXiv*, 2023.

[267] P. D. Nguyen, T. D. Ngo *et al.*, "Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance," *arXiv*, 2023.

[268] H. Wang, C. Yan *et al.*, "Towards open-vocabulary video instance segmentation," in *ICCV*, 2023.

[269] Z. Cheng, K. Li *et al.*, "Instance brownian bridge as texts for open-vocabulary video instance segmentation," *arXiv*, 2024.

[270] T. Zhang, X. Tian *et al.*, "Dvis++: Improved decoupled framework for universal video segmentation," *arXiv*, 2023.

[271] D. Kim, T.-Y. Lin *et al.*, "Learning open-world object proposals without learning to classify," *Robotics and Automation*, 2022.

[272] O. Siméoni, É. Zablocki *et al.*, "Unsupervised object localization in the era of self-supervised vits: A survey," *arXiv*, 2023.

[273] H. Zhou, T. Shen *et al.*, "Rethinking evaluation metrics of open-vocabulary segmentaion," *arXiv*, 2023.

[274] K. Gao, L. Chen *et al.*, "Compositional prompt tuning with motion cues for open-vocabulary video relation detection," in *ICLR*, 2023.

[275] L. Li, J. Xiao *et al.*, "Zero-shot visual relation detection via composite visual cues from large language models," *arXiv*, 2023.

[276] X. Gu, Y. Cui *et al.*, "Dataseg: Taming a universal multi-dataset multi-task segmentation model," *arXiv*, 2023.

[277] Y. Zang, W. Li *et al.*, "Contextual object detection with multimodal large language models," *arXiv*, 2023.

[278] R. Pi, J. Gao *et al.*, "Detgpt: Detect what you need via reasoning," *arXiv*, 2023.

[279] W. Wang, Z. Chen *et al.*, "Visionllm: Large language model is also an open-ended decoder for vision-centric tasks," *NeurIPS*, 2024.
[280] X. Lai, Z. Tian *et al.*, "Lisa: Reasoning segmentation via large language model," *arXiv*, 2023.
[281] T. Chen, S. Saxena *et al.*, "Pix2seq: A language modeling framework for object detection," *arXiv*, 2021.
[282] ——, "A unified sequence interface for vision tasks," *NeurIPS*, 2022.
[283] W. Lv, S. Xu *et al.*, "Detrs beat yolos on real-time object detection," *arXiv*, 2023.
[284] R. Mottaghi, X. Chen *et al.*, "The role of context for object detection and semantic segmentation in the wild," in *CVPR*, 2014.
[285] S. Shao, Z. Li *et al.*, "Objects365: A large-scale, high-quality dataset for object detection," in *ICCV*, 2019.
[286] A. Kuznetsova, H. Rom *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *IJCV*, 2020.
[287] H. Caesar, J. Uijlings *et al.*, "Coco-stuff: Thing and stuff classes in context," in *CVPR*, 2018.
[288] B. Zhou, H. Zhao *et al.*, "Scene parsing through ade20k dataset," in *CVPR*, 2017.
[289] M. Cordts, M. Omran *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.
[290] S. Song, S. P. Lichtenberg *et al.*, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *CVPR*, 2015.
[291] A. Dai, A. X. Chang *et al.*, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *CVPR*, 2017.
[292] D. Rozenberszki, O. Litany *et al.*, "Language-grounded indoor 3d semantic segmentation in the wild," in *ECCV*, 2022.
[293] L. Yang, Y. Fan *et al.*, "Video instance segmentation," in *ICCV*, 2019.
[294] A. Athar, J. Luiten *et al.*, "Burst: A benchmark for unifying object recognition, segmentation and tracking in video," in *WACV*, 2023.
[295] C. Szegedy, S. Ioffe *et al.*, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI*, 2017.
[296] K. He, X. Zhang *et al.*, "Deep residual learning for image recognition," in *CVPR*, 2016.
[297] A. Farhadi, I. Endres *et al.*, "Describing objects by their attributes," in *CVPR*, 2009.
[298] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv*, 2014.
[299] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *CVPR*, 2018, pp. 4490–4499.
[300] J. Schult, F. Engelmann *et al.*, "Mask3d: Mask transformer for 3d semantic instance segmentation," in *ICRA*, 2023.
[301] L. Qi, J. Kuen *et al.*, "High-quality entity segmentation," *arXiv*, 2022.
[302] A. Bewley, Z. Ge *et al.*, "Simple online and realtime tracking," in *ICIP*, 2016.
[303] Y. Shen, R. Ji *et al.*, "Enabling deep residual networks for weakly supervised object detection," in *ECCV*, 2020.
[304] G. Zhang, Z. Luo *et al.*, "Accelerating detr convergence via semantic-aligned matching," in *CVPR*, 2022.
[305] Y. Liu, M. Ott *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv*, 2019.
[306] T. Wang, "Learning to detect and segment for open vocabulary object detection," in *CVPR*, 2023.
[307] C. Schuhmann, R. Vencu *et al.*, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv*, 2021.
[308] V. Ordonez, G. Kulkarni *et al.*, "Im2text: Describing images using 1 million captioned photographs," *NeurIPS*, 2011.
[309] C.-Y. Wang, H.-Y. M. Liao *et al.*, "Cspnet: A new backbone that can enhance learning capability of cnn," in *CVPRW*, 2020.
[310] [Online]. Available: https://github.com/ultralytics/yolov5
[311] S. Zhang, C. Chi *et al.*, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *CVPR*, 2020.
[312] X. Zhou, V. Koltun *et al.*, "Probabilistic two-stage detection," *arXiv*, 2021.
[313] A. Brock, S. De *et al.*, "High-performance large-scale image recognition without normalization," in *ICML*, 2021.
[314] X. Zhai, X. Wang *et al.*, "Lit: Zero-shot transfer with locked-image text tuning," in *CVPR*, 2022.
[315] B. Thomee, D. A. Shamma *et al.*, "Yfcc100m: The new data in multimedia research," *ACM Communications*, 2016.
[316] X. Dai, Y. Chen *et al.*, "Dynamic head: Unifying object detection heads with attentions," in *CVPR*, 2021.
[317] H. Zhang, F. Li *et al.*, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv*, 2022.
[318] F. Li, H. Zhang *et al.*, "Mask dino: Towards a unified transformer-based framework for object detection and segmentation," in *CVPR*, 2023.
[319] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *CVPR*, 2018.
[320] X. Chen, X. Wang *et al.*, "Pali: A jointly-scaled multilingual language-image model," *arXiv*, 2022.
[321] O. Ronneberger, P. Fischer *et al.*, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.
[322] B. Zhang, Z. Tian *et al.*, "Segvit: Semantic segmentation with plain vision transformers," *NeurIPS*, 2022.
[323] M. Ding, B. Xiao *et al.*, "Davit: Dual attention vision transformers," in *ECCV*, 2022.

Chaoyang Zhu currently is a Ph.D. student at the Department of Computer Science and Engineering, HKUST. He received the M.Sc degree in Computer Technology from Xiamen University in 2023, and the B.Eng. degree in Computer Science and Technology from Hangzhou Dianzi University in 2019. His research interests are computer vision and multimodal learning.

Long Chen received the Ph.D. degree in Computer Science from Zhejiang University in 2020, and the B.Eng. degree in Electrical Information Engineering from Dalian University of Technology in 2015. He is currently an assistant professor at the Department of Computer Science and Engineering, HKUST. He was a postdoctoral research scientist at Columbia University and a senior researcher at Tencent AI Lab. His research interests are computer vision and multimedia.

APPENDIX

TABLE 2: Datasets and evaluation metrics.

Tasks	Datasets (Split of Base/Novel Categories)	Evaluation Metrics
ZSD	Pascal VOC [15] (16/4) COCO [16] (48/17, 65/15) ILSVRC-2017 Detection [176] (177/23) Visual Genome [74] (478/130)	AP_{50}^{N}, AP_{50}^{B}, AP_{50} R@100
ZSSS	Pascal VOC [15] (15/5) Pascal Context [284] (29/4)	$mIoU_B$, $mIoU_N$, hIoU
OVD	COCO [16] (48/17) LVIS [17] (866/337) Objects365 [285] OpenImages [286]	AP_{50}^{N}, AP_{50}^{B}, AP_{50}, AP_r, AP_c, AP_f, AP
OVSS	Pascal VOC [15] (15/5) COCO stuff [287] (156/15) ADE20K-150 [288] (135/15) ADE20K-847 [288] (572/275) Pascal Context-59 [284] Pascal Context-459 [284]	mIoU, $mIoU_N$, $mIoU_B$, hIoU
OVIS	COCO [16] (48/17) ADE20k [288] (135/15) OpenImages [286] (200/100)	mask AP_{50}, mask AP_{50}^{N}, mask AP_{50}^{B}
OVPS	COCO Panoptic [11] (119/14) ADE20k [288] Cityscapes [289]	PQ, SQ, RQ
OV3D	SUN RGB-D [290] ScanNet [291] nuScenes [1]	AP_{25}^{N}, AP_{25}^{B}, AP_{25}
OV3SS	ScanNet [291] nuScenes [1]	mIoU, $mIoU_N$, $mIoU_B$, hIoU
OV3IS	ScanNet200 [292]	AP, AP_{25}, AP_{50}
OVVU	Youtube-VIS [293] BURST [294] LV-VIS [268]	AP, AP^B, AP^N

In this supplementary material, we provide as much as detailed, comprehensive, and fair comparisons of methods for different tasks and settings. We keep track of new works at awesome-ovd-ovs. However, note that the benchmark does not differentiate the subtle nuances such as image backbone initialization weights, with or without background evaluation, and different version of validation sets, *etc*. For precise details, please refer to the original paper.

Evaluation Protocols, Metrics, and Datasets

ZSD and ZSS mainly use two evaluation protocols for assessment: 1) evaluating only on novel classes (**non-generalized**); 2) evaluating on both base and novel classes (**generalized**). The generalized assessment is more challenging and realistic than non-generalized evaluation. It competes novel with base classes and requires model not to overfit on base classes. OVD and OVS mainly adopt the generalized protocol for evaluation. Besides, OVD and OVS introduce the third protocol, termed cross-dataset transfer evaluation (CDTE). Namely, the model is trained on one source dataset and tested on other target datasets without adaptation. Vocabularies of source and target datasets may or may not partially overlap with each other.

The evaluation metric for object detection and instance segmentation is mainly box and mask AP at a certain IoU threshold (AP_{50}, AP_{25}) or integrated over a series of IoU threshold (0.5 to 0.95 with 0.05 as interval). The AP can be divided into AP_B and AP_N considering only base or novel classes. For LVIS [17] dataset, the rare categories are regarded as novel classes, its metric is denoted as AP_r, common and frequent classes are base classes. Additionally, object detection and instance segmentation use recall as a complementary metric. For semantic segmentation, the metric is mIoU only considering either base ($mIoU_B$) or novel ($mIoU_N$) classes. The hamonic mean (hIoU) between $mIoU_B$ and $mIoU_N$ is calculated as the following:

$$\text{hIoU} = \frac{2 * \text{mIoU}_B * \text{mIoU}_N}{\text{mIoU}_B + \text{mIoU}_N}. \tag{5}$$

Note that for ZSD, the AP_{50} may represent the harmonic mean of AP_{50}^{N} and AP_{50}^{B} in some work, we do not differentiate them here. For panoptic segmentation, the metric is panoptic quality [11] (PQ) which can be viewed as a multiplication between segmentation quality (SQ) and recognition quality (RQ). For 3D scene and video understanding, the metrics are mainly inherited from their counterparts in image domain. For a complete dataset and metric list, *c.f.* to Table 2.

TABLE 3: ZSD performance on COCO [16] dataset. IRv2 is InceptionResnetv2 [295], *c.f.* to Table 4 for other notations.

Method	Image Backbone	Semantic Embeddings	AP_{50}^{N}	$AP_{50}^{B}/AP_{50}^{N}/AP_{50}$
48/17 split [18]				
SAN [113]	R50	W2V	5.1	13.9/2.6/4.3
SB [18]	IRv2	-	0.7	-
LAB [18]	IRv2	-	0.3	-
DSES [18]	IRv2	-	0.5	-
MS-Zero [33]	R101	GloVe [24]	12.9	-/-/30.7
PL [32]	R50-FPN	W2V	10.0	35.9/4.1/7.4
CG-ZSD [121]	DN53 [119]	BERT [25]	7.2	-
BLC [116]	RN50	W2V	10.6	42.1/4.5/8.2
ContrastZSD [123]	R101	W2V	12.5	45.1/6.3/11.1
SSB [117]	R101	W2V	14.8	48.9/10.2/16.9
DELO [34]	DN19 [119]	W2VR [20]	7.6	-/-/13.0
RRFS [36]	R101	FT	13.4	42.3/13.4/20.4
65/15 split [32]				
PL [32]	R50-FPN	W2V	12.4	34.1/12.4/18.2
TL [80]	R50-FPN	W2V	14.6	28.8/14.1/18.9
CG-ZSD [121]	DN53	BERT [25]	10.9	-
BLC [116]	R50	W2V	14.7	36.0/13.1/19.2
DPIF-M [122]	R50	W2V	19.8	29.8/19.5/23.6
ContrastZSD [123]	R101	W2V	18.6	40.2/16.5/23.4
SSB [117]	R101	W2V	19.6	40.2/19.3/26.1
SU [126]	R101	FT	19.0	36.9/19.0/25.1
RRFS [36]	R101	FT	19.8	37.4/19.8/26.0

TABLE 4: ZSD performance on Pascal VOC [15] under the non-generalized and generalized evaluation protocol. R denote ResNet [296]. W2V and FT is Word2Vec [23] and FastText [70], respectively.

Method	Image Backbone	Semantic Embeddings	AP_{50}^{N}	$AP_{50}^{B}/AP_{50}^{N}/AP_{50}$
SAN [19]	R50	-	59.1	48.0/37.0/41.8
HRE [120]	DN19 [119]	aPY [297]	54.2	62.4/25.5/36.2
PL [32]	R50-FPN	aPY [297]	62.1	-
BLC [116]	R50	-	55.2	58.2/22.9/32.9
TL [80]	R50-FPN	W2V	66.6	-
MS-Zero [33]	R101	aPY [297]	62.2	-/-/60.1
CG-ZSD [121]	DN53 [119]	BERT [25]	54.8	-
SU [126]	R101	FT	64.9	-
DPIF [122]	R50	aPY [297]	-	73.2/62.3/67.3
ContrastZSD [123]	R101	aPY [297]	65.7	63.2/46.5/53.6
RRFS [36]	R101	FT	65.5	47.1/49.1/48.1

TABLE 5: ZSD performance on ILSVRC-2017 detection [176] and Visual Genome [74] dataset under non-generalized evaluation protocol. R@100 is Recall@100 at IoU threshold 0.5, *c.f.* to Table 4 and Table 3 for other notations.

Method	Image Backbone	Semantic Embeddings	R@100	AP_{50}^{N}
177/23 split [19] for ILSVRC-2017 Detection				
SAN [19]	R50	W2V	-	16.4
ZSDTD [115]	IRv2	Text-Desc	-	24.1
GTNet [35]	R101	FT	-	26.0
SU [126]	R101	FT	-	24.3
478/130 split [19] for Visual Genome				
SB [18]	IRv2	-	4.1	-
LAB [18]	IRv2	-	5.4	-
DESE [18]	IRv2	-	4.8	-
CA-ZSD [114]	R50	GloVe [24]	-	-
ZSDTD [115]	IRv2	Text-Desc	7.2	-
GTNet [35]	R101	W2V	11.3	-
S2V [124]	IRv2	GloVe [24]	11.0	-
DPIF-M [122]	R50	W2V	18.3	1.8

TABLE 6: Zero-shot semantic segmentation performance on Pascal VOC [15] and Pascal Context [284] datasets. ZS3Net [21] randomly samples 2 to 10 novel classes with step size 2, here we only show the results of 4 novel classes. HM denote hamonic mean (hIoU), for other notations, *c.f.* to Table 4 and Table 3.

Method	Image Backbone	Semantic Embeddings	Pascal VOC mIoU (B/N/HM)	Pascal Context mIoU (B/N/HM)
15/5 split [37] for Pascal VOC 29/4 split [40] for Pascal Context				
SPNet-C [37]	R101	W2V & FT	78.0/15.6/26.1	35.1/4.0/7.2
ZS3Net [21]	R101	W2V	77.3/17.7/28.7	33.0/7.7/12.5
VM [130]	VGG16 [298]	GloVe [24]	-/35.6/-	-
CaGNet [40]	R101	W2V & FT	78.4/26.6/39.7	36.1/14.4/20.6
SIGN [133]	R101	W2V & FT	75.4/28.9/41.7	33.7/14.9/20.7
Novel - 4 [21]				
SPNet [37]	R101	W2V & FT	67.3/21.8/32.9	36.3/18.1/24.2
ZS3Net [21]	R101	W2V	66.4/23.2/34.4	37.2/24.9/29.8
CSRL [132]	R101	-	69.8/31.7/43.6	39.8/23.9/29.9
JoEm [38]	R101	W2V	67.0/33.4/44.6	36.9/30.7/33.5
PMOSR [39]	R101	W2V	75.0/44.1/55.5	41.1/43.1/42.1

TABLE 7: Open-vocabulary 3D detection performance on SUN RGB-D [290], ScanNet [291], and nuScenes [1] datasets under the generalized evaluation and CDTE protocol.

Method	Detector	3D Annotations	2D Detector	SUN RGB-D AP_{25}^{N}	SUN RGB-D AP_{25}^{B}	SUN RGB-D AP_{25}	ScanNet AP_{25}^{N}	ScanNet AP_{25}^{N}	ScanNet AP_{25}	nuScenes AP_{25}
OV-3DET [65]	3DETR [7]	✗	Detic [46]	-	-	20.5	-	-	18.0	-
FM-OV3D [66]	3DETR [7]	✓	Grounded-SAM [258]	-	-	21.5	-	-	21.5	-
OpenSight [259]	VoxelNet [299]	✗	Grounding DINO [164]	-	-	-	-	-	-	23.5
CoDA [67]	3DETR [7]	✓	✗	6.7	38.7	13.7	6.5	21.6	9.0	-
L3Det [260]	L3Det [260]	✓	✗	24.6	-	-	25.4	-	-	-

TABLE 8: Open-vocabulary 3D instance segmentation performance on ScanNet200 [292] dataset.

Method	Segmentor	3D Annotations	2D Segmentor	ScanNet200 AP	ScanNet200 AP_{25}	ScanNet200 AP_{50}
OpenMask3D [264]	Mask3D [300]	✓	SAM [196]	12.8	19.0	16.8
MaskClustering [265]	-	✗	CropFormer [301]	12.0	30.1	23.3
Open3DIS [267]	Mask3D [300]	✓	Grounded-SAM [258]	23.7	32.8	29.4

TABLE 9: Open-vocabulary video instance segmentation performance on validation set of Youtube-VIS19 [293] (YTVIS-19), Youtube-VIS21 [293] (YTVIS-21), BURST [294], and LV-VIS [268].

Method	Segmentor	Tracker	Training Source	Prompts	YTVIS-19 AP	YTVIS-19 AP^{B}	YTVIS-19 AP^{N}	YTVIS-21 AP	YTVIS-21 AP^{B}	YTVIS-21 AP^{N}	BURST AP	BURST AP^{B}	BURST AP^{N}	LV-VIS AP	LV-VIS AP^{B}	LV-VIS AP^{N}
OV2Seg [268]	-	SORT [302]	LVIS [17]	L (cat)	37.6	41.1	21.3	33.9	36.7	18.2	4.9	5.3	3.0	21.1	27.5	16.3
OpenVIS [69]	M2F [14]	✗	YTVIS [293]	T (cat)	-	-	-	-	-	-	3.5	5.8	3.0	-	-	-
BriVIS [269]	M2F [14]	✗	LV-VIS	T (cat)	45.3	-	-	39.5	-	-	7.4	9.5	6.9	27.68	-	-

TABLE 10: OVD performance on COCO [16] under generalized evaluation protocol. "T" and "L" denote template and learnable prompts. "cat" and "desc" denote that the prompts are filled with class names or class descriptions (definitions, synonyms, *etc*). "Ensemble" represents that whether detector prediction is ensembled with CLIP prediction [26], [49] or not. COCO Cap is COCO Captions dataset [136], Visual Genome [74] is denoted as VG, Conceptual Captions [134] is denoted as CC3M.

Method	Image Backbone	Detector	Image-Text Pairs	Text Encoder	Prompts	Ensemble	AP_{50}^{N}	AP_{50}^{B}	AP_{50}
Region-Aware Training									
OVR-CNN [27]	R50	FRCNN	COCO Cap	BERT	✗	✗	22.8	46.0	39.9
LocOv [137]	R50	FRCNN	COCO Cap	BERT	✗	✗	28.6	51.3	45.7
MMC-Det [142]	R50	FRCNN	COCO Cap	BERT	✗	✗	33.5	-	47.5
WSOVOD [145]	DRN [303]	FRCNN	✗	CLIP	T (cat)	✗	35.0	27.9	29.8
RO-ViT [146]	ViT-B/16	MRCNN	ALIGN [102]	CLIP	T (cat)	✓	30.2	-	41.5
CFM-ViT [147]	ViT-B/16	MRCNN	ALIGN [102]	CLIP	T (cat)	✓	30.8	-	42.4
DITO [148]	ViT-B/16	FRCNN	ALIGN [102]	CLIP	T (cat)	✓	38.6	-	48.5
VLDet [42]	R50	FRCNN	COCO Cap	CLIP	T (cat)	✗	32.0	50.6	45.8
GOAT [149]	R50	FRCNN	COCO Cap	CLIP	T (cat)	✗	31.7	51.3	45.7
OV-DETR [43]	R50	Def-DETR [99]	✗	CLIP	T (cat)	✗	29.4	61.0	52.7
Prompt-OVD [150]	ViT-B/16	Def-DETR [99]	✗	CLIP	T (cat)	✓	30.6	63.5	54.9
CORA [151]	R50	SAM-DETR [304]	✗	CLIP	T (cat)	✗	35.1	35.5	35.4
EdaDet [152]	R50	Def-DETR [99]	CLIP [31]	CLIP	T (cat)	✓	35.1	35.5	35.4
SGDN [156]	R50	Def-DETR [99]	VG, Flickr30K [82]	RoBERTa [305]	✗	✗	37.5	61.0	54.9
Pseudo-Labeling									
RegionCLIP [157]	R50	FRCNN	CC3M	CLIP	T (cat)	✗	31.4	57.1	50.4
CondHead [306]	R50	RegionCLIP [157]	✗	CLIP	T (cat)	✗	33.7	58.0	51.7
VL-PLM [158]	R50	FRCNN	✗	CLIP	T (cat)	✗	34.4	60.2	53.5
PromptDet [165]	R50	MRCNN	LAION [307]	CLIP	L (cat+desc)	✗	26.6	-	50.6
SAS-Det [168]	R50	FRCNN	✗	CLIP	T (cat)	✓	37.4	58.5	53.0
PB-OVD [45]	R50	MRCNN	COCO Cap, VG, SBU [308]	CLIP	T (cat)	✗	30.8	46.1	42.1
CLIM [172]	R50	Detic [46]	COCO Cap	CLIP	T (cat)	✗	35.4	-	-
VTP-OVD [173]	R50	MRCNN	✗	CLIP	T (cat)	✗	31.5	51.9	46.6
ProxyDet [174]	R50	FRCNN	COCO Cap	CLIP	T (cat)	✓	30.4	52.6	46.8
CoDet [175]	R50	FRCNN	COCO Cap	CLIP	T (cat)	✓	30.6	52.3	46.6
Detic [46]	R50	FRCNN	COCO Cap	CLIP	T (cat)	✗	27.8	47.1	45.0
Knowledge Distillation									
ViLD [26]	R50	MRCNN	✗	CLIP	T (cat)	✓	27.6	59.5	51.3
ZSD-YOLO [181]	CSP [309]	YOLOv5x [310]	✗	CLIP	T (cat+desc)	✗	13.6	31.7	19.0
LP-OVOD [182]	R50	FRCNN	✗	CLIP	T (cat)	✓	40.5	60.5	55.2
EZSD [183]	R50	MRCNN	✗	CLIP	T (cat)	✗	31.6	59.9	52.1
SIC-CADS [184]	R50	BARON [48]	✗	CLIP	T (cat)	✓	36.9	56.1	51.1
BARON [48]	R50	FRCNN	COCO Cap	CLIP	T (cat)	✗	42.7	54.9	51.7
OADP [186]	R50	FRCNN	✗	CLIP	T (cat)	✓	35.6	55.8	50.5
RKDWTF [170]	R50	FRCNN	COCO Cap	CLIP	T (cat)	✗	36.6	54.0	49.4
DK-DETR [188]	R50	Def-DETR [99]	✗	CLIP	T (cat)	✗	32.3	61.1	-
HierKD [189]	R50	ATSS [311]	CC3M	CLIP	T (cat/desc)	✗	20.3	51.3	43.2
CLIPSelf [110]	ViT-B/16	F-VLM [49]	✗	CLIP	T (cat)	✓	37.6	-	-
Transfer Learning									
F-VLM [49]	R50	MRCNN	✗	CLIP	T (cat)	✓	28.0	-	39.6
DRR [194]	R50	FRCNN	CC3M	CLIP	T (cat)	✗	35.8	54.6	49.6

TABLE 11: OVD performance on LVIS [17] dataset under generalized evaluation protocol. Base classes are common and frequent classes in LVIS, rare classes in LVIS are novel class. Gray numbers denote mask AP. Conceptual 12M dataset [135] is denoted as CC12M. The subset of ImageNet-21K [176] (IN21K) that overlaps with LVIS vocabulary is IN-L [46] (*c.f.* to Table 10 for other abbreviatios).

Method	Image Backbone	Detector	Image-Text Pairs	Text Encoder	Prompts	Ensemble	AP_r	AP_c	AP_f	AP
Region-Aware Training										
MMC-Det [142]	R50	CN2 [312]	CC3M	CLIP	T (cat)	✗	21.1	30.9	35.5	31.0
RO-ViT [146]	ViT-B/16	MRCNN	ALIGN [102]	CLIP	T (cat)	✓	28.0	-	-	30.2
CFM-ViT [147]	ViT-B/16	MRCNN	ALIGN [102]	CLIP	T (cat)	✓	$29.6_{28.8}$	-	-	$33.8_{32.0}$
DITO [148]	ViT-B/16	FRCNN	ALIGN [102]	CLIP	T (cat)	✓	$34.9_{32.5}$	-	-	$36.9_{34.0}$
VLDet [42]	R50	CN2 [312]	CC3M	CLIP	T (cat)	✗	21.7	29.8	34.3	30.1
GOAT [149]	R50	CN2 [312]	CC3M	CLIP	T (cat)	✗	23.3	29.7	34.3	30.4
OV-DETR [43]	R50	Def-DETR [99]	✗	CLIP	T (cat)	✗	17.4	25.0	32.5	26.6
Prompt-OVD [150]	ViT-B/16	Def-DETR [99]	✗	CLIP	T (cat)	✗	$29.4_{23.1}$	-	-	$33.0_{24.2}$
CORA [151]	R50x4	CN2 [312]	✗	CLIP	T (cat)	✗	28.1	-	-	-
EdaDet [152]	R50	Def-DETR [99]	✗	CLIP	T (cat)	✓	23.7	27.5	29.1	27.5
SGDN [156]	R50	Def-DETR [99]	VG, Flickr30K [82]	RoBERTa [305]	✗	✗	23.6	29.0	34.3	31.1
Pseudo-Labeling										
RegionCLIP [157]	R50	MRCNN	CC3M	CLIP	T (cat)	✗	$17.1_{17.4}$	$27.4_{26.0}$	$34.0_{31.6}$	$28.2_{26.7}$
CondHead [306]	R50	RegionCLIP [157]	✗	CLIP	T (cat)	✗	$19.9_{20.0}$	$28.6_{27.3}$	$35.2_{32.2}$	$29.7_{27.9}$
PromptDet [165]	R50	MRCNN	LAION [307]	CLIP	L (cat+desc)	✗	21.4	23.3	29.3	25.3
SAS-Det [168]	R50	FRCNN	✗	CLIP	T (cat)	✓	20.9	26.1	31.6	27.4
CLIM [172]	R50	VLDet [46]	CC3M	CLIP	T (cat)	✗	22.2	-	-	-
ProxyDet [174]	R50	CN2 [312]	IN-L	CLIP	T (cat)	✓	26.2	-	-	32.5
CoDet [175]	R50	CN2 [312]	CC3M	CLIP	T (cat)	✗	23.4	30.0	34.6	30.7
Detic [46]	R50	CN2 [312]	IN-L	CLIP	T (cat)	✗	24.6	-	-	32.4
MMC [177]	R50	CN2 [312]	IN-L	CLIP	GPT-3 [178]	✗	27.3	-	-	33.1
3Ways [169]	NF-F0 [313]	FCOS [98]	CC12M	CLIP	T (cat)	✗	25.6	34.2	41.8	35.7
PLAC [179]	Swin-B	Def-DETR [99]	CC3M	CLIP	T (cat)	✗	27.0	40.0	44.5	39.5
Knowledge Distillation										
ViLD-ens [26]	R50	MRCNN	✗	CLIP	T (cat)	✓	$16.7_{16.6}$	$26.5_{24.6}$	$34.2_{30.3}$	$27.8_{25.5}$
LP-OVOD [182]	R50	MRCNN	✗	CLIP	T (cat)	✓	19.3	26.1	29.4	26.2
EZSD [183]	R50	MRCNN	✗	CLIP	T (cat)	✗	15.8	25.6	31.7	26.3
SIC-CADS [184]	R50	Detic [46]	IN21K	CLIP	T (cat)	✓	26.5	33.0	35.6	32.9
BARON [48]	R50	FRCNN	✗	CLIP	L (cat)	✗	$23.2_{22.6}$	$29.3_{27.6}$	$32.5_{29.8}$	$29.5_{27.6}$
OADP [186]	R50	FRCNN	✗	CLIP	T (cat)	✓	$21.9_{21.7}$	$28.4_{26.3}$	$32.0_{29.0}$	$28.7_{26.6}$
GridCLIP [187]	R50	FCOS [98]	✗	CLIP	T (cat)	✗	15.0	22.7	32.5	25.2
RKDWTF [170]	R50	CN2 [312]	IN21K	CLIP	T (cat)	✗	25.2	33.4	35.8	32.9
DK-DETR [188]	R50	Def-DETR [99]	✗	CLIP	T (cat)	✗	$22.2_{20.5}$	$32.0_{28.9}$	$40.2_{35.4}$	$33.5_{30.0}$
DetPro [47]	R50	MRCNN	✗	CLIP	L (cat)	✓	$20.8_{19.8}$	$27.8_{25.6}$	$32.4_{28.9}$	$28.4_{25.9}$
CLIPSelf [240]	ViT-B/16	F-VLM [49]	✗	CLIP	T (cat)	✓	25.3	-	-	-
Transfer Learning										
OWL-ViT [190]	ViT-H/14	DETR	LiT [314]	CLIP	T (cat)	✗	23.3	-	-	35.3
F-VLM [49]	R50	MRCNN	✗	CLIP	T (cat)	✓	18.6	-	-	24.2

TABLE 12: OVD performance under the CDTE protocol on Pascal VOC [15] (VOC), Obejcts365 [285] (O365), COCO [16], OpenImages [286] (OI), and LVIS [17] validation sets. Cap4M image-text pairs are crawled in [44]. GoldG denote the merged grounding datasets in [41], [44] (*c.f.* to Table 11 for other notations). Note that some methods evaluate on different versions of O365 and OI validation datasets, we do not differentiate them here. Gray Numbers denote the performance of LVIS *minival* set [41]. All metrics are box AP.

Method	Image Backbone	Detector	Training Source	VOC	COCO		O365		OI	LVIS
				AP_{50}	AP	AP_{50}	AP	AP_{50}	AP_{50}	AP_r/AP
Region-Aware Training										
MMC-Det [142]	R50	CN2 [312]	LVIS, CC3M	-	-	56.4	-	21.4	38.6	-
DetCLIP [143]	Swin-T	ATSS [311]	O365, GoldG, YFCC1M [315]	-	-	-	-	-	-	$25.0_{33.2}$/$28.4_{35.9}$
DetCLIPv2 [144]	Swin-T	ATSS [311]	O365, GoldG, CC3M, CC12M	-	-	-	-	-	-	36.0/40.4
RO-ViT [146]	ViT-B/16	MRCNN	LVIS, ALIGN [102]	-	-	-	17.1	26.9	-	-
CFM-ViT [147]	ViT-B/16	MRCNN	LVIS, ALIGN [102]	-	-	-	15.9	24.6	-	-
DITO [148]	ViT-L/16	FRCNN	LVIS, ALIGN [102]	-	-	-	19.8	30.4	-	-
OV-DETR [43]	R50	Def-DETR [99]	LVIS	76.1	38.1	58.4	-	-	-	-
EdaDet [152]	R50	Def-DETR [99]	LVIS	-	-	-	13.6	19.8	-	-
MDETR [41]	R101	DETR	GoldG+ [41]	-	-	-	-	-	-	$7.4_{20.9}$/$22.5_{24.2}$
MQ-Det [154]	Swin-T	GLIP [44]	O365	-	-	-	-	-	-	$15.4_{21.0}$/$22.6_{30.4}$
YOLO-World [155]	-	YOLOv8-L	O365, GoldG	-	-	-	-	-	-	27.1/35.0
SGDN [156]	R50	Def-DETR [99]	LVIS, Flickr30K, VG	-	40.5	-	-	-	-	-
Pseudo-Labeling										
GLIP [44]	Swin-T	DyHead [316]	O365, GoldG, Cap4M	-	46.3	-	-	-	-	$10.1_{20.8}$/$17.2_{26.0}$
GLIPv2 [159]	Swin-T	DyHead [316]	O365, GoldG, Cap4M	-	-	-	-	-	-	-/29.0
Grounding DINO [164]	Swin-T	DINO [317]	O365, GoldG, Cap4M	-	48.4	-	-	-	-	18.1/27.4
PB-OVD [45]	R50	MRCNN	COCO, COCO Cap VG, SBU [308]	59.2	-	-	6.9	-	-	-
VTP-OVD [173]	R50	MRCNN	COCO	61.1	-	-	-	7.4	-	-
ProxyDet [174]	R50	CN2 [312]	LVIS	-	-	57.0	-	19.1	-	-
CoDet [175]	R50	CN2 [312]	LVIS, CC3M	-	39.1	57.0	14.2	20.5	-	-
Detic [46]	Swin-B	CN2 [312]	LVIS, IN21K	-	-	-	-	21.5	55.2	-
MMC (text) [177]	R50	CN2 [312]	IN-L, LVIS	-	-	-	16.6	23.1	-	-
3Ways [169]	NF-F0 [313]	FCOS	LVIS	-	41.5	-	16.4	-	-	-
Knowledge Distillation										
ViLD [26]	R50	MRCNN	LVIS	72.2	36.6	55.6	11.8	18.2	-	-
CondHead [306]	R50	ViLD [26]	LVIS	74.6	39.1	59.1	13.2	20.4	-	-
SIC-CADS [184]	R50	Detic [46]	LVIS	-	-	-	-	31.2	54.7	-
BARON [48]	R50	FRCNN	LVIS	76.0	36.2	55.7	13.6	21.0	-	-
GridCLIP [187]	R50	FCOS	LVIS	70.9	34.7	52.2	-	-	-	-
RKDWTF [170]	R50	MRCNN	IN21K, LVIS	-	-	56.6	-	22.3	42.9	-
DK-DETR [188]	R50	Def-DETR [99]	LVIS	71.3	39.4	54.3	12.4	17.3	-	-
DetPro [47]	R50	MRCNN	LVIS	74.6	34.9	53.8	12.1	18.8	-	-
CLIPSelf [110]	ViT-L/14	F-VLM [49]	LVIS	-	40.5	63.8	19.5	31.3	-	-
Transfer Learning										
OWL-ViT [190]	ViT-B/16	DETR	O365, VG	-	-	-	-	-	-	23.6/26.7
UniDetector [191]	R50	FRCNN	COCO, O365, OI	-	-	-	-	-	-	18.0/19.8
F-VLM [49]	R50	MRCNN	LVIS	-	32.5	53.1	11.9	19.2	-	-
OpenSeeD [193]	Swin-T	Mask DINO [318]	COCO, O365	-	-	-	-	-	-	-/21.8
Sambor [195]	ViT-B	Cascade R-CNN [319]	O365	-	48.6	66.1	-	-	-	$20.9_{29.6}$/$26.3_{33.1}$

TABLE 13: Open-vocabulary semantic segmentation performance on the validation set of ADE20K [288] (A-847 and A-150), Pascal Context [15] (PC-459 and PC-59), Pascal VOC [15] (PAS-20), Cityscapes [289] (CS-19), COCO Stuff [287] (Stuff), and COCO [16] datasets under the CDTE protocol. MF is MaskFormer [9], *c.f.* to Tables 10 and 11 for other notations.

Method	Image Backbone	Segmentor	Training Source	Text Encoder	Prompts	Ensemble	mIoU							
							A-847	A-150	PC-459	PC-59	PAS-20	CS-19	Stuff	COCO
Region-Aware Training														
OpenSeg [29]	R101	MF	COCO Pan, COCO Cap	BERT	cat+desc	✗	4.0	15.3	6.5	36.9	60.0	-	-	-
SLIC [197]	ViT-B/16	CAT-Seg [234]	WebLI [320], COCO Stuff	CLIP	T (cat)	✗	13.4	36.6	22.0	61.2	95.9	-	-	-
GroupViT [50]	ViT-S	-	CC12M, YFCC14M [315]	CLIP	T (cat)	✗	-	-	-	22.4	52.3	-	-	-
ViL-Seg [199]	ViT-B/16	-	CC12M	CLIP	T (cat)	✗	-	-	-	16.3	34.4	-	16.4	-
SegCLIP [202]	ViT	-	COCO Cap, CC3M	CLIP	T (cat)	✗	-	-	-	24.7	52.6	-	-	26.5
OVSegmentor [203]	ViT-B	-	CC4M [203]	BERT	T (cat)	✗	-	5.6	-	20.4	53.8	-	-	25.1
PACL [205]	ViT-B/16	-	CC3M, CC12M, YFCC15M [315]	CLIP	T (cat)	✗	-	31.4	-	50.1	72.3	-	38.8	-
TCL [206]	ViT-B/16	-	CC3M, CC12M	CLIP	T (cat)	✗	-	17.1	-	33.9	83.2	24.0	22.4	31.6
SimSeg [207]	ViT-B/16	-	CC3M, CC12M	CLIP	T (cat)	✗	-	-	-	26.2	57.4	-	29.7	-
Knowledge Distillation														
GKC [52]	R50	MF	COCO Pan, COCO Cap	CLIP	T (cat+desc)	✗	3.2	17.5	6.5	41.9	78.7	34.3	-	-
SAM-CLIP [53]	ViT-B	SAM [196]	CC3M, CC12M, YFCC15M [315], IN21K, SA-1B [196]	CLIP	T (cat)	✗	-	-	-	29.2	60.6	17.1	31.5	-
ZeroSeg [209]	ViT	-	IN1K [176]	CLIP	T (cat)	✗	-	-	-	20.4	40.8	-	-	20.2
Transfer Learning														
LSeg+ [29]	R101	SRB [30]	COCO Pan	CLIP	T (cat)	✗	2.5	13.0	5.2	36.0	59.0	-	-	-
CEL [212]	R50	MF	COCO Cap, COCO Stuff	CLIP	T (cat)	✓	7.2	20.5	9.6	49.6	86.7	-	-	-
ZSSeg [217]	R101	MF	COCO Stuff	CLIP	L (cat)	✓	7.0	20.5	-	47.7	-	34.5	-	-
MaskCLIP [218]	R101	DL [100]	✗	CLIP	T (cat)	✗	-	-	-	25.5	-	-	14.6	-
CLIP-DINOiser [219]	ViT-B/16	-	Pascal VOC	CLIP	T (cat)	✗	-	20.0	-	35.9	80.2	31.7	-	-
MVP-SEG [221]	R50	DL [100]	COCO Stuff	CLIP	L (cat)	✗	-	-	-	38.7	-	-	-	-
ReCo [213]	-	-	-	CLIP	T (cat)	✗	-	-	-	-	-	19.3	26.3	-
OVDiff [223]	UNet [321]	-	CLIP [31], Stable Diffusion [108]	✗	T (cat)	✗	-	-	-	30.1	67.1	-	-	34.8
FOSSIL [224]	ViT-L/14	-	COCO Cap	CLIP	T (cat)	✗	-	-	-	35.8	-	23.2	24.8	-
POMP [225]	R101	MF	COCO Stuff	CLIP	L (cat)	✗	-	20.7	-	51.1	-	-	-	-
AttrSeg [226]	R101	-	COCO Stuff	CLIP	desc	✗	-	-	-	56.3	91.6	-	-	-
PnP-OVSS [227]	ViT-L/16	BLIP [230]	COCO Cap	BERT	T (cat)	✗	-	23.2	-	41.9	55.7	-	32.6	33.8
SCAN [216]	Swin-B	M2F	COCO Stuff	CLIP	T (cat)	✗	10.8	30.8	13.2	58.4	97.0	-	-	-
TagAlign [228]	ViT-B/16	-	CC12M	CLIP	T (cat)	✗	-	17.3	-	37.6	87.9	27.5	-	33.3
Self-Seg [229]	ViT-L	X-Dec [232]	COCO Cap	-	-	✗	6.4	-	-	-	-	41.1	-	-
OVSeg [55]	R101c [100]	MF	COCO Stuff, COCO Cap	CLIP	T (cat)	✓	7.1	24.8	11.0	53.3	92.6	-	-	-
CAT-Seg [234]	Swin-B	-	COCO Stuff	CLIP	T (cat)	✗	10.8	31.5	20.4	62.0	96.6	-	-	-
SED [235]	ConvNext-B [236]	-	COCO Stuff	CLIP	T (cat)	✗	11.4	31.6	18.6	57.3	94.4	-	-	-
MAFT [237]	ViT-B/16	FreeSeg [248]	COCO Stuff	CLIP	T (cat)	✗	10.1	29.1	12.8	53.5	90.0	-	-	-
SAN [56]	ViT-B/16	-	COCO Stuff	CLIP	T (cat)	✗	10.1	27.5	12.6	53.8	94.0	-	-	-
CaR [242]	ViT-B/16	-	✗	CLIP	T (cat)	✗	-	-	-	30.5	67.6	-	-	36.6

TABLE 14: Open-vocabulary semantic segmentation performance under generalized evaluation protocol. HM is the hamonic mean (hIoU), *c.f.* to Tables 10, 11 and 13 for other notations.

Method	Image Backbone	Segmentor	Image-Text Pairs	Text Encoder	Prompts	Ensemble	Pascal VOC mIoU(N/B/HM)	COCO Stuff mIoU(N/B/HM)
CEL [212]	R50	MF	COCO Cap	CLIP	T (cat)	✓	74.8/88.5/81.1	42.0/38.6/40.2
ZegFormer [54]	R101	MF	✗	CLIP	T (cat)	✓	63.6/86.4/73.3	33.2/36.6/34.8
ZSSeg [217]	R101	MF	✗	CLIP	L (cat)	✓	72.5/83.5/77.5	36.3/39.3/37.8
MVP-SEG+ [221]	R50	DL [100]	✗	CLIP	L (cat)	✗	87.4/89.0/88.2	55.8/38.3/45.5
POMP [225]	R101	MF	✗	CLIP	L (cat)	✗	76.8/93.6/84.4	38.2/39.9/39.1
MAFT [237]	ViT-B/16	FreeSeg [248]	✗	CLIP	T (cat)	✗	81.8/91.4/86.3	50.4/43.3/46.5
ZegCLIP [238]	ViT-B/16	SegViT [322]	✗	CLIP	T (cat)	✗	77.8/91.9/84.3	41.4/40.2/40.8
TagCLIP [239]	ViT-B/16	SegViT [322]	✗	CLIP	T (cat)	✗	85.2/93.5/89.2	43.1/40.7/41.9

TABLE 15: Open-vocabulary instance segmentation performance on COCO [16] and OpenImages [286] datasets under the gOVE protocol. The metric is mask AP. M2F is Mask2Former [14], *c.f.* to Tables 10 and 11 for other notations.

Method	Image Backbone	Segmentor	Image-Text pairs	Text Encoder	Prompts	Ensemble	COCO			OpenImages		
							AP^N_{50}	AP^B_{50}	AP_{50}	AP^N_{50}	AP^B_{50}	AP_{50}
CGG [57]	R50	M2F	COCO Cap	BERT	✗	✗	28.4	46.0	41.4	-	-	-
D^2Zero [244]	R50	M2F	✗	CLIP	T (cat)	✗	15.8	54.1	24.5	-	-	-
XPM [28]	R50	MRCNN	CC3M	BERT	✗	✗	21.6	41.5	36.3	22.7	49.8	40.7
Mask-free OVIS	R50	MRCNN	COCO, OI	ALBEF [141]	✗	✗	25.0	-	-	25.8	-	-

TABLE 16: Open-vocabulary panoptic segmentation performance on COCO Panoptic [11] and ADE20k [288] dataset. PQ^{st} and PQ^{th} represent PQ for stuff and thing classes, respectively. For other notations, *c.f.* to Tables 10, 11 and 15.

Method	Image Backbone	Segmentor	Text Encoder	Prompts	Ensemble	COCO Panoptic						ADE20K				
						PQ^B	SQ^B	RQ^B	PQ^N	SQ^N	RQ^N	PQ	PQ^{th}	PQ^{st}	SQ	RQ
PADing [60]	R50	M2F	CLIP	T (cat)	✗	41.5	80.6	49.7	15.3	72.8	18.4	-	-	-	-	-
FreeSeg [248]	R101	M2F	CLIP	L (cat)	✗	31.4	78.3	38.9	29.8	79.2	37.6	-	-	-	-	-
MaskCLIP [256]	R50	M2F	CLIP	cat	✗	-	-	-	-	-	-	15.1	13.5	18.3	70.5	19.2
OPSNet [257]	R50	M2F	CLIP	cat	✗	-	-	-	-	-	-	17.7	15.6	21.9	54.9	21.6

TABLE 17: Open-vocabulary panoptic segmentation performance under the CDTE protocol. For notations, *c.f.* to Table 16.

Method	Image Backbone	Segmentor	Text Encoder	Prompts	Training Source	Ensemble	COCO Pan			ADE20K			Cityscapes		
							PQ	SQ	RQ	PQ	SQ	RQ	PQ	SQ	RQ
Uni-OVSeg [247]	ConvNext-L [236]	M2F	CLIP	T (cat)	*c.f.* to [247]	✗	18.0	72.6	24.3	14.1	66.1	19.0	17.5	65.2	23.5
X-Decoder [232]	DaViT-B [323]	M2F	CLIP	T (cat)	CC3M, SBU [308], COCO Cap, VG	✗	-	-	-	21.1	-	-	39.5	-	-
APE [59]	ViT-L	-	CLIP	T (cat+desc)	*c.f.* to [59]	✗	-	-	-	26.1	-	-	32.8	-	-
FC-CLIP [61]	ConvNext-L [236]	M2F	CLIP	T (cat)	COCO Pan or ADE20K	✓	27.0	78.0	32.9	26.8	-	-	44.0	75.4	53.6
PosSAM [251]	ViT-B	M2F	CLIP	T (cat)	COCO Pan or ADE20K	✗	25.1	80.1	30.4	26.5	74.7	32.0	-	-	-
MasQCLIP [252]	R50	M2F	CLIP	cat	COCO Pan	✗	-	-	-	23.3	-	-	-	-	-
OMG-Seg [253]	ConvNext-L [236]	M2F	CLIP	T (cat)	*c.f.* to [253]	✗	53.8	-	-	-	-	-	65.7	-	-
ODISE [62]	UNet [321]	M2F	CLIP	T (cat+desc)	COCO Pan, COCO Cap	✓	-	-	-	23.4	-	-	-	-	-
HIPIE [255]	R50	Def-DETR [99], MaskDINO [318]	BERT	cat	*c.f.* to [255]	✗	-	-	-	18.4	-	-	-	-	-