Learning to Detect and Segment for Open Vocabulary Object Detection

Tao Wang
Sichuan University
twangnh@gmail.com

Abstract

Open vocabulary object detection has been greatly advanced by the recent development of vision-language pretrained model, which helps recognize novel objects with only semantic categories. The prior works mainly focus on knowledge transferring to the object proposal classification and employ class-agnostic box and mask prediction. In this work, we propose CondHead*, a principled dynamic network design to better generalize the box regression and mask segmentation for open vocabulary setting. The core idea is to conditionally parameterize the network heads on semantic embedding and thus the model is guided with class-specific knowledge to better detect novel categories. Specifically, CondHead is composed of two streams of network heads, the dynamically aggregated head and dynamically generated head. The former is instantiated with a set of static heads that are conditionally aggregated, these heads are optimized as experts and are expected to learn sophisticated prediction. The latter is instantiated with dynamically generated parameters and encodes general class-specific information. With such a conditional design, the detection model is bridged by the semantic embedding to offer strongly generalizable class-wise box and mask prediction. Our method brings significant improvement to the state-of-the-art open vocabulary object detection methods with very minor overhead,* e.g.*, it surpasses a RegionClip model by 3.0 detection AP on novel categories, with only 1.1% more computation.*

1. Introduction

Given the semantic object categories of interest, object detection aims at localizing each object instance from the input images. The prior research efforts mainly focus on the close-set setting, where the images containing the interested object categories are annotated and used to train a detector. The obtained detector only recognizes object categories that are annotated in the training set. In such a setting, more data needs to be collected and annotated if novel category[1] needs to be detected. However, data collection and annotation are very costly for object detection, which raises a significant challenge for traditional object detection methods.

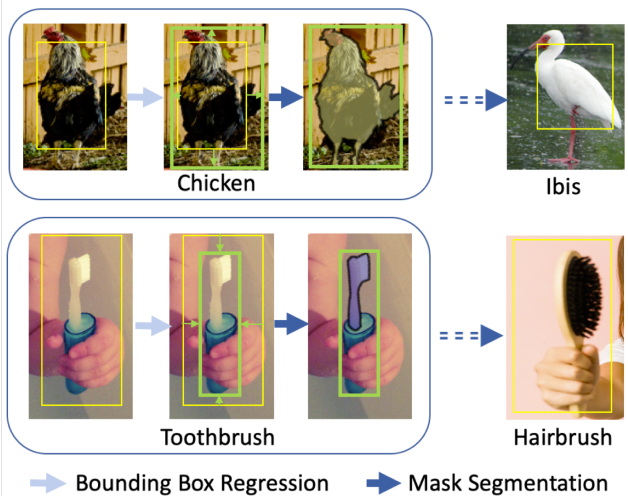

Figure 1. Illustration of our main intuition. Given the object proposals, the bounding box regression and mask segmentation learned from some object categories could generalize to the target category. For example, the knowledge learned from a chicken could help detect and segment the long thin feet and the small head of an ibis (upper row). Similarly for the hairbrush, the knowledge learned from the toothbrush could better handle the extreme aspect ratio and occlusion from the hand (lower row).

To address the challenge, the open vocabulary object detection methods are widely explored recently, these methods [1, 7, 9, 11, 14, 21, 23, 30–32] aim at generalizing object detection on novel categories by only training on a set of labeled categories. The core of these methods is transferring the strong image-text aligned features [16, 22] to classify objects of novel categories. To achieve bounding box regression and mask segmentation on novel categories, they simply employ the class-agnostic network heads. Although class agnostic heads can be readily applied to novel target object categories, they offer limited capacity to learn category-specific knowledge like object

[1] we treat *category* and *class* interchangeably in this paper

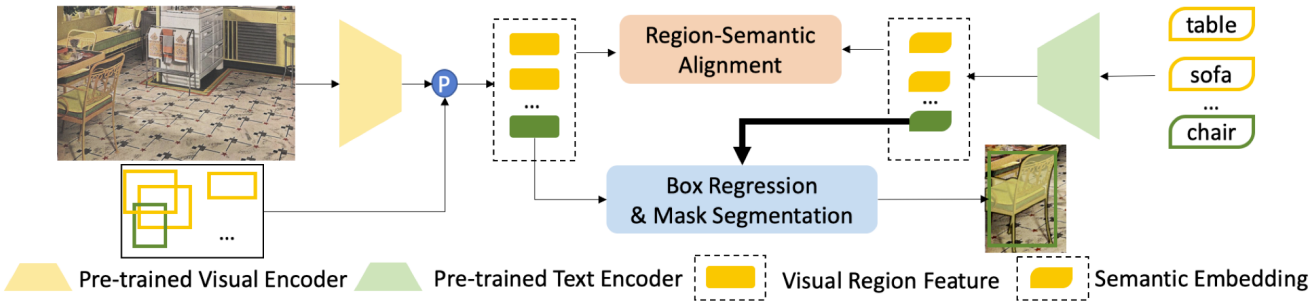

Figure 2. **Overview of CondHead.** To detect objects of novel categories, we aim at *conditionally parameterizing* the bounding box regression and mask segmentation based on the semantic embedding, which is strongly correlated with the visual feature and provides effective class-specific cues to refine the box and predict the mask.

shape, and thus provide sub-optimal performance. On the other hand, training class-wise heads is not feasible as we do not have bounding box and mask annotation for the target categories. As shown in Figure 1, our intuition is that the class-wise knowledge could naturally generalize across object categories, and may be leveraged to achieve much higher quality box regression and mask segmentation on the target categories, in a category-specific manner. However, we find a brute force way of training class-wise heads on the base categories and manually gathering the class-specific prediction with closet semantic similarity during inference provides limited gain. The reason is that there still remains gap between the base and target categories, such as appearance and context.

Motivated by the strong semantic-visual aligned representation [7, 11, 31, 32] in open vocabulary object detection, we propose to exploit the semantic embedding as a conditional prior to parameterize class-wise bounding box regression and mask segmentation. Such conditioning is learned on base categories and easily generalizes to novel categories with the semantic embedding. Our method, named *CondHead*, is based on dynamic parameter generation of neural networks [5, 6, 15, 17, 29]. To achieve strong efficiency, it exploits both large complex network head for their representative power and small light network head for their efficiency. The complex head is employed by conditional weight aggregation over a set of static heads. The light head is employed by dynamically basing its parameters on the semantic embedding. The final prediction is obtained by combining the predictions from the two stream results. Through optimization on the base categories, the set of static heads are expected to learn sophisticated expert knowledge to cope with complex shapes and appearance, the dynamic head is endowed with general class-specific knowledge such as color and context.

Our CondHead is flexible regarding the choice of semantic-visual representation. The stronger quality of the aligned representation is expected to bring higher performance, as the conditional knowledge from the semantic embedding could generalize better to the target visual features. This is demonstrated by the clear grow of improvement over three baselines with increasing quality of pre-trained semantic-visual encoder networks, OVR-CNN [31], ViLD [11] and RegionCLIP [32], on both COCO [19] and LVIS [12] datasets. Remarkably, CondHead brings an average 2.8 improvement w.r.t both box and mask AP for the strong RegionCLIP baseline, with only about 1% more computation. We also demonstrate intriguing qualitative results, showing how the semantic conditioning positively affects the regression and segmentation tasks.

Our contributions are three-fold. 1) To the best of our knowledge, we are the first to leverage semantic-visual aligned representation for open vocabulary box regression and mask segmentation. 2) We design a differentiable semantic-conditioned head design to efficiently bridge the strong category-specific prediction learned on base categories to the target novel categories. 3) We extensively validate the proposed method on various benchmark datasets and conduct thorough ablation and analysis to understand how the semantic conditioning helps detect and segment the novel categories.

2. Related Work

Detecting Objects with Zero-shot and Open Vocabulary Setting Despite the remarkable success of fully-supervised object detection [2, 3, 10, 20, 24–26, 36], the instance-level annotation for novel object categories is costly. Recent years, open vocabulary object detection emerges as an effective solution, which aims at detecting novel categories without corresponding annotation. The representative methods employ region-based alignment, *e.g.*, OVR-CNN [31], ViLD [11], RegionCLIP [32] and DetPro [7]. They are based on the two-stage object detection framework [26] and aim to match the object region feature with the generalizable semantic category embedding. They mainly differ in the learning of semantic-visual representa-

tion. Specifically, OVR-CNN [31] optimizes the semantic-visual grounding objective based multimodal transformer and applies the learned vision to language projection to facilitate novel category recognition. ViLD [11] transfers the stronger visual-semantic representation from large-scale vision-language pre-training [22]. Based on ViLD, DetPro [7] introduces a learnable prompt mechanism [33] to enhance the semantic representation and achieve higher performance. RegionCLIP [32] instead focuses on more effective region-wise pre-training to obtain more discriminative representation. In addition to pursuing more stronger region feature alignment, concurrent works [9, 14, 21, 30] explore other aspects of open vocabulary detection. For example, Gao *et al.* [9] and Huynh *et al.* [14] develop pseudo-labeling approach to generate pseudo ground-truth labels for object detection and instance segmentation. Instead of labeling perspective, some works focus on transformer architecture design for open vocabulary detection. OV-DETR [30] proposes a binary matching objective to enable query matching with novel objects on the end-to-end DETR framework DETR [3]. Minderer *et al.* [21] directly transfers pre-trained vision transformers to simplify open vocabulary detection.

The prior works generally focus on the recognition of open vocabulary objects, while the localization or segmentation are typically conducted in a class-agnostic fashion. In this work, we explore this orthogonal perspective and aim at developing effective methods for more accurate box regression and mask segmentation.

Dynamic Network Design Traditional neural network architectures employ static weights, *i.e.*, the weights are independent of input signal and are fixed after training. Such static neural networks are limited at flexibility, in dealing with diverse input signals like images. Some previous works explore the idea of dynamic instantiation of networks to improve the flexibility. For example, Jia *et al.* [17] proposes to actively predict the convolution filters based on the input. In addition to explicit network parameter generation, Jaderberg *et al.* [15] introduces a dynamic parametric transformation (STN) to adaptively transform the image feature map, recovering affine distortions in the input and thus facilitating recognition. STN [15] operates on global image space, the Deformable Convolution [6, 35] instead works in local scale. It learns spatial offsets to adjust the sampling locations of convolution kernels. The dynamic designs like STN [15] and Deformable Convolution [6, 35] introduce explicit control over the networks. Some recent works introduce implicit dynamics to achieve strong network capacity. Notably, [5, 29] develop conditionally parameterized convolution filters to increase the network capacity. Specifically, the convolution kernel is instantiated by dynamically aggregating a set of base filters.

Our work is related to the above dynamic network architecture designs, in that we also aim at achieving input-conditioned processing, to better adapt to the input data. Unlike prior works, we focus on the object instance-level design of dynamic processing. Moreover, instead of operating on the same visual signal, we instead introduce semantic signal for the dynamic network design.

3. Methodology

3.1. Problem Definition

Given the bounding box annotation on a set of base object categories $\mathbb{F} = \{C_1, C_2, ...C_N\}$, open vocabulary object detection aims at training on the base data and generalizing to any open set of target object categories $\mathbb{F}^* = \{C_1, C_2, ...C_M\}$. Open vocabulary instance segmentation is also expected if the mask annotation is given.

The state-of-the-art two-stage methods [7, 11, 31, 32] can be simplified as a region-semantic alignment framework. The model is initialized with an image encoder ($\mathcal{T}_i$) and a language encoder ($\mathcal{T}_l$), which extract image feature and semantic embedding: $\mathbf{F} = \mathcal{T}_i(I)$, $\mathbf{s}_n = \mathcal{T}_l(C_n)$, here I denotes the input image. These encoders [16, 22] are typically pre-trained on large-scale image-caption [4, 22, 28] data and thus extract strongly correlated visual-semantic representations. With region proposal network [26] trained on the base categories to extract object proposals, the open-vocabulary detection is trained in a region-wise manner.

Concretely, given an object proposal bounding box $\mathbf{p} = (x_1, y_1, x_2, y_2)$, feature pooling is conducted on the image feature map $\mathbf{F}$ to obtain the region feature $\mathbf{f}$. The object proposal box is then matched with the ground-truth object instance set $\aleph$ w.r.t IOU (Intersection Over Union) metric:

$$(\mathbf{b}, c) = \underset{(\mathbf{b}', c') \in \aleph}{\arg\max} \mathrm{IOU}(\mathbf{p}, \mathbf{b}') \tag{1}$$

where $\mathbf{b} = (x_1^*, y_1^*, x_2^*, y_2^*)$ means the obtained ground-truth bounding box and c is corresponding object category. Then, optimization is conducted to further minimize the alignment between ground-truth semantic embedding $\mathbf{s}_c$ and region feature $\mathbf{f}$, *e.g.*, through similarity-based cross entropy loss [11].

To generalize the bounding box regression, prior works employ class-agnostic network head $\mathcal{B}$ with parameter θ and optimize it on the base category data:

$$\min_{\theta} \mathcal{L}_{\mathcal{B}}(\mathcal{B}_\theta, \mathcal{X}) = \mathcal{L}_{\mathcal{B}}(\mathcal{B}_\theta(\boldsymbol{p}, \boldsymbol{f}), \boldsymbol{b}), \tag{2}$$

where $\mathcal{X} = (\mathbf{p}, \mathbf{f}, \mathbf{b})$ denotes the tuple data of proposal, region feature and bounding box. $\mathcal{L}_{\mathcal{B}}$ is typically defined as **L1** or **L2** error between the regressed bounding box and ground-truth box.

To learn instance segmentation, the mask region feature $\mathbf{v}$ is pooled with the refined bounding box from the box

head, Then the class-agnostic mask segmentation network head $\mathcal{M}$ with parameter ϑ is trained by:

$$\min_{\vartheta} \mathcal{L}_{\mathcal{M}}(\mathcal{M}_{\vartheta}, \mathcal{Y}) = \mathcal{L}_{\mathcal{M}}(\mathcal{M}_{\vartheta}(\boldsymbol{v}), \boldsymbol{m}), \tag{3}$$

where **m** denotes the corresponding ground-truth mask segmentation and $\mathcal{Y} = (\mathbf{v}, \mathbf{m})$ denotes the paired region feature and mask. $\mathcal{L}_{\mathcal{M}}$ is typically defined as pixel-wise segmentation error between the predicted mask and ground-truth.

Although the class-agnostic bounding box regression and mask segmentation heads readily generalize to the target novel object categories, they offer sub-optimal results due to the less representative capability. Since the semantic embeddings are strongly correlated with the visual features, we propose to achieve *generalizable class-wise prediction* by conditioning the network parameters on the base category semantic embeddings during training:

$$\min_{\alpha} \mathcal{L}_{\mathcal{B}}(\mathcal{B}_{\theta(\mathbf{s}_c)}, \mathcal{X}) \tag{4}$$

$$\min_{\beta} \mathcal{L}_{\mathcal{M}}(\mathcal{M}_{\vartheta(\mathbf{s}_c)}, \mathcal{Y}) \tag{5}$$

Where α and β are parameters of the conditioning function $\theta(\cdot)$ and $\vartheta(\cdot)$. Through this approach, the model learns to generalize regression and segmentation to novel categories in a class-wise manner, by predicting corresponding parameters $\theta(\mathbf{s}_m)$ and $\vartheta(\mathbf{s}_m)$ during inference.

3.2. CondHead Formulation

Our proposed CondHead aims to bridge the gap between base and target categories, with the semantic conditioning mechanism. We instantiate the conditioning function with neural networks to achieve flexible optimization and inference. To construct such conditioning framework, we observe the challenges in network capacity: 1) More complex head architecture (*e.g.*, head with several hidden layers) offers stronger representative ability, but it requires large number of parameters, which are not efficient to generate and hard to optimize. For example, a hidden fully layer with input and output dimension of 256 involves more than 60k scalar weights. 2) Less complex head architecture (*e.g.*, head with a single layer) is easier to optimize and can be efficiently generated, but provides limited capacity.

To address the challenge, we design a *dual conditioning framework*, to leverage both the complex heads and light heads. The framework is composed of dynamically aggregated head and dynamically generated head.

Dynamically Aggregated Head To leverage the large network capacity of complex heads, we propose to generate dynamic weights to aggregate a set of complex heads $\{\mathcal{B}_1, \mathcal{B}_2, ..., \mathcal{B}_H\}$, with associated parameters $\{\theta_1, \theta_2, ..., \theta_H\}$. These heads act as experts that are good at refining bounding box and predicting mask segmentation for objects categories with certain shapes or appearance. Specifically, the aggregation weights is generated by:

$$\mathbf{w} = \mathcal{A}_{\phi}(\mathbf{s}_c) \tag{6}$$

where $\mathbf{w} \in \mathbb{R}^H$, and $\mathcal{A}$ is a small neural network with parameter ϕ. To regularize the weight aggregation space as a convex hull for optimization, the weight **w** is then normalized with Softmax function:

$$w_h = \frac{e^{w_h}}{\sum_{h'=1}^{H} e^{w_{h'}}} \tag{7}$$

which ensures the weights sums to 1 and each element is larger than 0. With the normalized aggregation weight, the expert heads are then combined to a single head $\hat{\mathcal{B}}$ with parameter $\hat{\theta}$:

$$\hat{\theta} = \sum_{h=1}^{H} w_h * \theta_h \tag{8}$$

Similarly, for a set of mask heads $\{\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_H\}$, with associated parameters $\{\vartheta_1, \vartheta_2, ..., \vartheta_H\}$, the weight generation and aggregation can be conducted to obtain a single mask head $\hat{\mathcal{M}}$ with parameter $\hat{\vartheta}$. When the dynamic aggregation is applied to both box regression and mask segmentation, separate weight generation networks are used, as box regression and mask segmentation may not share the same attention across the expert models.

Dynamically Generated Head To introduce stronger conditioning on the semantic embedding, we propose to directly generate the network parameter. With bounding box regression head as an example, the parameter is first generated as:

$$\dot{\theta} = \mathcal{D}_{\varphi}(\mathbf{s}_c) \tag{9}$$

the parameter $\dot{\theta}$ is used to instantiate a new box head $\dot{\mathcal{B}}$. Since $\dot{\theta}$ is directly generated from the semantic embedding, it easily encodes general class-specific information like aspect ratio and color. Similarly, a dynamic mask segmentation head $\dot{\mathcal{M}}$ can be obtained with generated parameter $\dot{\vartheta}$.

Combining the Dynamic Predictions We then combine the above dynamically aggregated heads $\hat{\mathcal{B}}$, $\hat{\mathcal{M}}$ and dynamically generated heads $\dot{\mathcal{B}}$, $\dot{\mathcal{M}}$ to obtain the final result. We consider simple weighted averaging on their prediction:

$$\overline{\mathcal{B}}_{\theta(\mathbf{s}_c)}(\cdot) = \lambda * \hat{\mathcal{B}}(\cdot) + (1 - \lambda) * \dot{\mathcal{B}}(\cdot) \tag{10}$$

$$\overline{\mathcal{M}}_{\vartheta(\mathbf{s}_c)}(\cdot) = \mu * \hat{\mathcal{M}}(\cdot) + (1 - \mu) * \dot{\mathcal{M}}(\cdot) \tag{11}$$

$\overline{\mathcal{B}}$ and $\overline{\mathcal{M}}$ are for the box regression and mask segmentation head respectively. λ and μ are hyper-parameters.

After optimization on the base category data with equation 4 and equation 5, $\overline{\mathcal{B}}$ and $\overline{\mathcal{M}}$ are employed during inference on the target categories by conditioning on the target semantic embedding as $\overline{\mathcal{B}}_{\theta(\mathbf{s}_m)}(\cdot)$ and $\overline{\mathcal{M}}_{\vartheta(\mathbf{s}_m)}(\cdot)$

Optimizing the Expert Heads with Temperature Annealing During the early optimization stage, the expert heads within the dynamically aggregated head all require training signals to initiate the learning of basic regression and segmentation capability. While in the later optimization stage, the experts are expected to provide specialized capability in dealing with different objects. We thus facilitate the optimization of the expert heads by a temperature annealing Softmax strategy. Concretely, the normalization of aggregation weights is performed with temperature τ

$$w_h = \frac{e^{w_h/\tau}}{\sum_{h'=1}^{H} e^{w_{h'}/\tau}} \tag{12}$$

During the early training stages, the temperature is set as a large value and with the progress of training to provide nearly uniform gradients for all the expert heads, the temperature is gradually annealed to a smaller value to achieve the desired specialized learning.

3.3. Relation to Close Works

Conditionally Parameterized Convolution CondHead is closely related to Dynamic Convolution [5] and CondConv [29]. They differ from CondHead in motivation and application. They aim to improve the representation capacity of convolution kernels through conditional parameterization, the conditioning is performed on the visual features of each network layer. Moreover, they are applied to the image recognition task. While CondHead aims to bridge the perception gap between the base training object categories and target novel object categories by conditioning on the strongly visual-aligned semantic embedding. CondHead is applied to the more challenging bounding box regression and mask segmentation task, which is not explored before.

Partially Supervised Instance Segmentation Recent works [13, 18] explore partially supervised instance segmentation, where the base categories have both box and mask annotation while the novel categories have only box annotation. Mask^X R-CNN [13] propose a parameterized transformation function to transfer the box regression weights to mask segmentation weights. Shapemask [18] explores strong shape priors to achieve better class-agnostic object segmentation, which generalizes better to novel categories without mask annotation. Our focused open vocabulary object detection and instance segmentation are much more challenging than the partially supervised instance segmentation, which is based on good detection quality and strong object region representation. Mask^X R-CNN [13] cannot be applied here as box annotation is not available. Shapemask [18] could be applied here, we compare to it for instance segmentation and evaluate an augmented version of CondHead with Shapemask.

Head	$\mathcal{A}$	$\mathcal{D}$	$\{\mathcal{B}_h\}$	$\{\mathcal{M}_h\}$	$\dot{\mathcal{B}}$	$\dot{\mathcal{M}}$
Architecture	2fc	2fc	2fc	3conv	1fc	1conv

Table 1. Architectural instantiate of CondHead. fc and conv denote fully connected and convolutional networks, respectively, with a hidden dimension of 256. The digit means the number of layers.

4. Experiments

We study four questions in experiments. 1) Is CondHead able to improve the performance of open vocabulary object detection? is it efficient? 2) How does the quality of semantic-visual representation affects the performance? 3) How does CondHead improves the box regression and mask segmentation? 4) How does each component and hyper-parameter take effect? Throughout the experiments, unless otherwise stated, we adopt the architectural instantiation shown in Table 1.

4.1. Datasets and Setup

We adopt object detection benchmark datasets COCO [19] and LVIS [12] to evaluate our method. Following prior works [11, 31], we manually split the datasets into base and target categories[2]. We also evaluate the generalization through inference on two other common object benchmarks, PASCAL VOC [8] and Objects365 [27].

Setup Since the prior works mainly focus on the classification of novel object categories, we evaluate CondHead by adopting the prior state-of-the-art open vocabulary methods and treating them as baselines. We employ the following methods with increasing quality of semantic-visual alignment:

- **OVR-CNN** [31] proposes to pre-train the multi-modal encoder with image-caption pair data. The pre-trained rich visual-semantic representation is then transferred to recognize the objects with open vocabulary.
- **ViLD** [11] achieves much higher performance than OVR-CNN by transferring the stronger visual-semantic representation of CLIP [22], which was pre-trained on a much larger scale of image-caption data.
- **RegionCLIP** [32] further improves the alignment between the semantic and visual embeddings by conduct-

[2] we use the pre-processed data from [31] and [32].

Method	Object Detection			Instance Segmentation		
	Novel	Base	All	Novel	Base	All
DELO [34] (CVPR20)	3.41	13.8	13.0	-	-	-
PL [23] (AAAI20)	4.12	35.9	27.9	-	-	-
OVR-CNN [31] (CVPR21)	22.8	46.0	39.9	-	-	-
OVR-CNN*	22.6	45.9	39.8	20.1	42.3	36.5
CondHead (Ours)	**24.0** (+1.4)	**46.5** (+0.6)	**40.6** (+0.8)	**21.4** (+1.3)	**42.9** (+0.6)	**37.3** (+0.8)
ViLD [16] (ICLR22)	27.6	59.5	51.3	-	-	-
ViLD*	27.7	59.7	51.4	24.1	56.7	48.2
CondHead (Ours)	**29.8** (+2.1)	**60.8** (+1.1)	**52.7** (+1.3)	**25.9** (+1.8)	**57.9** (+1.2)	**49.5** (+1.3)
RegionCLIP [32] (CVPR22)	31.4	57.1	50.4	-	-	-
RegionCLIP*	31.3	56.5	49.9	27.5	54.1	47.1
CondHead (Ours)	**33.7** (+2.4)	**58.0** (+1.5)	**51.7** (+1.8)	**29.7** (+2.2)	**55.8** (+1.7)	**49.0** (+1.9)

Table 2. **Open vocabulary object detection results on COCO.** * denotes our re-evaluation by adding the class-agnostic mask head.

Method	Backbone	Object Detection				Instance Segmentation			
		AP_r	AP_c	AP_f	AP	AP_r	AP_c	AP_f	AP
Supervised	RN50-FPN	3.3	22.5	34.5	23.3	4.2	22.8	32.7	21.8
Supervised +RFS [12]		11.6	23.5	32.5	24.3	12.8	23.3	31.0	23.1
ViLD [11]	RN50-FPN	16.7	26.5	34.2	27.8	16.6	24.6	30.3	25.5
ViLD*		16.6	27.0	33.0	27.5	16.9	25.0	29.1	25.2
CondHead (Ours)		**18.8** (+2.2)	**28.3** (+1.3)	**33.7** (+0.7)	**28.8** (+1.3)	**19.1** (+2.2)	**26.2** (+1.1)	**29.9** (+0.8)	**26.4** (+1.2)
RegionCLIP [32]	RN50-C4	17.1	27.4	34.0	28.2	-	-	-	-
RegionCLIP*		17.0	27.2	34.3	28.2	17.4	26.0	31.6	26.7
CondHead (Ours)		**19.9** (+2.9)	**28.6** (+1.4)	**35.2** (+0.9)	**29.7** (+1.5)	**20.0** (+2.6)	**27.3** (+1.3)	**32.2** (+0.6)	**27.9** (+1.2)
ViLD [11]	RN152-FPN	19.8	27.1	34.5	28.7	-	-	-	-
ViLD*		20.0	27.0	34.0	28.5	19.8	25.1	32.1	26.9
CondHead (Ours)		**21.9** (+1.9)	**28.4** (+1.4)	**34.6** (+0.6)	**29.7** (+1.2)	**21.6** (+1.8)	**26.2** (+1.1)	**33.0** (+0.9)	**28.1** (+1.2)
RegionCLIP [32]	RN50x4-C4	22.0	32.1	36.9	32.3	-	-	-	-
RegionCLIP*		22.1	31.8	37.0	32.2	21.8	30.2	35.1	30.7
CondHead (Ours)		**25.1** (+3.0)	**33.4** (+1.6)	**37.8** (+0.8)	**33.7** (+1.5)	**24.4** (+2.6)	**31.6** (+1.4)	**35.9** (+0.8)	**32.0** (+1.3)

Table 3. **Open vocabulary object detection result on LVIS.** * denotes our re-evaluation by adding the mask head. Supervised and Supervised + RFS are baselines that have access to the training data of target categories. RFS means the repeat factor sampling method [12].

ing region-wise pre-training. The improved representation further improves the recognition of novel objects and thus brings significant performance gain.

4.2. Results

Results on COCO We re-evaluate the baseline methods by adding class-agnostic mask segmentation heads. Then we replace the bounding box regression and mask segmentation heads with the proposed CondHead and re-run the experiments. As shown in Table 2, our method surpasses all baseline methods, especially on the novel object category set. We observe the improvements grow with the three baselines, *e.g.*, the object detection AP improvements are 1.4 with OVR-CNN, 2.1 with ViLD, and 2.4 with RegionCLIP. A similar observation holds for the instance mask segmentation task. This phenomenon verifies that CondHead can leverage semantic embeddings that better align with visual representations to achieve stronger bounding box regression and mask segmentation ability.

Results on LVIS Following the evaluation on COCO, We then evaluate CondHead on LVIS dataset. As shown in Table 3, our method brings about 2.0-3.0 absolute improvement on the target novel object categories (*i.e.*, AP_r), for both bounding box detection and instance segmentation. The improvement holds for stronger backbone networks. We also observe a similar growth of improvement as COCO, *i.e.*, the RegionCLIP method gains more improvement with CondHead than that of ViLD. This is likely because RegionCLIP provides stronger semantic-visual alignment with its region-wise pre-training, and thus CondHead better generalizes to the target categories.

Cross Dataset Evaluation We further evaluate CondHead

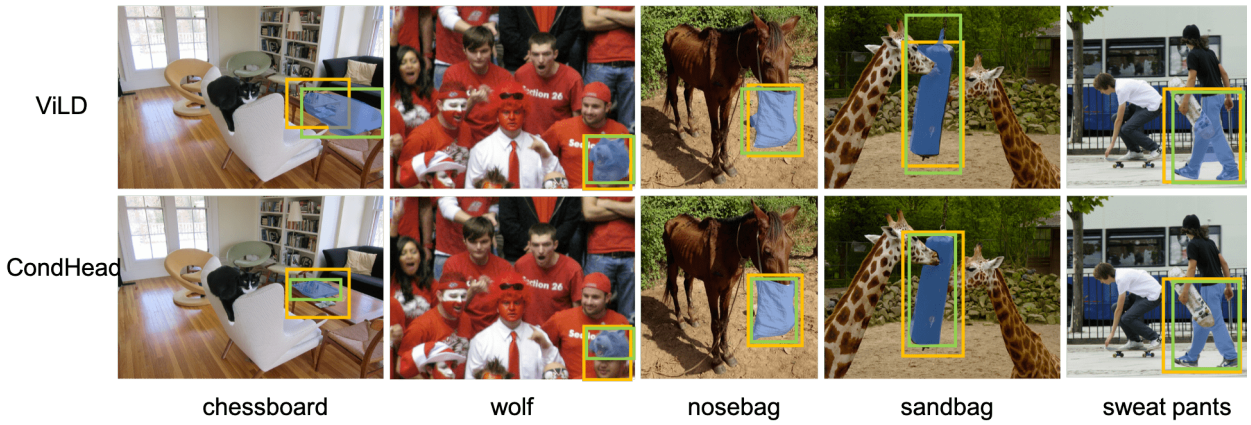

Figure 3. Qualitative comparison with baseline ViLD [11]. The bounding box regression and mask segmentation results are overlaid on the images (Yellow: Proposals. Green: Regressed bounding box. Blue: segmentation mask). Best viewed with zoom-in.

	PASCAL VOC		COCO			Objects365		
Method	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Supervised	78.5	49.0	46.5	67.6	50.9	25.6	38.6	28.0
ViLD [11]	72.2	56.7	36.6	55.6	39.8	11.8	18.2	12.6
CondHead (Ours)	**74.6** (+2.4)	**58.5** (+1.8)	**39.1** (+2.5)	**59.1** (+3.5)	**42.2** (+2.4)	**13.2** (+1.4)	**20.4** (+2.2)	**14.2** (+1.6)

Table 4. Cross-dataset evaluation results. The model is trained on LVIS and directly evaluated on the target datasets.

with the cross dataset scenarios, where the open vocabulary object detector is expected to generalize on datasets that are different from the one used for training. We simply replace the vocabulary of a trained model with target dataset vocabulary to evaluate the performance. As shown in Table 4, we observe the improvement of CondHead transfers well. For example, 3.5, 2.4, and 2.2 absolute improvements in AP_50 for COCO, PASCAL VOC, and Objects365 respectively.

Qualitative Results To demonstrate how *CondHead* improves bounding box regression and mask segmentation, we visualize the example qualitative results and compare that with the baseline method. As shown in Figure 3, CondHead is better at refining the proposal box and segmenting the target objects. For example, the baseline ViLD wrongly regresses the target chessboard to the table, while CondHead correctly predicts the chessboard (Figure 3 first column). In addition, CondHead better segments the challenging sweat pants while ViLD predicts inferior segmentation mask (Figure 3 last column).

4.3. Analysis

The Effects of Language Descriptions We examine how CondHead is affected by language descriptions. As shown in Figure 4, when the input text is for a wrong category label, the box regression deteriorates, *e.g.*, the proposal for sunflower is misguided to predict the carrot. On the other hand, it is interesting that when supplied with manually tuned descriptions, the detection and segmentation quality is improved, *e.g.*, when employing snowboard and motorcycle as descriptions, the snowmobile is predicted better. This intriguing observation suggests CondHead learns strong category-conditioned prediction and a better performance could be achieved by simply tuning the semantic descriptions during inference.

	AP_r^b	AP_r^m
ViLD*	16.6	16.9
ClassWise	**17.1** (+0.5)	**17.3** (+0.4)
CondHead	**18.8** (+2.2)	**19.1** (+2.2)
RegionCLIP*	17.0	17.4
ClassWise	**17.3** (+0.3)	**17.9** (+0.5)
CondHead	**17.9** (+2.9)	**20.0** (+2.6)

Table 5. The results of naive class-wise transfer experiment (discussed in Section 1) on LVIS. (AP_r^b: box AP, AP_r^m: mask AP).

Naive Class-wise Transfer We also establish another intuitive experiment as discussed in Section 1 to demonstrate the effectiveness of the proposed method. Since our intuition (Figure 1) is that the class-specific knowledge could be shared across categories, we train the class-wise heads on the base categories and then perform inference on the target categories in a semantically guided way. This is achieved by gathering the class-specific predictions with semantic similarity to the base categories. We use cosine similarity on the semantic embedding and select the category with the high-

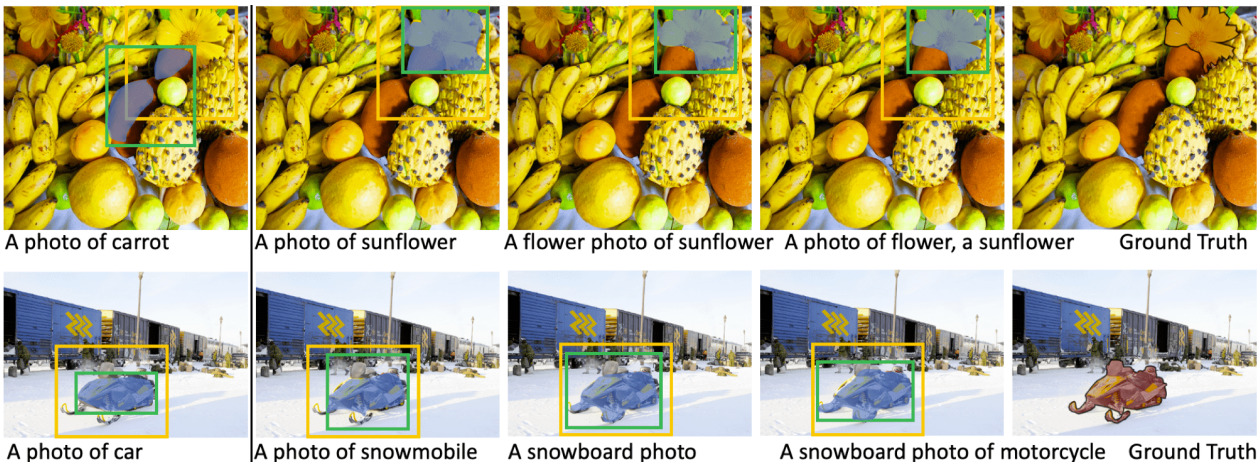

Figure 4. Effect of tuning language descriptions. We select some intriguing examples for which tuning the input language descriptions could deteriorate or help object detection. Yellow: Proposals. Green: Regressed bounding box.

est similarity. As shown in Table 5, we train the ClassWise baseline on both ViLD and RegionCLIP. ClassWise method provides a small improvement to class-agnostic heads compared to CondHead, demonstrating that CondHead better generalizes to the target categories.

	V	SH	CH	CH_S	R	SH	CH	CH_S
AP_r^m	16.9	19.6	19.1	20.4	17.4	20.3	20.0	21.0
Δ	-	2.7	2.2	**3.5**	-	2.9	2.6	**3.6**

Table 6. Results with Shapemask. V and R denote ViLD and RegionCLIP, SH, CH, and CH_S denote Shapemask head, CondHead and CondHead augmented with Shapemask, respectively.

Augmenting CondHead with Shape Priors As discussed in Section 3.3, we compare to Shapemask [18] which introduces explicit class-agnostic shape priors. We then explore an augmented version of CondHead by integrating the Shapemask, the semantic embedding is utilized to estimate the initial shape prior, then the dynamic expert aggregation is conducted on the latter coarse mask estimation and refinement. Please refer to the supplementary for more implementation details. As shown in Table 6, Shapemask brings comparable improvement to CondHead, this is attributed to its strong prior and complex multi-stage mask refinement capability. Further integrating Shapemask into CondHead offers larger gain, *e.g.*, 3.5 and 3.6 AP improvement for the two baselines. This implies CondHead can be improved with explicit shape priors.

Component Analysis As shown in Table 7, we perform an ablation study by applying CondHead on the box and mask heads independently, we find they improve box AP and mask AP independently. We also observe CondHead introduces a very small computation overhead. To understand how the number of expert heads and aggregation

CondBox	CondMask	AP_r^b	AP_r^m	FLOPs
		17.0	17.4	193.5
✓		19.6	17.8	194.1
✓	✓	19.9	20.0	195.8

Table 7. Component analysis. CondBox and CondMask mean applying CondHead on box and mask heads, respectively.

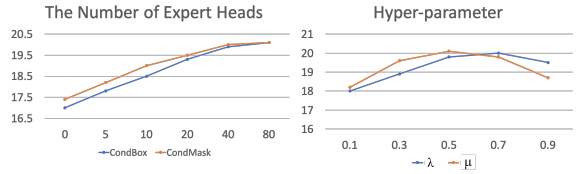

Figure 5. Component analysis. Effect of expert number, λ and μ.

hyper-parameter affects CondHead, a detailed analysis is conducted as shown in Figure 5. We find the performance gain is diminishing beyond 40 expert heads and the hyper-parameters λ and μ work best at between 0.5 and 0.7. More analysis experiments are presented in the supplementary.

5. Conclusion

We introduce a conditionally parameterized neural network design to improve open vocabulary bounding box regression and mask segmentation, which is not explored before. Specifically, we leverage the pre-trained semantic embedding to guide the parameterization of the box and mask heads. The semantic embedding is strongly aligned with visual representation and thus provides effective cues for refining the bounding box and segmenting the objects. Our method named *CondHead* is extensively validated on different datasets and setups. We hope our findings provide insights for future work on open vocabulary detection.

References

[1] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018. 1

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020. 2, 3

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 3

[5] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11030–11039, 2020. 2, 3, 5

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 2, 3

[7] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 1, 2, 3

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 5

[9] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Open vocabulary object detection with pseudo bounding-box labels. 1, 3

[10] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[11] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 1, 2, 3, 5, 6, 7

[12] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 2, 5, 6

[13] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4233–4241, 2018. 5

[14] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7031, 2022. 1, 3

[15] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 2, 3

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1, 3, 6

[17] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Advances in neural information processing systems*, 29, 2016. 2, 3

[18] Weicheng Kuo, Anelia Angelova, Jitendra Malik, and Tsung-Yi Lin. Shapemask: Learning to segment novel objects by refining shape priors. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9207–9216, 2019. 5, 8

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 5

[20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2

[21] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection with vision transformers. *arXiv preprint arXiv:2205.06230*, 2022. 1, 3

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 3, 5

[23] Shafin Rahman, Salman Khan, and Nick Barnes. Improved visual-semantic alignment for zero-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11932–11939, 2020. 1, 6

[24] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[25] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. *arXiv preprint*, 2017. 2

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2, 3

[27] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 5

[28] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 3

[29] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3, 5

[30] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. *arXiv preprint arXiv:2203.11876*, 2022. 1, 3

[31] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 1, 2, 3, 5, 6

[32] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 1, 2, 3, 5, 6

[33] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3

[34] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Don't even look once: Synthesizing features for zero-shot detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11693–11702, 2020. 6

[35] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019. 3

[36] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *The International Conference on Learning Representations*, 2020. 2