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

Chaoyang Zhu, Long Chen

Abstract—As the most fundamental tasks of computer vision, object detection and segmentation have made tremendous progress in the deep learning era. Due to the expensive manual labeling, the annotated categories in existing datasets are often small-scale and pre-defined, *i.e.*, state-of-the-art detectors and segmentors fail to generalize beyond the closed-vocabulary. To resolve this limitation, the last few years have witnessed increasing attention toward Open-Vocabulary Detection (OVD) and Segmentation (OVS). In this survey, we provide a comprehensive review on the past and recent development of OVD and OVS. To this end, we develop a taxonomy according to the type of task and methodology. 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-based, and transfer learning-based. The proposed taxonomy is universal across different tasks, covering object detection, semantic/instance/panoptic segmentation, 3D scene and video understanding. In each category, its main principles, key challenges, development routes, strengths, and weaknesses are thoroughly discussed. In addition, we benchmark each task along with the vital components of each method. Finally, several promising directions are provided to stimulate future research.

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

1 Introduction

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

The past decade has witnessed a steady progress in object detection and segmentation tasks brought by advanced deep neural architectures, such as Convolutional Neural Networks (CNNs) [6], [8], [9], [11], [12], [13], [14], [15], [16], [17] and Transformer-based models [18], [19], [20], [21], [22], [23], [24], [25]. However, existing object detectors and segmentors can only localize pre-defined semantic concepts (or categories) in each specific dataset, the number of which is typically at small-scale, *e.g.*, 20 classes in Pascal VOC [26], 80 in COCO [27], even the largest dataset LVIS [28] 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 localization limitation hinders the utilization of current detectors and segmentors in the wild.

To resolve the closed-vocabulary constraint for object localization tasks, research endeavors have been devoted to zero-shot or open-vocabulary detection and segmentation. In the early stage of development, zero-shot detection (ZSD) [29], [30], [31] and segmentation (ZSS) [32], [33] is first proposed as an attempt without accessing any unannotated unseen visual samples. To achieve this goal, current mainstream ZSD and ZSS methods always replace the learnable weights of the "classifier" with fixed class semantic embeddings, *e.g.*, Word2Vec [34] (W2V), FastText [35] (FT), GloVe [36], or from BERT [37], which can be leveraged to transfer knowledge from seen (base) categories to unseen (novel) ones. However, due to the unsupervised training only on text corpus, these semantic embeddings lack the alignment with visual features thus they are noisy to serve as anchors for the visual space to calibrate [38], [39]. Later the newly formulated open-vocabulary detection (OVD) [40] and segmentation (OVS) [41], [42], [43] allow the model to train on images with unannotated novel objects. They typically address the closed-set limitation via weak supervision signals, *i.e.*, image-text pairs (image-caption or image-level labels), or large pretrained Vision-Language Models (VLMs), such as CLIP [44]. The text embeddings from the text encoder of CLIP [44] can well align with the visual modality. Therefore, OVD and OVS achieve a huge leap in performance compared to ZSD and ZSS. Due to its 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, which focus on limited tasks and settings[1], hence a more comprehensive survey covering all tasks and settings is of urgent need.

In this paper, we provide a comprehensive review

• *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.

1. In this survey, we regard "zero-shot" and "open-vocabulary" as two different settings.

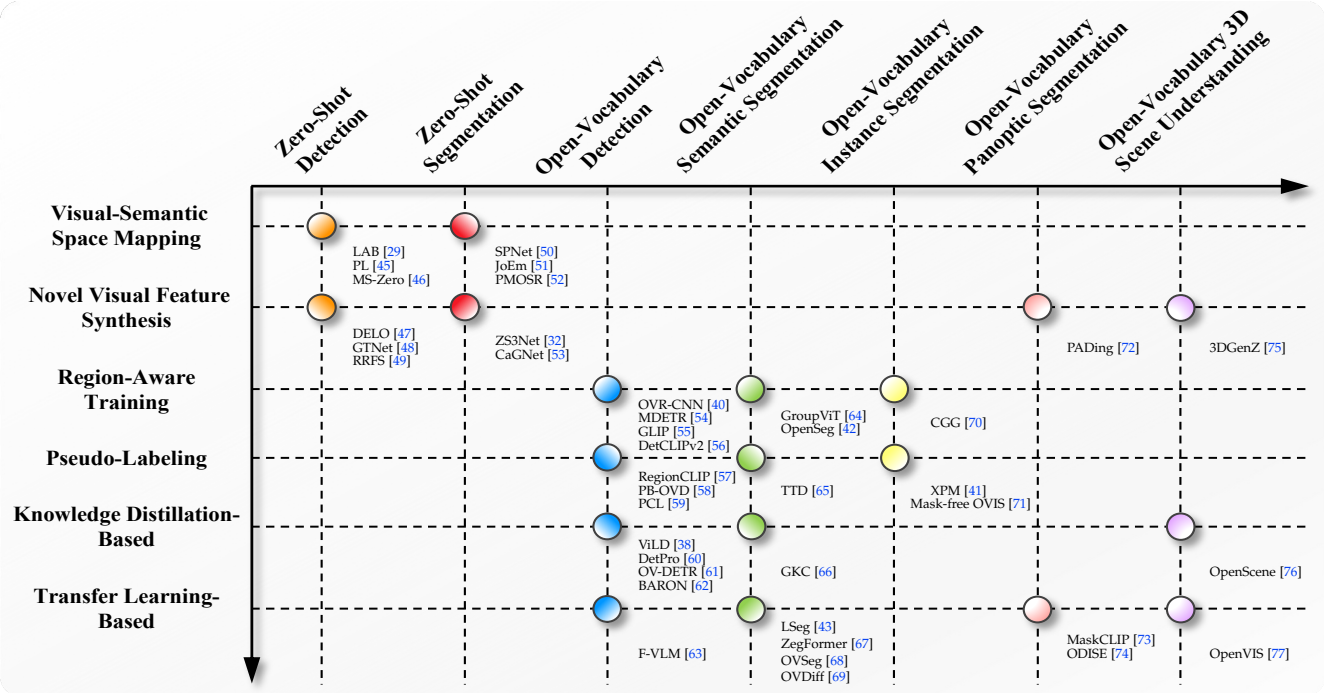

Fig. 1: The taxonomy based on tasks and methodologies. Typical models are shown in each category.

on different 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, in this survey, 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, how to utilize them is key to categorization. Hence, we build a taxonomy according to tasks and methodologies shown in Fig. 1. Zero-shot and open-vocabulary settings are differentiated by the permission of weak supervision signals, and different tasks share the same methodologies under each setting. The general frameworks of different methodologies are summarized in Fig. 2.

Concretely, ZSD and ZSS can be coarsely grouped into:

1) Visual-Semantic Space Mapping. Though the visual and semantic space may bear discriminative capabilities in one modality, there are no direct cross-modality training mechanisms mining mutual relationships between the two spaces. Therefore, 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 a reliable cross-space similarity measurement. Two main architectural modifications are made to canonical closed-set detectors/segmentors: 1) the learnable classifier is substituted with fixed semantic embeddings (*e.g.*, W2V [36]/GloVe [36]/FT [35]/BERT [37]); 2) the class-specific localization branch is switched to class-agnostic, relying on its generalization ability to discover novel objects.

2) Novel Visual Feature Synthesis. Due to the lack of annotations of unseen categories, the confidence of unseen classes in previous methodology is always overwhelmed by seen classes. To alleviate the bias issue [78], this solution utilizes an additional generative model [79], [80], [81] to synthesize fake unseen visual features conditioned on semantic embeddings and random noise vectors. The generation loss is to approximate the underlying distribution of real visual features. Then the classifier embedded in the detection or segmentation head is retrained on both pristine real seen and generated unseen visual features.

Once allowed to access the weak supervision signals, OVD and OVS methodologies can be mainly categorized into four types:

1) Region-Aware Training. This line of work aims to align regions and words implicitly on the cheap and abundant image-caption pairs besides the ground-truth datasets. Its main characteristic is the imposed bi-directional weakly-supervised grounding or contrastive losses that pull regions and words within the same image-text pair close while pushing away other negatives inside a batch. Through these additional region-aware losses, models can learn the cross-modality alignment and expand the vocabulary. In terms of architecture, it adopts text embeddings from the text encoder of VLMs instead of semantic embeddings trained only on text corpus in the zero-shot setting.

2) Pseudo-Labeling. Pseudo-labeling methodology also leverages image-text pairs besides ground-truths but it explicitly constructs pseudo region-text pairs to learn the correspondence 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, instead of the soft alignment in region-aware training where one word may correspond to multiple regions weighted by *softmax*. Pesudo-labeling can adopt VLMs as the teacher to produce pseudo labels, but it can also be deemed as self-training without VLMs (the teacher is the model itself). Note that region-aware training does not utilize the image encoder of VLMs either.

3) Knowledge Distillation-Based. VLMs (such as CLIP)

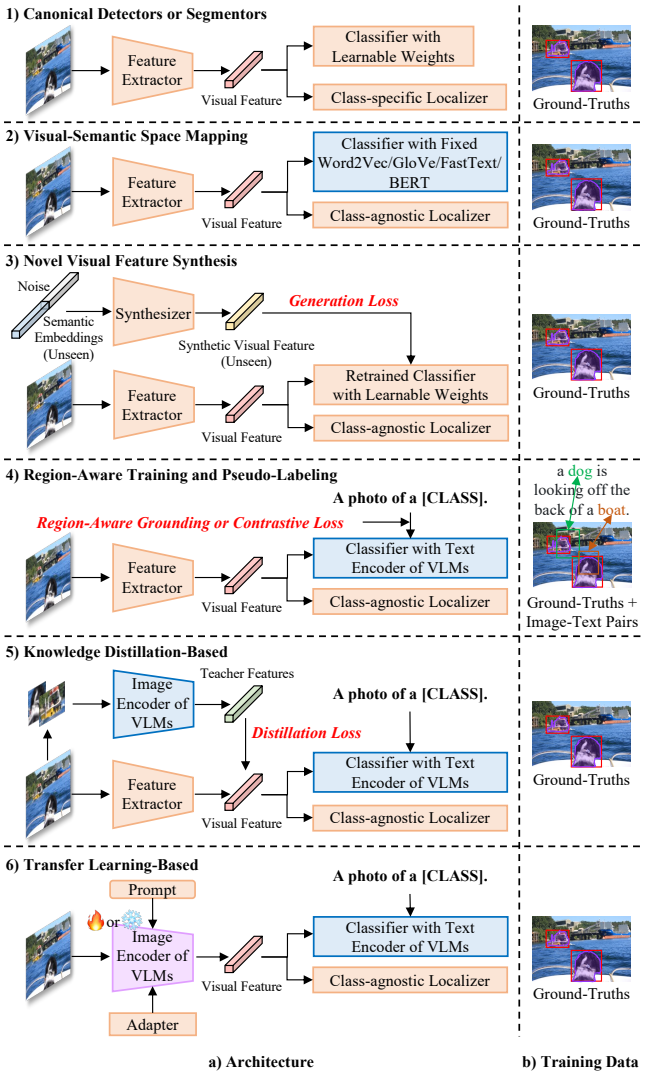

Fig. 2: A general comparison of each methodology.

trained via contrastive learning yield superior zero-shot recognition ability across a variety of downstream tasks. Methods in this group mainly distill the region embeddings from the teacher model, *i.e.*, VLMs image encoder, into the student model to make them compatible with text embeddings of VLMs using detection or segmentation data. Compared to region-aware training and pseudo-labeling, it does not train on image-text pairs but requires the image encoder of VLMs.

4) Transfer Learning-Based. Since knowledge distillation requires repeatedly forwarding each region-of-interest (RoI) into the VLMs image encoder, it inevitably induces a heavy memory consumption. Transfer learning-based models add negligible extra computation overhead to VLMs image encoder. They can be further categorized into: 1) the frozen image encoder of VLMs as feature extractor; 2) fine-tune the VLMs image encoder on downstream data; 3) freeze the image encoder of VLMs and train learnable visual prompts [82] on ground-truths; 4) train a lightweight adapter attached to the frozen image encoder of VLMs on downstream datasets. A more detailed framework for transfer learning-based models is given in Fig. 5.

In this survey, we use the term "open-vocabulary" to summarize methods from both traditional zero-shot and the newly emerged open-vocabulary settings for the following reasons: 1) both zero-shot and open-vocabulary settings enable detection and segmentation beyond a fixed vocabulary; 2) the methodology between the two settings can be shared, *e.g.*, the novel visual feature synthesis can be transferred into open-vocabulary setting seamlessly; 3) open-vocabulary setting is more realistic and promising given the image-text pairs and arise of large VLMs.

Current experiment settings for OVD and OVS vary from method to method, leading to the performance comparison incomplete, and direct comparison would be unfair to certain methods. To mitigate this issue, we additionally provide a comprehensive benchmark along with the vital components of each method. Following our taxonomy, the remainder of the paper is organized as follows: Section 2 describes the formal definition of OVD and OVS, related domains and tasks, canonical closed-set detectors and segmentors, large VLMs, as well as common datasets and evaluation protocols. Then we review ZSD and ZSS in Section 3 and Section 4, OVD and OVS in Section 5 and Section 6, respectively. Open-vocabulary 3D scene understanding and video instance segmentation are also covered in Section 7. The final Section 8 draws the conclusion, challenges, and promising future directions.

2 Preliminaries

2.1 Problem Definition

The goal of open-vocabulary detection and segmentation is to detect and segment unseen or novel classes that occupy a particular semantically-coherent region or volume within an image, video, or a set of point clouds. The constraint, during its early stages of development, *i.e.*, inductive zero-shot detection and segmentation, is that training images do not contain any unseen objects even if they are unannotated. Later open-vocabulary (transductive zero-shot) detection and segmentation removes such harsh restriction. Nevertheless, both settings avoid novel objects with annotations appearing in the training set. To achieve this requirement, the task 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 is $\mathcal{C}_{train} = \mathcal{C}_B$, and $\mathcal{C}_{train} \cap \mathcal{C}_N = \varnothing$, while $\mathcal{C}_N$ is mainly for testing. With this definition, the difference with closed-set detection and segmentation task is clear, where $\mathcal{C}_{test} = \mathcal{C}_{train} = \mathcal{C}$.

2.2 Related Domains and Tasks

We briefly describe highly-related domains with OVD and OVS, and summarize their differences as follows:

1) Visual Grounding. Visual grounding [83], [84], [85], [86], [87] grounds semantic concepts to visual regions. It can be divided into 1) phrase localization [86] that grounds all nouns in the sentence; 2) referring expression comprehension and segmentation [83], [84], [85] that only grounds the referent in the sentence, the referent is labeled not with a class name but freeform natural language describing instance attributes, positions, and relationships with other objects or background. Visual grounding also accepts arbitrary language queries, it greatly expands the vocabulary

but still is closed-set at inference. 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 hinders the development of visual grounding, besides, the referring expressions only refer to salient objects, barely considering background or the complex scene.

2) Unknown Detection and Segmentation. We regard open-set detection [88], [89], open-set segmentation [90], [91], out-of-distribution detection [92], [93], and anomaly segmentation [94] tasks as one unified unknown detection and segmentation task, from the perspective that all four tasks identify novel objects as one single "unknown" class. Besides, open-world detection [95], [96] and segmentation [97] take a step further to incrementally learn novel categories with a human-in-the-loop strategy, *i.e.*, the system forwards unknowns to an oracle (human annotator) for labeling then adds them back to known classes. These tasks only need to differentiate between known and unknown classes without recognizing different unknown classes, which is simpler than OVD/OVS.

2.3 Canonical Closed-Set Detectors and Segmentors

Faster R-CNN [6] (FRCNN) 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 [98] detection head predicts per-class probability and refines the locations of positive proposals. Meanwhile, one-stage detectors directly refine the positions of anchors without the proposal stage. FCOS [99] 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. With the development of Transformers in NLP, Transformer-based detectors have dominated the literature recently. DETR [21] 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 [22] (Def-DETR) designs a multi-scale deformable attention mechanism that sparsely attends sampled points around queries to accelerate convergence.

For segmentors, DeepLab [14], [100] enhances FCN [8] with dilated convolution, conditional random field (CRF), and atrous spatial pyramid pooling. Mask R-CNN [9] (MR-CNN) adds a parallel mask branch to FRCNN and proposes RoI Align for instance segmentation. Following DETR, MaskFormer [23] (MF) obtains mask embeddings from object queries, which then perform 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 [25] (M2F) follows the same meta-architecture of MF [23] but introduces a masked cross-attention module that only attends to the predicted mask regions. It achieves SoTA on semantic/instance/panoptic segmentation tasks.

2.4 Large Vision-Language Models (VLMs)

Large VLMs [44], [101], [102] have demonstrated superior transfer capability on classification tasks without finetuning in recent years. Particularly, CLIP [44] makes the first breakthrough via contrastively pretraining on an unprecedented 400M image-text pairs crawled from the internet. The pretraining simply predicts which image goes with which text in a batch, which makes an efficient and scalable way to learn transferable representations. During inference, template prompts filled with class names such as "a photo of a [CLASS]" are fed into the CLIP text encoder, the [EOS] token at the last transformer layer is taken as text embedding, the image embedding is obtained via a multi-head self-attention pooling layer at the top of image encoder, then both text and image embeddings are l_2 normalized to compute pairwise cosine similarity. One can fill the template prompts with as many classes as there may be and simply choose the one with the highest similarity score as the prediction. Later ALIGN [101] leverages one billion noisy image alt-text pairs for pretraining, both architecture and objective are the same as CLIP. Another type of large VLMs, *i.e.*, text-to-image diffusion models [103], [104] trained on internet-scale data such as LAION-5B [105] have also attracted a lot of attention. The step-by-step denoising process gradually evolves pure noise tensors into realistic images conditioned on languages, suggesting the internal feature representations are correlated with high-level semantic concepts which could be exploited for segmentation tasks.

2.5 Evaluation Protocols and Datasets

There are three prevalent evaluation protocols for OVD and OVS, namely, 1) open-vocabulary evaluation (OVE), this protocol only evaluates performance on novel classes, *i.e.*, $\mathcal{C}_{test} = \mathcal{C}_N$; 2) generalized open-vocabulary evaluation (gOVE), which tests the model on both base and novel classes, *i.e.*, $\mathcal{C}_{test} = \mathcal{C}_B \cup \mathcal{C}_N$, it is more challenging compared to OVE as the model tends to predict overly confident scores on base classes; 3) cross-dataset transfer evaluation (CDTE), contrary to the above two protocols, the model is trained on one dataset and tested on other datasets without finetuning, for example, one can train on LVIS v1.0 [28] and test on COCO [27] dataset for the former has a larger vocabulary.

Common datasets and evaluation metrics for OVD and OVS are given in Table 1. For open-vocabulary detection and instance segmentation, the metric is box and mask mean average precision (mAP) evaluated at intersection-over-union (IoU) threshold 0.5, respectively. For open-vocabulary semantic segmentation, the metric is mean intersection-over-union (mIoU), the harmonic mean (HM) of mIoU_B and mIoU_N is hIoU [50]. Panoptic quality (PQ), segmentation quality (SQ), and recognition quality (RQ) are used for open-vocabulary panoptic segmentation.

3 Zero-Shot Detection

The harsh constraint of removing training images containing unannotated unseen objects induces the main challenge for ZSD models. Hence, they resort to semantic embeddings [34], [35], [36] unsupervisedly trained on text corpus alone to transfer knowledge from seen to unseen classes. Methods in this section can be discriminative or generative: 1) Discriminative models seek to maximize separation between the decision boundaries of ambiguous classes espe-

TABLE 1: Common datasets and evaluation metrics used for open-vocabulary detection and image segmentation.

Tasks	Datasets (Split of Base/Novel Categories)	Evaluation Metrics
Open-Vocabulary Detection	Pascal VOC [26] (16/4) ILSVRC-2017 Detection [106] (177/23) COCO [27] (48/17, 65/15) LVIS v1.0 [28] (866/337) Visual Genome [107] (478/130) Objects365 [108] OpenImages [109]	Recall@100, AP, AP_{50}, AP_B, AP_N, AP_r
Open-Vocabulary Semantic Segmentation	Pascal VOC [26] (15/5) COCO-stuff [110] (156/15) ADE20K-150 [111] (135/15) ADE20K-847 [111] (572/275) Pascal Context-59 [112] Pascal Context-459 [112]	mIoU, hIoU
Open-Vocabulary Instance Segmentation	COCO [27] (48/17) ADE20k [111] (135/15) OpenImages [109] (200/100)	AP, AP_B, AP_N
Open-Vocabulary Panoptic Segmentation	COCO [27] (119/14) ADE20k [111]	PQ, SQ, RQ

cially the background and unseen classes in visual, semantic, or common embedding space. 2) Generative models use 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 can well reflect the inter-class relationships. Besides the two modifications made to canonical detectors in Fig. 2, a linear layer (mapping function) is additionally added to the backbone to make the dimension of visual features the same as semantic embeddings.

The ZSD task is first proposed by Bansal *et al.* [29]. The RoI features are linearly projected into the semantic space driven by a max-margin loss, then classified by GloVe [36]. Bansal *et al.* propose two techniques to remedy the ambiguity between background and unseen concepts, one is SB which adopts a fixed label vector ([1,0,...,0]) with norm one for modeling background, which is hard to cope with the high background visual variances, the other is LAB that dynamically assigns multiple classes from WordNet [113] belonging to neither seen nor unseen classes to background using an EM-like algorithm. A contemporaneous work SAN [30], [114] 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.* [115] provide external relationship knowledge graph as pairwise potentials besides unary potentials in CRF to achieve context-aware zero-shot detection. ZSDTD [116] leverages textual descriptions instead of a single word vector 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 [45] (described in the next paragraph), BLC [117] 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.* [118] explore transductive generalized ZSD via fixed and dynamic pseudo-labeling strategies to promote training in unseen samples. SSB [39] recently establishes a simple but strong baseline. It carefully ablates model characteristics, learning dynamics, and inference procedures from a myriad of design options.

Besides FRCNN [6], applying one-stage detectors such as YOLO [13], [119] or RetinaNet [120] to ZSD is also explored. Another concurrent work with LAB [29] and SAN [30] is ZS-YOLO [31]. It conditions the objectness branch of YOLO on the combination of semantic attributes, visual features, and localization output instead of visual features alone to improve the low recall rate of novel objects. HRE [121] constructs two parallel visual-to-semantic mapping branches for classification, one is a convex combination of class embeddings, while the other maps grid features associated with positive anchors into the semantic space, and the final prediction is a summation of the two. Later Rahman *et al.* [45] design a polarity loss (PL) that explicitly maximizes the margin between predictions of target and negative classes based on focal loss [120]. A vocabulary metric learning approach is also proposed to provide a richer and more complete semantic space for learning the mapping. Li *et al.* [122] perform the prediction of superclasses and fine-grained classes in parallel, similar to the hybrid branches in HRE [121].

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. Gupta *et al.* [46] demonstrate that classes can have poor separation in semantic space but are well separated in visual space, and vice versa. They propose MS-Zero [46] which exploits this complementary information via two unidirectional mapping functions. Similarity metrics are calculated in both spaces which are then averaged as the final prediction for better discrimination. Similar to previous works [121], [122], DPIF [123] proposes a dual-path inference fusion module that 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 [124] proposes RRCL contrasting seen region features to make the visual space more discriminative, and RCCL contrasts seen region features with both seen and unseen class embeddings under the guidance of the semantic relationship matrix. It maps seen and unseen classes separately into the joint embedding space.

3.1.3 Learning a Mapping from Semantic to Visual Space

Zhang *et al.* [125] 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 [126], *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, and perform k nearest neighbor search to find the most suitable unseen category from candidate textual descriptions.

3.2 Novel Visual Feature Synthesis

Novel visual feature synthesis produces "fake" unseen visual features as training samples for a new classifier to

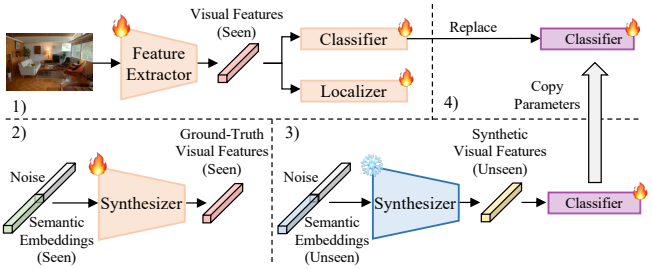

Fig. 3: Flowchart of novel visual feature synthesis.

enable recognition of novel concepts. This methodology follows a multi-stage pipeline shown in Fig. 3: 1) train the base model only on annotations of seen classes in a fully-supervised manner; 2) train the feature synthesizer $G : \mathcal{W} \times \mathcal{Z} \mapsto \tilde{\mathcal{F}}$ on seen class 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, upon which a new classifier is retrained 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.

Zhu *et al.* [47] mainly attribute the missed detection of unseen objects to the low confidence scores assigned by the objectness branch in YOLOv2 [119]. They leverage conditional variational auto-encoder [81] 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, DELO [47] retrains the objectness branch to assign high confidence scores on both seen and unseen objects. Later on, Hayat *et al.* [127] identify the same issue and use the same class consistency loss to make the generated features more discriminative. In addition, they adopt the mode seeking regularization [128] which maximizes the distances of generated data points *w.r.t* their noise vectors. At the same time with DELO [47], GTNet [48] proposes an IoU-Aware synthesizer based on the Wasserstein generative adversarial network [129]. The RoIs refined by RPN may not entirely overlap with ground-truths, however, the synthesizer used in DELO [47] can not generate unseen RoI features with diverse spatial context clues because it is only trained on ground-truths which perfectly align with the boundary of objects. To mitigate the context gap between unseen RoI features from RPN and those synthesized by the generator, Zhao *et al.* [48] randomly sample foreground and background RoIs according to max-IoU label assignment as the additional generation target, thus making the new classifier robust to various degrees of context information. Huang *et al.* [49] propose RRFS that consists of an intra-class semantic diverging loss and an inter-class structure preserving loss. The former pulls positive synthesized features lying within the hyper-sphere 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.

TABLE 2: ZSD performance on Pascal VOC [26] under gOVE protocol. The metric is mAP at IoU threshold 0.5. R and FPN denotes ResNet [11] and feature pyramid network [130]. DN is DarkNet [13], [119].

Method	Image Backbone	Semantic Embeddings	OVE AP	gOVE $AP_B/AP_N/AP$
SAN [30]	R50	-	59.1	48.0/37.0/41.8
HRE [121]	DN19	aPY [131]	54.2	62.4/25.5/36.2
PL [45]	R50-FPN	aPY [131]	62.1	-
BLC [117]	R50	-	55.2	58.2/22.9/32.9
TL [118]	R50-FPN	W2V	66.6	-
MS-Zero [46]	R101	aPY [131]	62.2	-/-/60.1
CG-ZSD [122]	DN53	BERT	54.8	-
SU [127]	R101	FT	64.9	-
DPIF [123]	R50	aPY [131]	-	73.2/62.3/67.3
ContrastZSD [124]	R101	aPY [131]	65.7	63.2/46.5/53.6
RRFS [49]	R101	FT	65.5	47.1/49.1/48.1

TABLE 3: ZSD performance on COCO [27] dataset. IRv2 denotes InceptionResnetv2 [132].

Method	Image Backbone	Semantic Embeddings	OVE AP	gOVE $AP_B/AP_N/AP$
48/17 split [29]				
SAN [114]	R50	W2V	5.1	13.9/2.6/4.3
SB [29]	IRv2	-	0.7	-
LAB [29]	IRv2	-	0.3	-
DSES [29]	IRv2	-	0.5	-
MS-Zero [46]	R101	GloVe	12.9	-/-/30.7
PL [45]	R50-FPN	W2V	10.0	35.9/4.1/7.4
CG-ZSD [122]	DN53	BERT	7.2	-
BLC [117]	RN50	W2V	10.6	42.1/4.5/8.2
ContrastZSD [124]	R101	W2V	12.5	45.1/6.3/11.1
SSB [39]	R101	W2V	14.8	48.9/10.2/16.9
DELO [47]	DN19	W2VR [31]	7.6	-/-/13.0
RRFS [49]	R101	FT	13.4	42.3/13.4/20.4
65/15 split [45]				
PL [45]	R50-FPN	W2V	12.4	34.1/12.4/18.2
TL [118]	R50-FPN	W2V	14.6	28.8/14.1/18.9
CG-ZSD [122]	DN53	BERT	10.9	-
BLC [117]	R50	W2V	14.7	36.0/13.1/19.2
DPIF-M [123]	R50	W2V	19.8	29.8/19.5/23.6
ContrastZSD [124]	R101	W2V	18.6	40.2/16.5/23.4
SSB [39]	R101	W2V	19.6	40.2/19.3/26.1
SU [127]	R101	FT	19.0	36.9/19.0/25.1
RRFS [49]	R101	FT	19.8	37.4/19.8/26.0

3.3 Discussion

Although the visual-semantic space mapping methods are the earliest attempts toward "open-vocabulary", they own the following intrinsic drawbacks: 1) the bias issue, *i.e.*, the model tends to misclassify unseen categories to seen ones (overfitting to seen classes) due to the lack of unseen training samples; 2) the confusion between unseen and background classes, and the difficulty to model background; 3) the hubness problem [126] that only predictions of a few unseen classes are highly confident in most cases, induced by the shrinking dimension of semantic space or the joint space with limited capacity to encompass visual variations. Research endeavors are gradually shifted to the novel visual feature synthesis methodology. However, it has its own limitations, *i.e.*, the synthetic visual features are unrealistic and have limited variations to reflect the fine-grained complexity of visual objects. Nevertheless, the appearance gap between real and fake objects we believe will be ultimately mitigated given the impressive progress on diffusion models [103], [104]. A quantitative comparison of ZSD models is given in Tables 2 to 4.

TABLE 4: ZSD performance on ILSVRC-2017 detection [106] and Visual Genome [107] dataset. R@100 is Recall@100 at IoU threshold 0.5.

Method	Image Backbone	Semantic Embeddings	OVE R@100	OVE AP
177/23 split [30] for ILSVRC-2017 Detection				
SAN [30]	R50	W2V	-	16.4
ZSDTD [116]	IRv2	Text-Desc	-	24.1
GTNet [48]	R101	FT	-	26.0
SU [127]	R101	FT	-	24.3
478/130 split [30] for Visual Genome				
SB [29]	IRv2	-	4.1	-
LAB [29]	IRv2	-	5.4	-
DESE [29]	IRv2	-	4.8	-
CA-ZSD [115]	R50	GloVe	-	-
ZSDTD [116]	IRv2	Text-Desc	7.2	-
GTNet [48]	R101	W2V	11.3	-
S2V [125]	IRv2	GloVe	11.0	-
DPIF-M [123]	R50	W2V	18.3	1.8

4 Zero-Shot Segmentation

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

4.1 Zero-Shot Semantic Segmentation

4.1.1 Visual-Semantic Space Mapping

Learning a Mapping from Visual to Semantic Space. SPNet [50] is the first work that proposes zero-shot semantic segmentation task. It directly maps pixel features into the 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.* [133] address the visual-semantic correspondence from the rarely noticed uncertainty perspective. They argue that the noisy and outlying samples in seen classes have adverse effects on the correspondence establishment. An uncertainty-aware loss is proposed to adaptively strengthen representative samples while attenuating loss for uncertain samples with high variance estimation. JoEm [51] learns a joint embedding space via the proposed boundary-aware regression loss and semantic consistency loss. At the test time, the semantic embeddings are transformed into semantic prototypes acting as a nearest-neighbor classifier without the classifier retraining stage in Section 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.* [134] propose variational mapping from semantic space to visual space via sampling the conditions (mimicking the support images in few-shot semantic segmentation [135]) from the predicted distribution. PMOSR [52] abstracts a set of seen visual prototypes, then trains a projection network mapping seen semantic embeddings to these prototypes. Similar to JoEm [51], new unseen classes can be flexibly added in inference without classifier retraining, since one can simply project unseen semantic embeddings to unseen prototypes for classification. An open-set rejection module is further proposed to prevent unseen classes from directly competing with seen classes.

TABLE 5: Zero-shot semantic segmentation performance on Pascal VOC [26] and Pascal Context [112] datasets. ZS3Net [32] randomly samples 2 to 10 novel classes with step size 2, here we only show the results of 4 novel classes.

Method	Image Backbone	Semantic Embeddings	Pascal VOC mIoU (B/N/HM)	Pascal Context mIoU (B/N/HM)
15/5 split [50] for Pascal VOC 29/4 split [53] for Pascal Context				
SPNet-C [50]	R101	W2V & FT	78.0/15.6/26.1	35.1/4.0/7.2
ZS3Net [32]	R101	W2V	77.3/17.7/28.7	33.0/7.7/12.5
VM [134]	VGG16	GloVe	-/35.6/-	-
CaGNet [53]	R101	W2V & FT	78.4/26.6/39.7	36.1/14.4/20.6
SIGN [137]	R101	W2V & FT	75.4/28.9/41.7	33.7/14.9/20.7
Novel - 4 [32]				
SPNet [50]	R101	W2V & FT	67.3/21.8/32.9	36.3/18.1/24.2
ZS3Net [32]	R101	W2V	66.4/23.2/34.4	37.2/24.9/29.8
CSRL [136]	R101	-	69.8/31.7/43.6	39.8/23.9/29.9
JoEm [51]	R101	W2V	67.0/33.4/44.6	36.9/30.7/33.5
PMOSR [52]	R101	W2V	75.0/44.1/55.5	41.1/43.1/42.1

4.1.2 Novel Visual Feature Synthesis

Concurrent with SPNet [50], Bucher *et al.* [32] propose ZS3Net, they condition the synthesizer [79] on adjacency graph encoding structural object arrangement to capture contextual cues for the generation process. CSRL [136] 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 attributes and produces limited visual diversity, hindering the effectiveness of the generative models. CaGNet [53] 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 [53], Cheng *et al.* [137] also substitute the noise vector, but with a spatial latent code incorporating the relative positional encoding. While previous ZS3Net [32] and CaGNet [53] simply discard pseudo-labels whose confidence scores are below a threshold and weight the importance of the remaining pseudo-labels equally, SIGN [137] utilizes all pseudo annotations but assigns different loss weights according to the confidence scores of pseudo-labels.

4.2 Zero-Shot Instance Segmentation

Zheng *et al.* [33] 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 word vectors. 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 [29], [117], they design a background-aware RPN and a synchronized background strategy to adaptively represent background.

4.3 Discussion

The performance comparison of zero-shot semantic segmentation is given in Table 5. In general, the novel visual

feature synthesis approaches perform better than the visual-semantic space mapping methods because the former avoid the hubness problem without mapping to a smaller dimensional space, and can well alleviate the bias problem and confusion between background and novel concepts in Section 3.3.

5 Open-Vocabulary Detection

OVD removes the stringent restriction on unannotated novel samples in Section 3. From this section on, we discuss methods resorting to weak supervision signals.

5.1 Region-Aware Training

Weakly-Supervised Grounding or Contrastive Loss. Methods in this category leverage image-text pairs to establish a coarse and noisy correspondence between regions and words. During training, they first measure local similarity scores between each word in the sentence (extracted by an off-the-shelf language parser) and each proposal, which is accumulated to form a global image-text level grounding score, then the bidirectional grounding or contrastive loss seeks to maximize the matching scores of positive paired image-texts, while minimizing that of negative pairs (for current paired image and caption, other captions and images from other pairs in the batch are negatives). Zareian *et al.* [40] first formulate the open-vocabulary detection task. Previous ZSD methods typically only train the vision-to-language (V2L) mapping layer from scratch on base classes, which is prone to overfitting. They instead learn the V2L layer during pretraining on a rich and complete visual-semantic space provided by image-caption datasets. The bidirectional grounding loss forces every word and every region within a positive image-caption pair to be close. LocOv [138] basically follows OVR-CNN [40] in terms of the main grounding objective plus the auxiliary image-text matching and masked language modeling (MLM) objectives. One difference is that LocOv utilizes both region and grid features for measuring local similarity scores with words while OVR-CNN only adopts the former one. Coherent with OVR-CNN, LocOv founds that only using the learned dictionary (embeddings before BERT [37]) is better than using BERT to encode contextualized text embeddings (embeddings after BERT). RO-ViT [139] randomly crops and resize regions of positional embeddings (PE) instead of the holistic image positional embeddings to match the use of PE in the detection fine-tuning phase. It utilizes focal loss [120] instead of softmax to put emphasis on hard negatives and a better proposal network OLN [140]. DetCLIP [141] constructs a concept dictionary with knowledge enrichment to achieve parallel concept formulation, avoiding unnecessary category interactions similar to the finding of LocOv [138]. DetCLIPv2 [56] selects a single region that best fits the current word via *argmax* instead of aggregating the scores of all regions to the current word in previous methods [40], [138], [139], [141]. It also excludes the image-to-text matching in the bidirectional loss due to the partial labeling problem put in their paper, *i.e.*, the caption usually describes a small fraction of objects in the image, hence most proposals can not find their matching words in the caption.

Ground-Truth Region-Word Correspondence. This line of work trains on ground-truth region-word pairs in visual grounding datasets to expand the vocabulary. MDETR [54] combines Flickr30K [86], Visual Genome (VG) [107], ReferItGame [85], and RefCOCO/+/g [83], [84] into a dataset containing 1.3M aligned region-text pairs. It proposes soft token prediction to predict the span of tokens in text and region-word-level contrastive loss to enforce alignment in the latent feature space. GLIP [55] in Section 5.2 also uses these golden region-word pairs and contrastive loss of MDETR. MAVL [142] improves MDETR by multi-scale deformable attention [22] and late fusion. MQ-Det [143] augments language queries in GLIP [55] 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 MLM forces the model to align with vision cues to reduce the learning inertia problem. With these techniques, MQ-Det only requires one epoch training upon GLIP. SGDN [144] also leverages Flickr30K and VG [107] datasets, but it exploits additional object relations in a scene graph to facilitate discovering, classifying, and localizing novel objects. Note that SGDN uses RoBERTa [145] instead of the CLIP text embeddings.

Others. MEDet [146] mines region-word relationships online via concept augmentation, noisy pair removal, and fragmented proposal merging steps on image captioning datasets. VLDet [147] formulates the region-word alignment as a set-matching problem that can be automatically learned on image-caption pairs and solved via an off-the-shelf Hungarian algorithm. CORA [148] augments region features with learned prompts to better align with CLIP visual-semantic space. The proposed anchor pre-matching can avoid repetitive per-class inference in [61] to improve efficiency. Currently, CORA sets SoTA performance on OVD.

5.2 Pseudo-Labeling

Models advocating pseudo-labeling also leverage abundant image-text pairs as in Section 5.1, but additionally, they adopt large pretrained VLMs or themselves (via self-training) to generate pseudo labels. Detectors are then trained on the unification of base annotations and new pseudo labels. According to the type and granularity of pseudo labels, methods can be grouped into: pseudo region-caption pairs, region-word pairs, and pseudo captions.

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 methods that develop assignment strategies propagating image-level labels to corresponding proposals, Detic [149] side-steps this error-prone label assignment process building region-word correspondence and simply trains the max-size proposal to predict all image-level labels. The max-size proposal is assumed to be big enough to cover all image-level labels (containing multiple object nouns like captions). Thus the classifier encounters various novel classes during training on ImageNet21K (IN21K) [106] and can generalize to novel objects at inference. 3Ways [150] regards the top-scoring bounding box per image as correspondence to the whole

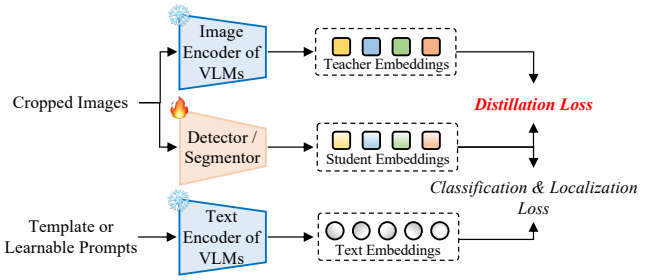

Fig. 4: A general teacher-student framework for knowledge distillation-based approaches. Teacher embeddings can be computed offline and cached to save training costs.

caption, again, one region corresponds to multiple concepts. They also augment text embeddings to avoid overfitting and include trainable gated shortcuts to stabilize training.

Pseudo Region-Word Pairs. In contrast to the soft alignment in pseudo region-caption pairs and the bidirectional grounding or contrastive losses in Section 5.1, which allows one region/word connects to multiple words/regions weighted by *softmax*, pseudo region-word pairs only allows one region/word to correspond to one word/region. RegionCLIP [57] leverages CLIP to create more fine-grained pseudo region-word pairs to pretrain the image encoder. However, proposals with the highest CLIP score yield low localization performance. Targeting this problem, VL-PLM [151] fuses CLIP scores with objectness scores and repeatedly applies the RoI head to remove redundant proposals. GLIP [55] reformulates object detection into phrase grounding and trains the model on the unification 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 which may not effectively localize novel objects. GLIPv2 [152] further reformulates visual question answering and image captioning into a grounded vision-language understanding task. Following GLIP [55], Grounding DINO [153] upgrades the detector into a transformer-based one, enhancing the capacity of the teacher model. Following Detic [149], Rasheed *et al.* [154] select as many pseudo-boxes from proposals as image-level labels using MViT [155] instead of a single max-size proposal corresponding to all image labels. Instead of generating pseudo labels once, PromptDet [156] iteratively learns region prompts and sources uncurated web images in two rounds, leading to more accurate pseudo boxes even though the heuristic rule [57], [150], [154] (proposals with highest scores are regarded as ground-truths) is also adopted.

Interpretability-based Pseudo Region-Word Pairs. Instead of thresholding [55], [152], [153] or using the top-scoring heuristic rule [57], [150], [154], [156], PB-OVD [58] generates pseudo annotations via employing GradCAM [157] to compute the activation map in the cross-attention layer of VLM (ALBEF [158]) *w.r.t.* an object of interest in the caption, then the proposals that overlap the most with the activation map are regarded as pseudo ground-truths.

Pseudo Captions. Cho *et al.* [59] propose to generate another type of pseudo label, *i.e.*, pseudo captions describing objects in natural languages instead of bounding boxes. They leverage an image captioning model to generate captions for each object, which are then fed into the text encoder of CLIP, encoding the class attributes and relationships with the surrounding environment. It can also be seen as better prompting compared to the template prompts in CLIP.

5.3 Knowledge Distillation-Based

Knowledge distillation-based (KD) methodology (*c.f.* Fig. 4), employs a teacher-student framework. It is divided into two subcategories, *i.e.*, distilling region embeddings individually or collectively, where the former feeds individual RoI, and the latter forwards a bag of RoIs into VLMs image encoder. The distillation loss is typically a $\mathcal{L}_1$ loss.

5.3.1 *Distilling Region Embeddings Individually*

Distilling Knowledge into Two-Stage Detectors. The foundational work ViLD proposed by Gu *et al.* [38] is the first to distill knowledge from CLIP into Faster R-CNN. Gu *et al.* verified that region proposals from RPN can well generalize to novel objects, though the RPN is only trained on base categories. ViLD consists of ViLD-text and ViLD-image. The former substitutes the classifier weights with text embeddings obtained via feeding prompt templates filled with base class names into the text encoder of CLIP. On the contrary, ViLD-image aligns region embeddings of both base and novel categories with teacher embeddings and discards the teacher model during inference. To alleviate the bias problem, Gu *et al.* further take a weighted geometric mean between the predictions of teacher and student as an ensemble method, which is inherited in many subsequent works. The follow-up work DetPro [60] bypasses the laborious prompt engineering in ViLD. ViLD maintains a curated list of manual prompt templates, *i.e.*, "A photo of the [CLASS]." As pointed out in CoOp [159], identifying proper prompts is crucial to downstream tasks. Hence, Du *et al.* [60] learn prompt representations automatically following CoOp. Particularly, DetPro forces all negative proposals equally unlike any object class to interpret the background concept. In addition, DetPro divides foreground proposals into disjoint groups according to their IoUs with ground-truths and learns prompt representations for each group separately to describe different levels of contexts accurately. Rasheed *et al.* [154] distill region embeddings from CLIP with an inter-embedding relationship matching (IRM) loss besides the usual $\mathcal{L}_1$ loss. The IRM loss (a Frobenius norm, $||\cdot||_F$) forces student embeddings to share the same inter-embedding similarity structure as teacher embeddings. Another work EZSD [160] contradicts ViLD in that predefined anchor boxes with high objectness scores (more chance of covering novel objects) are deemed as distillation regions instead of proposals from RPN. The distillation loss is the same as ViLD ($\mathcal{L}_1$ loss) but is weighted by the objectness. EZSD finetunes the layer normalization layers to adapt CLIP features to downstream detection data and boosts distillation efficiency. A semantic-based regressor is also proposed to improve the regression performance, while previous works generally adopt a class-agnostic localization module. OADP [161] improves the efficiency of knowledge transfer by a pyramid architecture including three granularities, *i.e.*, object-, image-, and block-level, each with a $\mathcal{L}_1$ loss.

Distilling Knowledge into One-Stage Detectors. Instead of region embeddings, ZSD-YOLO [162] aligns the embeddings of positive anchor points with their corresponding cropped ground-truth embeddings generated by CLIP via the $\mathcal{L}_1$ loss. HierKD [163] follows ZSD-YOLO [162] with several modifications: 1) a more advanced one-stage detector ATSS [164] is adopted instead of YOLOv5 [165]; 2) besides distilling the embeddings of positive grid features, a global KD further distills knowledge in a language-to-visual manner on image-captions pairs using a symmetrical contrastive loss. GridCLIP [166] aligns global image-level embeddings between two identical image encoders of CLIP (one is trainable and the other is frozen) also with $\mathcal{L}_1$ loss. Zang *et al.* [61] propose OV-DETR based on DETR [21]. OV-DETR conditions object queries on concept embeddings (embeddings of ground-truth class names or bounding boxes forwarded to CLIP) and reformulates the set matching objective into a conditional binary matching, which measures the matchability between detection outputs and the conditional object queries. However, the conditioned object queries are class-specific, the number of which is linearly proportional to the number of object classes. Prompt-OVD [167] addresses the slow inference speed by prepending class prompts instead of repeatedly adding to object queries and changing the binary matching goal to a multi-label classification cost. It further proposes RoI-based Masked Attention and RoI Pruning to extract region embeddings from the teacher in one forward pass instead of sending regions into the CLIP image encoder one by one.

5.3.2 Distilling Region Embeddings Collectively

Pretraining images of CLIP contain multiple concepts and their compositional structure, *e.g.*, co-occurrence of objects, is implicitly captured in the image encoder. However, in previous work [38], [60], [161], individual RoI with limited spatial clues does not provide such prior. Inspired by MaskCLIP [73], BARON [62] aligns the embedding of bag-of-regions to harness this knowledge embedded in the teacher. For each proposal, it samples nearby regions to form multiple groups of bag-of-regions, which are then encoded as student embeddings by CLIP text encoder. The image crop enclosing the bag-of-regions (hence the knowledge is preserved) for each group is encoded as teacher embedding by CLIP image encoder. Finally the teacher and student embeddings are aligned via InfoNCE loss [168].

5.4 Transfer Learning-Based

Transfer learning-based models differ from KD-based methodology in the usage of VLMs. Specifically, it mainly leverages the VLMs image encoder as a feature extractor. For example, directly fine-tuning it on detection data, or extracting visual features via the frozen image encoder of VLMs. OWL-ViT [178] removes the final token pooling layer of the image encoder and attaches a lightweight detection head to each transformer output token. Then it fine-tunes the whole model end-to-end on detection data through a bunch of dedicated techniques. UniDetector [179] also trains the whole model, it initializes its image backbone using RegionCLIP [57] pretrained weights. During training, UniDetector leverages images of multiple sources and heterogeneous label spaces. During inference, it calibrates the base and novel probabilities via a class-specific prior probability recording how the network biases towards that category. Instead of fine-tuning the whole model, F-VLM [63] leverages the frozen CLIP image encoder as the image backbone to extract features and only trains the detection head. It ensembles predictions of the detector and CLIP via geometric mean (dual-path inference). Another line of work only transfers the CLIP text encoder for detection and discards the CLIP image encoder. ScaleDet [180] unifies the multi-dataset label space under hard and soft assignment, which disambiguates classes and relates similar classes, respectively. OpenSeeD [181] unifies open-vocabulary detection, and semantic/instance/panoptic segmentation in one network architecture. It proposes decoupled foreground-background decoding and conditioned mask decoding to compensate for task and data discrepancies, respectively. Similar to DetCLIP [141] that enriches the concept name with definitions, Kaul *et al.* [182] employ a large language model (LLM), *i.e.*, GPT-3 [183], to generate rich descriptions which are then fed into CLIP text encoder for classification.

5.5 Discussion

The OVD performance is given in Tables 6 to 8. Though the region-aware training and pseudo-labeling methodologies leverage relatively cheap and abundant image-text pairs, how to counteract the negative effect of noisy and incorrect region-word pairs which is crucial to improving data efficiency is still untouched. Integrating the pseudo-labeling process into model training like semi-supervised object detection [184], [185] instead of generating them once and done could gradually improve the quality of pseudo region-word pairs. For knowledge distillation-based and transfer learning-based models, the context discrepancy hinders unleashing the full potential of VLMs. Pretraining images of CLIP are full scenes with resolution 224×224, while proposals (image crops) contain limited spatial clues with non-square sizes or extreme aspect ratios, the preprocessing step of CLIP (resize the shorter edge to 224 and center crop) further distorts proposals and aggravates this gap. Besides, fine-tuning CLIP leads to degraded performance on novel objects, and ensembling CLIP during inference is unacceptable for edge devices.

6 Open-Vocabulary Segmentation

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

6.1 Open-Vocabulary Semantic Segmentation

6.1.1 Region-Aware Training

Weakly-Supervised Grounding or Contrastive Loss. Similar to Section 5.1, models under this category mainly bidirectionally ground regions and words on image-text pairs. Following OVR-CNN [40], OpenSeg [42] leverages the weakly supervised grounding loss with a random drop on each word to encourage each word/region to be aligned to one or a few regions/words. SimSeg [205] identifies that CLIP heavily relies on contextual pixels and contextual words instead of entity words during the pretraining phase. Hence, instead of aligning all image patches with all words (dense

TABLE 6: OVD performance on COCO [27] dataset. The number of base/novel classes is 48/17. "T (cat)" denotes template prompts filled with category names, while "L" denotes learnable prompts, and desc is category descriptions obtained from definitions of WordNet [113], Wikipedia, or the dataset itself. Prompts with ✗ does not utilize CLIP text encoder, instead they use BERT [37]. Ensemble suggests whether the final prediction is ensembled with CLIP.

Method	Image Backbone	Detector	Image-Text pairs	Teacher	Prompts	Ensemble	OVE	gOVE		
							AP_N	AP_N	AP_B	AP
Region-Aware Training										
OVR-CNN [40]	R50-C4	FRCNN	COCO Cap [169]	✗	✗	✗	27.5	22.8	46.0	39.9
LocOv [138]	R50-C4	FRCNN	COCO Cap [169]	✗	✗	✗	30.1	28.6	51.3	45.7
MEDet [146]	R50-C4	FRCNN	COCO Cap [169] CC [170]	✗	T (cat)	✗	-	32.2	53.3	47.8
VLDet [147]	R50-C4	FRCNN	COCO Cap [169]	✗	T (cat)	✗	-	32.0	50.6	45.8
RO-ViT [139]	ViT-B/16	MRCNN	ALIGN [101]	✗	T (cat)	✓	-	30.2	-	41.5
CORA [148]	R50	DAB-DETR [171]	CLIP [44]	✗	T (cat)	✗	-	35.1	35.5	35.4
SGDN [144]	R50	Def-DETR [22]	Flickr30K [86], VG [107]	✗	✗	✗	-	37.5	61.0	54.9
Pseudo-Labeling										
RegionCLIP [57]	R50-C4 ($CLIP_V$)	FRCNN	CC3M [170]	✗	T (cat)	✗	35.2	31.4	57.1	50.4
Detic [149]	R50-C4	FRCNN	COCO Cap [169]	✗	T (cat)	✗	-	27.8	47.1	45.0
PromptDet [156]	R50-FPN	MRCNN	LAION-novel [172]	✗	L (cat+desc)	✗	-	26.6	-	50.6
PB-OVD [58]	R50	MRCNN	COCO Cap [169], VG [107], SBU [173]	✗	T (cat)	✗	-	30.8	46.1	42.1
CondHead [174]	Same as RegionCLIP [57]			✗	T (cat)	✗	-	33.7	58.0	51.7
XPM [41]	R50-C4	FRCNN	CC [170]	✗	✗	✗	29.9	27.0	46.3	41.2
Knowledge Distillation-based										
ViLD [38]	R50-FPN	MRCNN	✗	CLIP (ViT-B/32)	T (cat)	✓	-	27.6	59.5	51.3
ZSD-YOLO [162]	CSP-DN53 [175]	YOLOv5x [165]	✗	CLIP (ViT-B/32)	T (cat+desc)	✗	13.4	13.6	31.7	19.0
HierKD [163]	R50-FPN	ATSS [164]	CC [170]	CLIP (ViT-B/32)	T (cat/desc)	✗	25.3	20.3	51.3	43.2
OV-DETR [61]	R50-C4	Def-DETR [22]	✗	CLIP (ViT-B/32)	T (cat)	✗	-	29.4	61.0	52.7
RKDWTF [154]	R50-C4	FRCNN	COCO Cap [169]	CLIP (ViT-B/32)	T (cat)	✗	-	36.6	54.0	49.4
OADP [161]	R50-FPN (SoCo [176])	FRCNN	✗	CLIP (ViT-B/32)	T (cat)	✓	-	30.0	53.3	47.2
BARON [62]	R50-FPN (SoCo [176])	FRCNN	✗	CLIP (ViT-B/32)	T (cat)	✗	-	34.0	60.4	53.5
Prompt-OVD [167]	ViT-B/16 (ViTDet [177])	Def-DETR [22]	✗	CLIP (ViT-L/14)	T (cat)	✓	-	30.6	63.5	54.9
Transfer Learning-based										
F-VLM [63]	R50-FPN	MRCNN	✗	✗	T (cat)	✓	-	28.0	-	39.6

sampling), SimSeg sparsely samples a portion of patches and words used for the bidirectional contrastive losses. PACL [209] finds that the CLIP image encoder produces similar patch representations for semantically coherent regions [215]. Thus, PACL aligns the weighted sum of patch embeddings with text embeddings instead of the sole [CLS] image token to encourage patch-level alignment. TCL [211] introduces a region-level text grounder to produce text-grounded masks, then performs matching on grounded image regions and texts via symmetric InfoNCE loss [168].

Grouping in ViT without Densely Annotated Masks. This line of work contrasts image-text pairs, *i.e.*, contrasting only one image and text token instead of aligning all words and regions within an image-text pair. But the model learns to group semantically-coherent pixels into segments in plain ViT [19] by appending a set of learnable segment tokens to aggregate patch tokens without mask annotations. The image token used to contrast is obtained from pooling these segment tokens. GroupViT [64] is the first work that hierarchically shrinks the number of learnable group tokens in different stages of ViT and groups patch tokens into arbitrary-shaped segments according to their assigned groups via gumbel-softmax [216]. It extracts object nouns from captions and prompts them to increase positive text samples for image-text contrasting. ViL-Seg [203] trains the image encoder jointly with a vision-based contrasting and a cross-modal contrasting, then groups and classifies segments through an online clustering head trained by mutual information maximization. Following the same group-based principle [64], [203], SegCLIP [204] adds a reconstruction loss [217] and a superpixel-based KL loss to the normal image-text contrastive loss. OVSegmentor [207] proposes a slot attention-based binding module to group patch tokens then aligns averaged group tokens (the image embedding) with the text embedding via an image-text contrast and the proposed cross-image mask consistency loss.

6.1.2 Pseudo-Labeling

Zabari *et al.* [65] leverage a transformer interpretability method [218] to generate coarse relevance maps for each category, which are then refined by test-time-augmentation techniques (identity, 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-Based

GKC [66] proposes a text diversification strategy that enriches the template prompts with synonyms from WordNet [113] instead of relying only a single category name to guess what the object looks like. The text-guided knowledge distillation loss transfers the inter-class distance relationships in semantic space into visual space with the same $\mathcal{L}_1$ loss as in Section 5.3.

TABLE 7: OVD performance on LVIS v1.0 [28] dataset under the gOVE protocol. Base classes are common and frequent, while rare classes are novel. Gray denotes mask AP. Backbones suffixed with ($CLIP_V$) are initialized from CLIP.

Method	Image Backbone	Detector	Image-Text pairs	Teacher	Prompts	gOVE AP_r	gOVE AP_c	gOVE AP_f	gOVE AP
Region-Aware Training									
MEDet [146]	R50-FPN	CN2 [186]	COCO Cap [169] CC [170]	✗	T (cat)	22.4	-	-	34.4
VLDet [147]	R50-FPN	CN2 [186]	CC3M [170]	✗	T (cat)	21.7	29.8	34.3	30.1
RO-ViT [139]	ViT-B/16	MRCNN	ALIGN [101]	✗	T (cat)	28.0	-	-	30.2
SGDN [144]	R50	Def-DETR [22]	VG [107], Flickr30K [86]	✗	✗	23.6	29.0	34.3	31.1
Pseudo-Labeling									
RegionCLIP [57]	R50-C4 ($CLIP_V$)	MRCNN	CC3M [170]	✗	T (cat)	$17.1_{17.4}$	$27.4_{26.0}$	$34.0_{31.6}$	$28.2_{26.7}$
RegionCLIP [57]	R50x4-C4 ($CLIP_V$)	MRCNN	CC3M [170]	✗	T (cat)	$22.0_{21.8}$	$32.1_{30.2}$	$36.9_{35.1}$	$32.3_{30.7}$
Detic [149]	R50-FPN	MRCNN	IN21K [106]	✗	T (cat)	17.8	26.3	31.6	26.8
PromptDet [156]	R50-FPN [130]	MRCNN	LAION-novel [172]	✗	L (cat+desc)	21.4	23.3	29.3	25.3
3Ways [150]	NF-F0 [187]-FPN	FCOS [99] (T-Head [188])	CC12M [189]	✗	T (cat) + dropout	25.6	34.2	41.8	35.7
CondHead [174]	R50-C4 ($CLIP_V$)	Same as RegionCLIP [57]		✗	T (cat)	$19.9_{20.0}$	$28.6_{27.3}$	$35.2_{32.2}$	$29.7_{27.9}$
PCL [59]	Swin-L [20]	Def-DETR [22]	VG [107]	✗	GPT-2 [190]	29.1	-	-	32.9
Knowledge Distillation-based									
ViLD-ens [38]	R50-FPN	MRCNN	✗	CLIP (ViT-B/32)	T (cat)	$16.7_{16.6}$	$26.5_{24.6}$	$34.2_{30.3}$	$27.8_{25.5}$
ViLD-ens [38]	EN-b7 [191]	MRCNN	✗	ALIGN [101] (EN-l2 [191])	T (cat)	$27.0_{26.3}$	$29.4_{27.2}$	$36.5_{32.9}$	$31.8_{29.3}$
DetPro [60]	R50-FPN (SoCo [176])	MRCNN	✗	CLIP (ViT-B/32)	L (cat)	$20.8_{19.8}$	$27.8_{25.6}$	$32.4_{28.9}$	$28.4_{25.9}$
OV-DETR [61]	R50-C4	Def-DETR [22]	✗	CLIP (ViT-B/32)	T (cat)	$18.0_{17.4}$	25.0	32.5	$27.4_{26.6}$
RKDWTF [154]	R50-FPN	CN2 [186]	IN21K [106]	CLIP (ViT-B/32)	T (cat)	25.2	33.4	35.8	32.9
GridCLIP [166]	$CLIP_V$ (R50)	FCOS [99]	✗	CLIP (ViT-B/32)	T (cat)	15.0	22.7	32.5	25.2
OADP [161]	R50 (SoCo [176])	FRCNN	✗	CLIP (ViT-B/32)	T (cat)	$21.9_{21.7}$	$28.4_{26.3}$	$32.0_{29.0}$	$28.7_{26.6}$
EZSD [160]	R50-FPN	MRCNN	✗	CLIP (ViT-B/32)	T (cat)	15.8	25.6	31.7	26.3
BARON [62]	R50-FPN (SoCo [176])	FRCNN	✗	CLIP (ViT-B/32)	L (cat)	$23.2_{22.6}$	$29.3_{27.6}$	$32.5_{29.8}$	$29.5_{27.6}$
Prompt-OVD [167]	ViT-B/16 (ViTDet [177])	Def-DETR [22]	✗	CLIP (ViT-L/14)	T (cat)	$29.4_{23.1}$	-	-	$33.0_{24.2}$
Transfer Learning-based									
OWL-ViT [178]	ViT-H/14	DETR	LiT [192]	✗	T (cat)	23.3	-	-	35.3
F-VLM [63]	R50-FPN	MRCNN	✗	✗	T (cat)	18.6	-	-	24.2
MMC (Text) [182]	R50-FPN	CN2 [186]	✗	✗	GPT-3 [183]	19.3	-	-	30.3

TABLE 8: OVD performance under the CDTE protocol in Section 2.5 on the test set of Pascal VOC [26] and validation set of Obejcts365 [108] (O365), COCO [27], OpenImages [109] (OI), and LVIS v1.0 [28]. Gray denotes the performance is evaluated on LVIS minival [54] or LVIS 0.5. The metric is box AP.

Method	Image Backbone	Detector	Training Source	VOC AP_{50}	COCO AP	COCO AP_{50}	O365 AP	O365 AP_{50}	OI AP_{50}	LVIS v1.0 [28] AP_r/AP_c/AP_f/AP
MDETR [54]	R101	DETR	GoldG [55], COCO Cap [169]	-	-	-	-	-	-	20.9/24.9/24.3/24.2
ViLD [38]	R50-FPN	MRCNN	LVIS v1.0	72.2	36.6	55.6	11.8	18.2	-	-
GLIP-T [55]	Swin-T [20]	DyHead [193]	O365, GoldG [55], Cap4M [55]	-	46.3	-	-	-	-	$10.1_{20.8}$/$12.5_{21.4}$/$25.5_{31.0}$/$17.2_{26.0}$
Detic [149]	Swin-B [20]	CN2 [186]	LVIS v1.0, IN21K	-	-	-	-	21.5	55.2	-
OV-DETR [61]	R50-C4	Def-DETR [22]	LVIS v1.0	76.1	38.1	58.4	-	-	-	-
DetPro [60]	R50-FPN (SoCo [176])	MRCNN	LVIS v1.0	74.6	34.9	53.8	12.1	18.8	-	-
OWL-ViT [178]	ViT-B/16 ($CLIP_V$)	DETR	O365, VG [107]	-	-	-	-	-	-	23.6/-/-/26.7
DetCLIP-T [141]	Swin-T [20]	ATSS [164]	O365, GoldG [55], YFCC1M [194]	-	-	-	-	-	-	33.2/35.7/36.4/35.9
F-VLM [63]	R50-FPN	MRCNN	LVIS v1.0	-	32.5	53.1	11.9	19.2	-	-
GLIPv2-T [152]	Swin-T [20]	DyHead [193]	O365, GoldG [55], Cap4M [55]	-	-	-	-	-	-	-/-/-/29.0
PB-OVD [58]	R50	MRCNN	COCO, COCO Cap [169], VG [107], SBU [173]	59.2	-	-	6.9	-	-	-/-/-/8.0
RKDWTF [154]	R50-FPN	MRCNN	IN21K, LVIS v1.0	-	-	56.6	-	22.3	42.9	-
GridCLIP [166]	CLIP (R50)	FCOS	LVIS v1.0	70.9	34.7	52.2	-	-	-	-
UniDetector [179]	R50-C4 (RegionCLIP [57])	FRCNN	COCO, O365, OI	-	-	-	-	-	-	18.0/19.2/21.2/19.8
RO-ViT [139]	ViT-B/16 [19]	MRCNN	LVIS v1.0	-	-	-	14.0	22.3	-	-
3Ways [150]	NF-F0 [187]-FPN	FCOS [99] (T-Head [188])	LVIS v1.0	-	41.5	-	16.4	-	-	-
DetCLIPv2-T [56]	Swin-T [20]	ATSS [164]	O365, GoldG [55], CC3M [170], CC12M [189]	-	-	-	-	-	-	36.0/41.7/40.0/40.4
OpenSeeD [181]	Swin-T [20]	Mask DINO [195]	COCO, O365	-	-	-	-	-	-	21.8/-/-/-
MMC (Text) [182]	R50-FPN	CN2 [186]	IN21K, LVIS v1.0	-	-	-	16.6	23.1	-	-
Grounding DINO-T [153]	Swin-T [20]	DINO [196]	O365, GoldG [55], Cap4M [55]	-	48.4	-	-	-	-	18.1/23.3/32.7/27.4
MQ-Det [143]	Swin-T [20]	GLIP [55]	O365	-	-	-	-	-	-	$15.4_{21.0}$/$18.4_{27.5}$/$30.4_{34.6}$/$22.4_{30.4}$

6.1.4 *Transfer Learning-Based*

This methodology aims to transfer VLMs text and image encoder to downstream segmentation tasks. The transfer strategy is explored in the following aspect: 1) only adopting VLMs text encoder for open-vocabulary classification; 2) leveraging frozen VLMs image encoder as a feature extractor to the segmentor backbone; 3) directly fine-tuning VLMs image encoder on segmentation datasets; 4) employing visual prompts or attaching a lightweight adapter to frozen VLMs image encoder for feature adaptation. A detailed

TABLE 9: Open-vocabulary semantic segmentation performance under the gOVE protocol. The base/novel split is 156/15 [50], 15/5 [50], 572/275 [67] for COCO-Stuff [110], Pascal VOC [26], and ADE20K [111], respectively.

Method	Image Backbone	Segmentor	Image-Text pairs	Prompts	Ensemble	Pascal VOC mIoU (B/N/HM)	COCO-Stuff mIoU (B/N/HM)	ADE20K mIoU (B/N/HM)
ZegFormer [67]	R101-FPN	MF [23]	✗	T (cat)	✓	86.4/63.6/73.3	36.6/33.2/34.8	-
ZegFormer [67]	R50-FPN	MF [23]	✗	T (cat)	✓	-	-	17.4/5.3/8.1
ZSSeg [197]	R101	MF [23]	✗	L (cat)	✗	83.5/72.5/77.5	39.3/36.3/37.8	-
ZegCLIP [198]	ViT-B/16 [19]	-	✗	T (cat)	✗	91.9/77.8/84.3	40.2/41.4/40.8	-
MVP-SEG+ [199]	CLIP_V (R50)	DLv2 [100]	MaskCLIP+ [200]	T (cat)	✗	89.0/87.4/88.2	38.3/55.8/45.5	-
TagCLIP [201]	CLIP_V (ViT-B/16)	SegViT [202]	✗	T (cat)	✗	93.5/85.2/89.2	40.7/43.1/41.9	-

TABLE 10: Open-vocabulary semantic segmentation performance on the validation set of ADE20K [111] (A-847 and A-150), Pascal Context [26] (PC-459 and PC-59), and Pascal VOC [26] (PAS-20) datasets under the CDTE protocol. Cat-ens uses synonyms or subcategories with class names to fill prompt templates.

Method	Image Backbone	Segmentor	Training Source	Prompts	mIoU A-847	A-150	PC-459	PC-59	PAS-20
GroupViT [64]	ViT-S [19]	-	CC12M [189], YFCC14M [194]	T (cat)	-	-	-	22.4	52.3
LSeg+ [42]	R101-FPN	SRB [43]	COCO Panoptic [10]	T (cat)	2.5	13.0	5.2	36.0	59.0
ViL-Seg [203]	ViT-B/16 [19]	-	CC12M [189]	T (cat)	-	-	-	15.9	33.6
SegCLIP [204]	ViT (CLIP_V)	-	COCO [27], CC [170]	T (cat)	-	-	-	24.7	52.6
OpenSeg [42]	R101-FPN	-	COCO Panoptic [10], COCO Cap [169]	T (cat-ens)	4.0	15.3	6.5	36.9	60.0
SimSeg [205]	ViT-S [19]	-	CC3M [170], CC12M [189]	T (cat)	-	-	-	25.8	56.6
ZegFormer [67], [206]	R101	MF [23]	COCO-Stuff [110]	T (cat)	5.6	18.0	10.4	45.5	89.5
ZSSeg [197]	R101	MF [23]	COCO-Stuff [110]	L (cat)	7.0	20.5	-	47.7	-
OVSegmentor [207]	ViT-B [19]	-	CC12M [189]	✗	-	5.6	-	20.4	53.8
OVSeg [68]	R101c [100]	MF [23]	COCO-Stuff [110], COCO Cap [169]	T (cat)	7.1	24.8	11.0	53.3	92.6
SAN [208]	CLIP (ViT-B/16)	-	COCO-Stuff [110]	T (cat)	10.1	27.5	12.6	53.8	94.0
PACL [209]	CLIP_V (ViT-B/16)	-	CC3M [170], CC12M [189], YFCC15M [194]	T (cat)	-	31.4	-	50.1	72.3
CAT-Seg [206]	Swin-B [20]	-	COCO-Stuff [110]	T (cat)	10.8	31.5	20.4	62.0	96.6
OVDiff [69]	UNet [210]	-	CLIP [44], StableDiffusion [104]	T (cat)	-	-	-	30.1	67.1
TCL [211]	CLIP_V (ViT-B/16)	-	CC3M [170], CC12M [189]	T (cat)	-	17.1	-	33.9	83.2

TABLE 11: Open-vocabulary instance segmentation performance on COCO [27] and OpenImages [109] datasets under the gOVE protocol. The base/novel split is 48/17 [40] for COCO and 200/100 [41] for OpenImages.

Method	Image Backbone	Segmentor	Image-Text pairs	Prompts	COCO AP_N	AP_B	AP	OpenImages AP_N	AP_B	AP
XPM [41]	R50-C4	MRCNN	CC [170]	✗	21.6	41.5	36.3	22.7	49.8	40.7
Mask-free OVIS	R50-C4	MRCNN	COCO, OpenImages	T (cat)	25.0	-	-	25.8	-	-
CGG [70]	R50	M2F [25]	COCO Cap [169]	✗	28.4	46.0	41.4	-	-	-
D^2Zero [212]	R50	M2F [25]	-	T (cat)	15.8	54.1	24.5	-	-	-

TABLE 12: Open-vocabulary panoptic segmentation performance on COCO [27] and ADE20k [111] dataset.

Method	Image Backbone	Segmentor	Prompts	COCO PQ^s	SQ^s	RQ^s	PQ^u	SQ^u	RQ^u	ADE20K PQ	PQ^{th}	PQ^{st}	SQ	RQ
FreeSeg [213]	R101	M2F [25]	L (cat)	31.4	78.3	38.9	29.8	79.2	37.6	-	-	-	-	-
ODISE [74]	UNet [210]	M2F [25]	T (cat-desc)	-	-	-	-	-	-	22.6	-	-	-	-
MaskCLIP [73]	R50	M2F [25]	cat	-	-	-	-	-	-	15.1	13.5	18.3	70.5	19.2
OPSNet [214]	CLIP_V (R50)	M2F [25]	cat	-	-	-	-	-	-	17.7	15.6	21.9	54.9	21.6
PADing [72]	R50	M2F [25]	T (cat)	41.5	80.6	49.7	15.3	72.8	18.4					

comparison is given in Fig. 5.

VLMs Text Encoder as Classifier. Methods in this category only adopt the CLIP text encoder as the classifier and discard the CLIP image encoder, while being simple, it prevents unleashing the full potential of the CLIP image encoder. LSeg [43] simply replaces the learnable weights of the classifier with text embeddings from the CLIP text encoder. The image backbone is initialized from ImageNet pretrained weights and trained on downstream segmentation datasets. SAZS [219] focuses on improving boundary segmentation performance supervised by ground-truth boundaries. During inference, SAZS fuses the predictions with eigensegments obtained through spectral analysis on a self-supervised DINO [215].

VLMs Frozen Image Encoder as Feature Extractor. Besides using VLMs text encoder, this line of work also adopts the frozen image encoder as a feature extractor. ZegFormer [67] decouples the per-pixel semantic segmentation into a class-agnostic grouping and a segment-level recognition stage using MaskFormer [23]. It forwards cropped

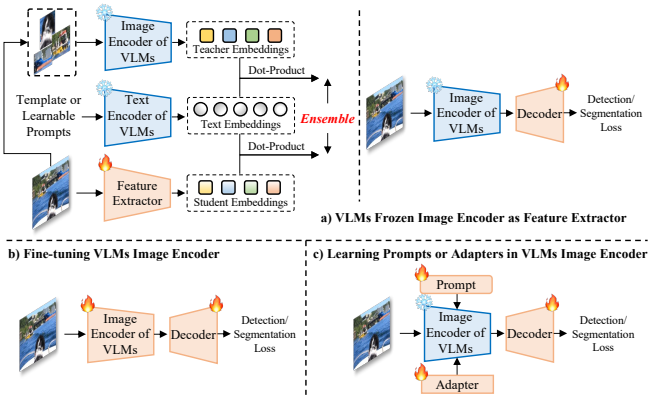

Fig. 5: Framework for transfer learning-based models.

masked regions to the frozen CLIP image encoder and ensembles its predicted scores via geometric mean with the from segmentor. ZSSeg [197] proposes the same architecture as ZegFormer [67] except that ZSSeg adopts learnable prompts similar to CoOp [159] without ensembling. MaskCLIP+ [200] removes the final attention pooling layer in the CLIP image encoder and directly classifies patch tokens using text embeddings without segmentor backbone. Key smoothing and prompt denoising are further proposed. MVP-SEG [199] also directly regards CLIP frozen image encoder as the only backbone and proposes multi-view prompt learning optimized by orthogonal contrastive loss to focus on different object parts. ReCo [220] first retrieves an archive of exemplar images for each class from unlabelled images, then leverages MoCo [221] to extract seed pixels across exemplar images to construct reference embeddings as the 1×1 convolution classifier on the fly. The prediction is ensembled with DenseCLIP [222]. Peekaboo [223] explores how off-the-shelf StableDiffusion (SD) [104] can perform grouping pixels with the proposed dream loss. OVDiff [69] is also entirely training-free, the same as Peekaboo [223], it relies only on SD [104]. 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 within the same modality. POMP [224] condenses semantic concepts over twenty-thousand classes into the learned prompts via the added prompt pretraining stage. It introduces local contrast and local correction strategy to reduce memory consumption (CoOp [159] requires memory proportional to the number of classes.) and improve generalization ability.

Fine-tuning VLMs Image Encoder. This group finetunes the CLIP image encoder to adapt its feature representations to the segmentation task. DenseCLIP [222] establishes pixel-text matching in latent feature space by removing the final multi-head self-attention pooling layer in the image encoder of CLIP. OVSeg [68] proposes mask prompt tuning along with finetuning CLIP image encoder on the constructed mask-category pairs (self-generated pseudo labels) from caption dataset to address the performance bottleneck in the domain gap between masked image crops with blank areas and natural images used to pretrain CLIP. CAT-Seg [206] devises a cost aggregation module including spatial and class aggregation to produce the segmentation map. It finetunes the CLIP image encoder but only the attention layers as finetuning all parameters of CLIP harms its open-vocabulary capabilities evidenced by previous work.

Learning Prompts or Adapters in VLMs Image Encoder. Visual prompts [82] or lightweight adapters [225], [226] is a trainable module inserted into the CLIP image encoder. Compared to fine-tuning the CLIP image encoder, learning prompts or adapters can better preserve the generalization ability in novel classes. TagCLIP [201] adopts deep prompt tuning (DPT) [82] and a learnable trusty token generating trusty maps used to weigh the raw segmentation map to adapt CLIP and judge the reliability. ZegCLIP [198] also proposes DPT that prepends trainable prompt tokens as additional input to each layer instead of finetuning all model parameters. CLIPSeg [227] attaches a lightweight decoder to CLIP image encoder with U-Net-like skip connections conditioned on text embeddings using FiLM [228]. SAN [208] attaches a lightweight vision transformer called side adapter network 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 [229] discovers that CLIP has opposite visualization results similar to the findings of SimSeg [205] and has noisy activations. The proposed architecture surgery replaces Q-K self-attention with V-V self-attention without FFN, forming a dual-path inference route alongside the CLIP image encoder. It avoids Q-K interaction that causes the opposite visualization problem. Another feature surgery identifies and removes redundant features to reduce noisy activations.

6.2 Open-Vocabulary Instance Segmentation

Region-Aware Training. CGG [70] achieves the region-text alignment via a grounding loss, but not with the whole caption as in OVR-CNN [40]. 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, which is complementary to the caption grounding loss. D^2Zero [212] proposes an unseen-constrained feature extractor and an input-conditional classifier to address the bias issue in Section 3.3. It further proposes image-adaptive background representations, which compared to the static BARPN [33] can better generalize to proposing novel foreground instances.

Pseudo-Labeling. XPM [41] first trains a teacher model using available base annotations, then self-trains a student model. The pseudo regions are selected as the most compatible region *w.r.t* the object nouns in the caption. However, pseudo masks contain noises that degrade performance, hence the student is trained to predict the noise level (each pixel in pseudo masks is assumed to be corrupted by a Gaussian noise) in pseudo masks to downweight incorrect teacher predictions. Mask-free OVIS [71] performs iterative masking using ALBEF [158] and GradCAM [157] to generate pseudo-instances both for base and novel categories. It avoids training base categories using strong supervision and novel categories using weak supervision, thus alleviating the overfitting issue.

6.3 Open-Vocabulary Panoptic Segmentation

FreeSeg [213] accomplishes semantic, instance, and panoptic segmentation in the same architecture. It directly feeds

masked crops into CLIP image encoder for classification. He *et al.* [72] argue that synthesizers [80], [81], [129] in Section 4.1.2 with several linear layers do not consider the feature granularity gap between image and text modality. They employ 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 [72] decouples visual features into semantic-related and semantic-unrelated parts and only aligns the semantic-related parts with the inter-class structure in semantic space. OPSNet [214] modulates mask embeddings and CLIP embeddings via the domain similarity coefficient, together with several meticulous components. ODISE [74] resorts to text-to-image diffusion models [104] as the mask feature extractor instead of training from scratch only on base categories. It also proposes an implicit captioner via CLIP image encoder to map images into pseudo words. The training is driven by a bidirectional grounding loss similar to the region-aware training methodology. MaskCLIP [73] designs mask class tokens to extract dense image features corresponding to each mask area via the proposed relative mask attention mechanism similar to relative positional encoding. Same as ODISE [74], HIPIE [230] ensembles classification logits with CLIP. It can hierarchically segments things, stuff, and object parts such as "human ear" or "cat head". It employs two separate decoders for things and stuff instead of one unified decoder.

7 Open-Vocabulary Beyond Images

Besides images, we cover open-vocabulary 3D scene understanding and video instance segmentation in this section.

7.1 Open-Vocabulary 3D Scene Understanding

Open-vocabulary 3D scene understanding suffers a more severe data scarcity issue, even pairing point clouds with text descriptions is not available up to now, hence the methods typically bridge the point-cloud and text modality via image modality, where VLMs (*e.g.*, CLIP) step in to guide the association. OV-3DET [231] is the earliest work that proposes the open-vocabulary point-cloud detection task. It first leverages pseudo-boxes from an open-vocabulary 2D detector Detic [149] to address 3D point-cloud localization without any manually annotated boxes. Then, DTCC is proposed to correct the biased contrastive learning and connect image, text, and point-cloud modalities to enable open-vocabulary classification. SeCondPoint [232] and 3DGenZ [75] are the first attempts at open-vocabulary point-cloud segmentation task, and they basically follow the pipeline of novel visual feature synthesis Section 3.2 in 2D scenarios. Following LSeg [43], PLA [233] first establishes a baseline model termed LSeg-3D, upon which a calibration module is added to avoid over-confident predictions on base classes regardless of their correctness. Then, PLA builds hierarchical coarse-to-fine point-caption pairs, *i.e.*, scene-, view-, and entity-level point-caption association via a pretrained captioning model [190], effectively facilitating learning from vocabulary-rich language supervisions. However, the pseudo-captions at the view level only cover sparse and salient objects in a scene, failing to provide fine-grained language descriptions. To enable dense regional point-language associations, RegionPLC [234] (authors of PLA) proposes region-level visual prompts (image patch via sliding-window and object proposal via 2D detector) to improve eliciting knowledge from foundation models via captioning. A point-discriminative contrastive learning objective is further proposed that makes the gradient of each point unique. OpenScene [76] embeds point features into the feature space of CLIP, 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 features, point features can be aligned with text features. OpenMask3D [235] aggregates per-mask features via multi-view fusion of CLIP-based image embeddings instead of embedding point-cloud features into a common space or distilling knowledge into the 3D model for the open-vocabulary 3D instance segmentation task.

7.2 Open-Vocabulary Video Instance Segmentation

MindVLT [236] first proposes the 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 (LV-VIS) covering 1,212 categories for benchmarking the task. The proposed MindVLT architecture leverages CLIP text encoder to classify queries from its memory-induced tracking module. The concurrent work OpenVIS [77] first proposes instances in a frame exhaustively based on Mask2Former [25], then in the second stage designs SquareCrop that avoids distorting the aspect ratio of instances to better conform to the CLIP image encoder. The open-vocabulary classification is enabled by sending cropped images and template prompts filled with class names to CLIP image and text encoder, respectively.

8 Conclusions and Outlook

8.1 Conclusions

We cover a broad and concrete development of OVD and OVS in this survey. First, background including definition, related domains and tasks, canonical closed-set detectors and segmentors, large VLMs, datasets, and evaluation protocols are given. Then we detail more than hundreds of methods. At the task level, both 2D detection and semantic/instance/panoptic segmentation tasks are discussed, along with 3D scene and video understanding. At the methodology level, we pivot on the permission and usage of weak supervision signals and group methods into six main categories which are universal across tasks. We give a general overview (strengths and weaknesses) of each methodology. In addition, we benchmark the performance of state-of-the-art methods for each task. In the following, challenges and future promising directions are discussed to facilitate future research.

8.2 Challenges

1) Accurate Region-Word Correspondence. Though image-text pairs are cheap and abundant, however, due to the partial labeling problem, *i.e.*, the object nouns in the caption may only cover salient objects, the number of which is far less than the proposals. Thus, many objects may not find the

matching words, and training efficiency is compromised. In addition, the region-word correspondence is weak and noisy, how to improve the quality of region-word pairs and counteract the negative effect of false region-word pairs is non-negligible to further scale up training data.

2) Efficient VLMs Utilization. Besides image-text pairs, large VLMs provide well-aligned visual-semantic space for downstream models to learn. However, there is a huge domain gap between image crops, masked regions, and natural images used to pretrain VLMs, which prevents unleashing the full potential of VLMs on open-vocabulary recognition. Better fine-tuning, prompting, or learning adapters to avoid catastrophic forgetting without losing generalization ability is also in urgent need. Besides, ensembling predictions of VLMs during inference hinders open-vocabulary detectors and segmentors moving toward real-time.

8.3 Future Directions

1) Enabling Open-Vocabulary on Other Tasks. Currently, OVD and OVS in 2D are vigorously studied, however, other tasks including open-vocabulary 3D scene understanding, video analysis [237], action recognition, object tracking, and human-object interaction [238], *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 become a mainstream trend.

2) Improving Region-Word Correspondence. Though the pseudo-labeling methodology can build a coarse region-word correspondence, it fails to compete against human annotations. Recently SAM [239] demonstrates that an iterative strategy of pseudo-labeling and model training can achieve astonishing segmentation performance, semi-supervised detection [184], [185] also shares the same spirit which evolves the teacher online to refine pseudo-labels. On the contrary, current OVD and OVS only generate pseudo-labels in one go without multi-round or online refining. Another perspective to improve region-word correspondence is harnessing LLMs for automatic and thorough caption generation [182] thereby alleviating the partial labeling problem.

3) Reducing Training and Inference Cost. Current methods typically adopt a heavy backbone (ResNet [11], ViT [19], or Swin [20]) and neck architecture unsuitable for real-time applications. There are many issues to deploy OVD and OVS in real-time scenarios. For example, the low recall rate of novel objects in real-time models, and distilling the knowledge of large VLMs into these small-scale models remains questionable, of which the architectures have limited learning capacity due to the trade-off between performance and speed. Another weakness is that, although knowledge distillation-based methods can reduce the inference time with student model, the heavy training cost for current OVD and OVS is still ineluctable.

4) Endowing Reasoning and Interactive Capability. Endowing reasoning capability of user intentions and enabling interactive detection within a language context have long been standing toward artificial general intelligence. ContextDET [240] and DetGPT [241] are two pioneering works toward this goal, in which multimodal LLMs readily serve to infer concise class names from ambiguous user instructions. Besides the visual input, using languages as the interaction medium, detecting and segmenting by the point of mouse like SAM [239], and any other different input formats are also worth exploring.

5) Unifying OVD and OVS. Unification is an inevitable trend for CV. Though there are several works addressing different segmentation tasks simultaneously [74], [213], [230], [242] or training on multiple detection datasets [179], [180], a universal foundational model for all tasks and datasets [181] remains barely untouched, or even further, accomplishing 2D and 3D open-vocabulary perception simultaneously can be more challenging.

References

[1] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.

[2] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *ECCV*, 2022.

[3] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *MIA*, 2017.

[4] F. Zhu, Y. Zhu, V. Lee, X. Liang, and X. Chang, "Deep learning for embodied vision navigation: A survey," *arXiv*, 2021.

[5] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *CVPR*, 2018.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *NeurIPS*, 2015.

[7] I. Misra, R. Girdhar, and A. Joulin, "An end-to-end transformer model for 3d object detection," in *ICCV*, 2021.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017.

[10] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *CVPR*, 2019.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[12] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *CVPR*, 2022.

[13] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv*, 2018.

[14] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.

[15] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017.

[16] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang *et al.*, "Hybrid task cascade for instance segmentation," in *CVPR*, 2019.

[17] B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, and L.-C. Chen, "Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation," in *CVPR*, 2020.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, 2017.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv*, 2020.

[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021.

[21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020.

[22] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *ICLR*, 2021.

[23] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *NeurIPS*, 2021.

[24] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *NeurIPS*, 2021.

[25] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *CVPR*, 2022.

[26] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *IJCV*, 2015.

[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[28] A. Gupta, P. Dollar, and R. Girshick, "Lvis: A dataset for large vocabulary instance segmentation," in *CVPR*, 2019.

[29] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, "Zero-shot object detection," in *ECCV*, 2018.

[30] S. Rahman, S. Khan, and F. Porikli, "Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts," in *ACCV*, 2019.

[31] P. Zhu, H. Wang, and V. Saligrama, "Zero shot detection," *TCSVT*, 2019.

[32] M. Bucher, T.-H. Vu, M. Cord, and P. Pérez, "Zero-shot semantic segmentation," *NeurIPS*, 2019.

[33] Y. Zheng, J. Wu, Y. Qin, F. Zhang, and L. Cui, "Zero-shot instance segmentation," in *CVPR*, 2021.

[34] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *NeurIPS*, 2013.

[35] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv*, 2016.

[36] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014.

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv*, 2018.

[38] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," *arXiv*, 2021.

[39] S. Khandelwal, A. Nambirajan, B. Siddiquie, J. Eledath, and L. Sigal, "Frustratingly simple but effective zero-shot detection and segmentation: Analysis and a strong baseline," *arXiv*, 2023.

[40] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, "Open-vocabulary object detection using captions," in *CVPR*, 2021.

[41] D. Huynh, J. Kuen, Z. Lin, J. Gu, and E. Elhamifar, "Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling," in *CVPR*, 2022.

[42] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *ECCV*, 2022.

[43] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," *arXiv*, 2022.

[44] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[45] S. Rahman, S. Khan, and N. Barnes, "Improved visual-semantic alignment for zero-shot object detection," in *AAAI*, 2020.

[46] D. Gupta, A. Anantharaman, N. Mamgain, V. N. Balasubramanian, C. Jawahar *et al.*, "A multi-space approach to zero-shot object detection," in *WACV*, 2020.

[47] P. Zhu, H. Wang, and V. Saligrama, "Don't even look once: Synthesizing features for zero-shot detection," in *CVPR*, 2020.

[48] S. Zhao, C. Gao, Y. Shao, L. Li, C. Yu, Z. Ji, and N. Sang, "Gtnet: Generative transfer network for zero-shot object detection," in *AAAI*, 2020.

[49] P. Huang, J. Han, D. Cheng, and D. Zhang, "Robust region feature synthesizer for zero-shot object detection," in *CVPR*, 2022.

[50] Y. Xian, S. Choudhury, Y. He, B. Schiele, and Z. Akata, "Semantic projection network for zero-and few-label semantic segmentation," in *CVPR*, 2019.

[51] D. Baek, Y. Oh, and B. Ham, "Exploiting a joint embedding space for generalized zero-shot semantic segmentation," in *ICCV*, 2021.

[52] H. Zhang and H. Ding, "Prototypical matching and open set rejection for zero-shot semantic segmentation," in *ICCV*, 2021.

[53] Z. Gu, S. Zhou, L. Niu, Z. Zhao, and L. Zhang, "Context-aware feature generation for zero-shot semantic segmentation," in *ACM MM*, 2020.

[54] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "Mdetr - modulated detection for end-to-end multi-modal understanding," in *ICCV*, 2021.

[55] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, "Grounded language-image pre-training," in *CVPR*, 2022.

[56] L. Yao, J. Han, X. Liang, D. Xu, W. Zhang, Z. Li, and H. Xu, "Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment," in *CVPR*, 2023.

[57] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li *et al.*, "Regionclip: Region-based language-image pretraining," in *CVPR*, 2022.

[58] M. Gao, C. Xing, J. C. Niebles, J. Li, R. Xu, W. Liu, and C. Xiong, "Open vocabulary object detection with pseudo bounding-box labels," in *ECCV*, 2022.

[59] H.-C. Cho, W. Y. Jhoo, W. Kang, and B. Roh, "Open-vocabulary object detection using pseudo caption labels," *arXiv*, 2023.

[60] Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, and G. Li, "Learning to prompt for open-vocabulary object detection with vision-language model," in *CVPR*, 2022.

[61] Y. Zang, W. Li, K. Zhou, C. Huang, and C. C. Loy, "Open-vocabulary detr with conditional matching," in *ECCV*, 2022.

[62] S. Wu, W. Zhang, S. Jin, W. Liu, and C. C. Loy, "Aligning bag of regions for open-vocabulary object detection," in *CVPR*, 2023.

[63] W. Kuo, Y. Cui, X. Gu, A. Piergiovanni, and A. Angelova, "F-vlm: Open-vocabulary object detection upon frozen vision and language models," *arXiv*, 2022.

[64] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, "Groupvit: Semantic segmentation emerges from text supervision," in *CVPR*, 2022.

[65] N. Zabari and Y. Hoshen, "Open-vocabulary semantic segmentation using test-time distillation," in *ECCV*, 2022.

[66] K. Han, Y. Liu, J. H. Liew, H. Ding, Y. Wei, J. Liu, Y. Wang, Y. Tang, Y. Yang, J. Feng *et al.*, "Global knowledge calibration for fast open-vocabulary segmentation," *arXiv*, 2023.

[67] J. Ding, N. Xue, G.-S. Xia, and D. Dai, "Decoupling zero-shot semantic segmentation," in *CVPR*, 2022.

[68] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, "Open-vocabulary semantic segmentation with mask-adapted clip," in *CVPR*, 2023.

[69] L. Karazija, I. Laina, A. Vedaldi, and C. Rupprecht, "Diffusion models for zero-shot open-vocabulary segmentation," *arXiv*, 2023.

[70] J. Wu, X. Li, H. Ding, X. Li, G. Cheng, Y. Tong, and C. C. Loy, "Betrayed by captions: Joint caption grounding and generation for open vocabulary instance segmentation," *arXiv*, 2023.

[71] V. VS, N. Yu, C. Xing, C. Qin, M. Gao, J. C. Niebles, V. M. Patel, and R. Xu, "Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations," in *CVPR*, 2023.

[72] S. He, H. Ding, and W. Jiang, "Primitive generation and semantic-related alignment for universal zero-shot segmentation," in *CVPR*, 2023.

[73] Z. Ding, J. Wang, and Z. Tu, "Open-vocabulary panoptic segmentation with maskclip," *arXiv*, 2022.

[74] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *CVPR*, 2023.

[75] B. Michele, A. Boulch, G. Puy, M. Bucher, and R. Marlet, "Generative zero-shot learning for semantic segmentation of 3d point clouds," in *3DV*, 2021.

[76] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, T. Funkhouser *et al.*, "Openscene: 3d scene understanding with open vocabularies," in *CVPR*, 2023.

[77] P. Guo, T. Huang, P. He, X. Liu, T. Xiao, Z. Chen, and W. Zhang, "Openvis: Open-vocabulary video instance segmentation," *arXiv*, 2023.

[78] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha, "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," in *ECCV*, 2016.

[79] Y. Li, K. Swersky, and R. Zemel, "Generative moment matching networks," in *ICML*, 2015.

[80] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *ACM Communications*, 2020.

[81] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *NeurIPS*, 2015.

[82] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *ECCV*, 2022.

[83] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *ECCV*, 2016.

[84] V. K. Nagaraja, V. I. Morariu, and L. S. Davis, "Modeling context between objects for referring expression understanding," in *ECCV*, 2016.

[85] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *EMNLP*, 2014.

[86] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *IJCV*, 2015.

[87] C. Zhu, Y. Zhou, Y. Shen, G. Luo, X. Pan, M. Lin, C. Chen, L. Cao, X. Sun, and R. Ji, "Seqtr: A simple yet universal network for visual grounding," in *ECCV*, 2022.

[88] A. Dhamija, M. Gunther, J. Ventura, and T. Boult, "The overlooked elephant of object detection: Open set," in *WACV*, 2020.

[89] D. Miller, L. Nicholson, F. Dayoub, and N. Sünderhauf, "Dropout sampling for robust object detection in open-set conditions," in *ICRA*, 2018.

[90] T. Pham, T.-T. Do, G. Carneiro, I. Reid *et al.*, "Bayesian semantic instance segmentation in open set world," in *ECCV*, 2018.

[91] J. Hwang, S. W. Oh, J.-Y. Lee, and B. Han, "Exemplar-based open-set panoptic segmentation network," in *CVPR*, 2021.

[92] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *NeurIPS*, 2020.

[93] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *arXiv*, 2021.

[94] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv*, 2016.

[95] K. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian, "Towards open world object detection," in *CVPR*, 2021.

[96] A. Gupta, S. Narayan, K. Joseph, S. Khan, F. S. Khan, and M. Shah, "Ow-detr: Open-world detection transformer," in *CVPR*, 2022.

[97] J. Cen, P. Yun, J. Cai, M. Y. Wang, and M. Liu, "Deep metric learning for open world semantic segmentation," in *ICCV*, 2021.

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

[99] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *ICCV*, 2019.

[100] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *TPAMI*, 2017.

[101] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021.

[102] J. Yang, C. Li, P. Zhang, B. Xiao, C. Liu, L. Yuan, and J. Gao, "Unified contrastive learning in image-text-label space," in *CVPR*, 2022.

[103] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, 2020.

[104] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[105] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *NeurIPS*, 2022.

[106] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *IJCV*, 2015.

[107] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, 2017.

[108] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun, "Objects365: A large-scale, high-quality dataset for object detection," in *ICCV*, 2019.

[109] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *IJCV*, 2020.

[110] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *CVPR*, 2018.

[111] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *CVPR*, 2017.

[112] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *CVPR*, 2014.

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

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

[115] R. Luo, N. Zhang, B. Han, and L. Yang, "Context-aware zero-shot recognition," in *AAAI*, 2020.

[116] Z. Li, L. Yao, X. Zhang, X. Wang, S. Kanhere, and H. Zhang, "Zero-shot object detection with textual descriptions," in *AAAI*, 2019.

[117] Y. Zheng, R. Huang, C. Han, X. Huang, and L. Cui, "Background learnable cascade for zero-shot object detection," in *ACCV*, 2020.

[118] S. Rahman, S. Khan, and N. Barnes, "Transductive learning for zero-shot object detection," in *ICCV*, 2019.

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

[120] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017.

[121] B. Demirel, R. G. Cinbis, and N. Ikizler-Cinbis, "Zero-shot object detection by hybrid region embedding," *arXiv*, 2018.

[122] Y. Li, Y. Shao, and D. Wang, "Context-guided super-class inference for zero-shot detection," in *CVPRW*, 2020.

[123] Y. Li, P. Li, H. Cui, and D. Wang, "Inference fusion with associative semantics for unseen object detection," in *AAAI*, 2021.

[124] C. Yan, X. Chang, M. Luo, H. Liu, X. Zhang, and Q. Zheng, "Semantics-guided contrastive network for zero-shot object detection," *TPAMI*, 2022.

[125] L. Zhang, X. Wang, L. Yao, L. Wu, and F. Zheng, "Zero-shot object detection via learning an embedding from semantic space to visual space," in *IJCAI*, 2020.

[126] G. Dinu, A. Lazaridou, and M. Baroni, "Improving zero-shot learning by mitigating the hubness problem," *arXiv*, 2014.

[127] N. Hayat, M. Hayat, S. Rahman, S. Khan, S. W. Zamir, and F. S. Khan, "Synthesizing the unseen for zero-shot object detection," in *ACCV*, 2020.

[128] Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, and M.-H. Yang, "Mode seeking generative adversarial networks for diverse image synthesis," in *CVPR*, 2019.

[129] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *ICML*, 2017.

[130] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017.

[131] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *CVPR*, 2009.

[132] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI*, 2017.

[133] P. Hu, S. Sclaroff, and K. Saenko, "Uncertainty-aware learning for zero-shot semantic segmentation," *NeurIPS*, 2020.

[134] N. Kato, T. Yamasaki, and K. Aizawa, "Zero-shot semantic segmentation via variational mapping," in *ICCVW*, 2019.

[135] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *ICCV*, 2019.

[136] P. Li, Y. Wei, and Y. Yang, "Consistent structural relation learning for zero-shot segmentation," *NeurIPS*, 2020.

[137] J. Cheng, S. Nandi, P. Natarajan, and W. Abd-Almageed, "Sign: Spatial-information incorporated generative network for generalized zero-shot semantic segmentation," in *ICCV*, 2021.

[138] M. A. Bravo, S. Mittal, and T. Brox, "Localized vision-language matching for open-vocabulary object detection," in *DAGM GCPR*, 2022.

[139] D. Kim, A. Angelova, and W. Kuo, "Region-aware pretraining for open-vocabulary object detection with vision transformers," in *CVPR*, 2023.

[140] D. Kim, T.-Y. Lin, A. Angelova, I. S. Kweon, and W. Kuo, "Learning open-world object proposals without learning to classify," *Robotics and Automation*, 2022.

[141] L. Yao, J. Han, Y. Wen, X. Liang, D. Xu, W. Zhang, Z. Li, C. Xu, and H. Xu, "Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection," *arXiv*, 2022.

[142] M. Maaz, H. Rasheed, S. Khan, F. S. Khan, R. M. Anwer, and M.-H. Yang, "Class-agnostic object detection with multi-modal transformer," in *ECCV*, 2022.

[143] Y. Xu, M. Zhang, C. Fu, P. Chen, X. Yang, K. Li, and C. Xu, "Multi-modal queried object detection in the wild," *arXiv*, 2023.

[144] H. Shi, M. Hayat, and J. Cai, "Open-vocabulary object detection via scene graph discovery," *arXiv*, 2023.

[145] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv*, 2019.

[146] P. Chen, K. Sheng, M. Zhang, Y. Shen, K. Li, and C. Shen, "Open vocabulary object detection with proposal mining and prediction equalization," *arXiv*, 2022.

[147] C. Lin, P. Sun, Y. Jiang, P. Luo, L. Qu, G. Haffari, Z. Yuan, and J. Cai, "Learning object-language alignments for open-vocabulary object detection," *arXiv*, 2022.

[148] X. Wu, F. Zhu, R. Zhao, and H. Li, "Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching," in *CVPR*, 2023.

[149] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *ECCV*, 2022.

[150] R. Arandjelović, A. Andonian, A. Mensch, O. J. Hénaff, J.-B. Alayrac, and A. Zisserman, "Three ways to improve feature alignment for open vocabulary detection," *arXiv*, 2023.

[151] S. Zhao, Z. Zhang, S. Schulter, L. Zhao, B. Vijay Kumar, A. Stathopoulos, M. Chandraker, and D. N. Metaxas, "Exploiting unlabeled data with vision and language models for object detection," in *ECCV*, 2022.

[152] H. Zhang, P. Zhang, X. Hu, Y.-C. Chen, L. Li, X. Dai, L. Wang, L. Yuan, J.-N. Hwang, and J. Gao, "Glipv2: Unifying localization and vision-language understanding," *NeurIPS*, 2022.

[153] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv*, 2023.

[154] H. Bangalath, M. Maaz, M. U. Khattak, S. H. Khan, and F. Shahbaz Khan, "Bridging the gap between object and image-level representations for open-vocabulary detection," *NeurIPS*, 2022.

[155] M. Maaz, H. B. Rasheed, S. H. Khan, F. S. Khan, R. M. Anwer, and M.-H. Yang, "Multi-modal transformers excel at class-agnostic object detection," *arXiv*, 2021.

[156] C. Feng, Y. Zhong, Z. Jie, X. Chu, H. Ren, X. Wei, W. Xie, and L. Ma, "Promptdet: Towards open-vocabulary detection using uncurated images," in *ECCV*, 2022.

[157] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017.

[158] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *NeurIPS*, 2021.

[159] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *IJCV*, 2022.

[160] Z. Liu, X. Hu, and R. Nevatia, "Efficient feature distillation for zero-shot detection," *arXiv*, 2023.

[161] L. Wang, Y. Liu, P. Du, Z. Ding, Y. Liao, Q. Qi, B. Chen, and S. Liu, "Object-aware distillation pyramid for open-vocabulary object detection," in *CVPR*, 2023.

[162] J. Xie and S. Zheng, "Zero-shot object detection through vision-language embedding alignment," in *ICDMW*, 2022.

[163] Z. Ma, G. Luo, J. Gao, L. Li, Y. Chen, S. Wang, C. Zhang, and W. Hu, "Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation," in *CVPR*, 2022.

[164] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *CVPR*, 2020.

[165] [Online]. Available: https://github.com/ultralytics/yolov5

[166] J. Lin and S. Gong, "Gridclip: One-stage object detection by grid-level clip representation learning," *arXiv*, 2023.

[167] H. Song and J. Bang, "Prompt-guided transformers for end-to-end open-vocabulary object detection," *arXiv*, 2023.

[168] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv*, 2018.

[169] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv*, 2015.

[170] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *ACL*, 2018.

[171] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "Dab-detr: Dynamic anchor boxes are better queries for detr," *arXiv*, 2022.

[172] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv*, 2021.

[173] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," *NeurIPS*, 2011.

[174] T. Wang, "Learning to detect and segment for open vocabulary object detection," in *CVPR*, 2023.

[175] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in *CVPRW*, 2020.

[176] F. Wei, Y. Gao, Z. Wu, H. Hu, and S. Lin, "Aligning pretraining for detection via object-level contrastive learning," *NeurIPS*, 2021.

[177] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," in *ECCV*, 2022.

[178] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen *et al.*, "Simple open-vocabulary object detection with vision transformers," *arXiv*, 2022.

[179] Z. Wang, Y. Li, X. Chen, S.-N. Lim, A. Torralba, H. Zhao, and S. Wang, "Detecting everything in the open world: Towards universal object detection," in *CVPR*, 2023.

[180] Y. Chen, M. Wang, A. Mittal, Z. Xu, P. Favaro, J. Tighe, and D. Modolo, "Scaledet: A scalable multi-dataset object detector," in *CVPR*, 2023.

[181] H. Zhang, F. Li, X. Zou, S. Liu, C. Li, J. Gao, J. Yang, and L. Zhang, "A simple framework for open-vocabulary segmentation and detection," *arXiv*, 2023.

[182] P. Kaul, W. Xie, and A. Zisserman, "Multi-modal classifiers for open-vocabulary object detection," *arXiv*, 2023.

[183] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *NeurIPS*, 2020.

[184] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda, "Unbiased teacher for semi-supervised object detection," in *ICLR*, 2020.

[185] M. Xu, Z. Zhang, H. Hu, J. Wang, L. Wang, F. Wei, X. Bai, and Z. Liu, "End-to-end semi-supervised object detection with soft teacher," in *ICCV*, 2021.

[186] X. Zhou, V. Koltun, and P. Krähenbühl, "Probabilistic two-stage detection," *arXiv*, 2021.

[187] A. Brock, S. De, S. L. Smith, and K. Simonyan, "High-performance large-scale image recognition without normalization," in *ICML*, 2021.

[188] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "Tood: Task-aligned one-stage object detection," in *ICCV*, 2021.

[189] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *CVPR*, 2021.

[190] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, 2019.

[191] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019.

[192] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "Lit: Zero-shot transfer with locked-image text tuning," in *CVPR*, 2022.

[193] X. Dai, Y. Chen, B. Xiao, D. Chen, M. Liu, L. Yuan, and L. Zhang, "Dynamic head: Unifying object detection heads with attentions," in *CVPR*, 2021.

[194] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *ACM Communications*, 2016.

[195] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask dino: Towards a unified transformer-based framework for object detection and segmentation," in *CVPR*, 2023.
[196] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv*, 2022.
[197] M. Xu, Z. Zhang, F. Wei, Y. Lin, Y. Cao, H. Hu, and X. Bai, "A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model," in *ECCV*, 2022.
[198] Z. Zhou, Y. Lei, B. Zhang, L. Liu, and Y. Liu, "Zegclip: Towards adapting clip for zero-shot semantic segmentation," in *CVPR*, 2023.
[199] J. Guo, Q. Wang, Y. Gao, X. Jiang, X. Tang, Y. Hu, and B. Zhang, "Mvp-seg: Multi-view prompt learning for open-vocabulary semantic segmentation," *arXiv*, 2023.
[200] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *ECCV*, 2022.
[201] J. Li, P. Chen, S. Qian, and J. Jia, "Tagclip: Improving discrimination ability of open-vocabulary semantic segmentation," *arXiv*, 2023.
[202] B. Zhang, Z. Tian, Q. Tang, X. Chu, X. Wei, C. Shen *et al.*, "Segvit: Semantic segmentation with plain vision transformers," *NeurIPS*, 2022.
[203] Q. Liu, Y. Wen, J. Han, C. Xu, H. Xu, and X. Liang, "Open-world semantic segmentation via contrasting and clustering vision-language embedding," in *ECCV*, 2022.
[204] H. Luo, J. Bao, Y. Wu, X. He, and T. Li, "Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation," *arXiv*, 2022.
[205] M. Xu, Z. Zhang, F. Wei, Y. Lin, Y. Cao, H. Hu, and X. Bai, "A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model," in *ECCV*, 2022.
[206] S. Cho, H. Shin, S. Hong, S. An, S. Lee, A. Arnab, P. H. Seo, and S. Kim, "Cat-seg: Cost aggregation for open-vocabulary semantic segmentation," *arXiv*, 2023.
[207] J. Xu, J. Hou, Y. Zhang, R. Feng, Y. Wang, Y. Qiao, and W. Xie, "Learning open-vocabulary semantic segmentation models from natural language supervision," in *CVPR*, 2023.
[208] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "Side adapter network for open-vocabulary semantic segmentation," in *CVPR*, 2023.
[209] J. Mukhoti, T.-Y. Lin, O. Poursaeed, R. Wang, A. Shah, P. H. Torr, and S.-N. Lim, "Open vocabulary semantic segmentation with patch aligned contrastive learning," in *CVPR*, 2023.
[210] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.
[211] J. Cha, J. Mun, and B. Roh, "Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs," in *CVPR*, 2023.
[212] S. He, H. Ding, and W. Jiang, "Semantic-promoted debiasing and background disambiguation for zero-shot instance segmentation," in *CVPR*, 2023.
[213] J. Qin, J. Wu, P. Yan, M. Li, R. Yuxi, X. Xiao, Y. Wang, R. Wang, S. Wen, X. Pan *et al.*, "Freeseg: Unified, universal and open-vocabulary image segmentation," in *CVPR*, 2023.
[214] X. Chen, S. Li, S.-N. Lim, A. Torralba, and H. Zhao, "Open-vocabulary panoptic segmentation with embedding modulation," *arXiv*, 2023.
[215] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *ICCV*, 2021.
[216] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv*, 2016.
[217] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022.
[218] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *CVPR*, 2021.
[219] X. Liu, B. Tian, Z. Wang, R. Wang, K. Sheng, B. Zhang, H. Zhao, and G. Zhou, "Delving into shape-aware zero-shot semantic segmentation," in *CVPR*, 2023.
[220] G. Shin, W. Xie, and S. Albanie, "Reco: Retrieve and co-segment for zero-shot transfer," *NeurIPS*, 2022.
[221] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020.
[222] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu, "Denseclip: Language-guided dense prediction with context-aware prompting," in *CVPR*, 2022.
[223] R. Burgert, K. Ranasinghe, X. Li, and M. S. Ryoo, "Peekaboo: Text to image diffusion models are zero-shot segmentors," *arXiv*, 2022.
[224] S. Ren, A. Zhang, Y. Zhu, S. Zhang, S. Zheng, M. Li, A. Smola, and X. Sun, "Prompt pre-training with twenty-thousand classes for open-vocabulary visual recognition," *arXiv*, 2023.
[225] Y.-L. Sung, J. Cho, and M. Bansal, "Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks," in *CVPR*, 2022.
[226] R. Zhang, R. Fang, W. Zhang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free clip-adapter for better vision-language modeling," *arXiv*, 2021.
[227] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *CVPR*, 2022.
[228] V. Dumoulin, E. Perez, N. Schucher, F. Strub, H. d. Vries, A. Courville, and Y. Bengio, "Feature-wise transformations," *Distill*, 2018.
[229] Y. Li, H. Wang, Y. Duan, and X. Li, "Clip surgery for better explainability with enhancement in open-vocabulary tasks," *arXiv*, 2023.
[230] X. Wang, S. Li, K. Kallidromitis, Y. Kato, K. Kozuka, and T. Darrell, "Hierarchical open-vocabulary universal image segmentation," *arXiv*, 2023.
[231] Y. Lu, C. Xu, X. Wei, X. Xie, M. Tomizuka, K. Keutzer, and S. Zhang, "Open-vocabulary point-cloud object detection without 3d annotation," in *CVPR*, 2023.
[232] B. Liu, S. Deng, Q. Dong, and Z. Hu, "Language-level semantics conditioned 3d point cloud segmentation," *arXiv*, 2021.
[233] R. Ding, J. Yang, C. Xue, W. Zhang, S. Bai, and X. Qi, "Pla: Language-driven open-vocabulary 3d scene understanding," in *CVPR*, 2023.
[234] J. Yang, R. Ding, Z. Wang, and X. Qi, "Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding," *arXiv*, 2023.
[235] A. Takmaz, E. Fedele, R. W. Sumner, M. Pollefeys, F. Tombari, and F. Engelmann, "Openmask3d: Open-vocabulary 3d instance segmentation," 2023.
[236] H. Wang, S. Wang, C. Yan, X. Jiang, X. Tang, Y. Hu, W. Xie, and E. Gavves, "Towards open-vocabulary video instance segmentation," *arXiv*, 2023.
[237] K. Gao, L. Chen, H. Zhang, J. Xiao, and Q. Sun, "Compositional prompt tuning with motion cues for open-vocabulary video relation detection," in *ICLR*, 2023.
[238] L. Li, J. Xiao, G. Chen, J. Shao, Y. Zhuang, and L. Chen, "Zero-shot visual relation detection via composite visual cues from large language models," *arXiv*, 2023.
[239] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv*, 2023.
[240] Y. Zang, W. Li, J. Han, K. Zhou, and C. C. Loy, "Contextual object detection with multimodal large language models," *arXiv*, 2023.
[241] R. Pi, J. Gao, S. Diao, R. Pan, H. Dong, J. Zhang, L. Yao, J. Han, H. Xu, and L. K. T. Zhang, "Detgpt: Detect what you need via reasoning," *arXiv*, 2023.
[242] X. Gu, Y. Cui, J. Huang, A. Rashwan, X. Yang, X. Zhou, G. Ghiasi, W. Kuo, H. Chen, L.-C. Chen *et al.*, "Dataseg: Taming a universal multi-dataset multi-task segmentation model," *arXiv*, 2023.