P2P: Transforming from Point Supervision to Explicit Visual Prompt for Object Detection and Segmentation

Guangqian Guo, **Dian Shao**, **Chenguang Zhu**, **Sha Meng**, **Xuan Wang**, **Shan Gao** *

Unmanned System Research Institute, Northwestern Polytechnical University

{guogq21, zcg23, mengsha}@mail.nwpu.edu.cn, {shaodian, wangxuan, gaoshan}@nwpu.edu.cn

Abstract

Point-supervised vision tasks, including detection and segmentation, aiming to learn a network that transforms from points to pseudo labels, have attracted much attention in recent years. However, the lack of precise object size and boundary annotations in the point-supervised condition results in a large performance gap between point- and fully-supervised methods. In this paper, we propose a novel iterative learning framework, Point to Prompt (P2P), for point-supervised object detection and segmentation, with the key insight of transforming from point supervision to explicit visual prompt of the foundation model. The P2P is formulated as an iterative refinement process of two stages: Semantic Explicit Prompt Generation (SEPG) and Prompt Guided Spatial Refinement (PGSR). Specifically, SEPG serves as a prompt generator for generating semantic-explicit prompts from point input via a group-based learning strategy. In the PGSR stage, prompts guide the visual foundation model to further refine the object regions, by leveraging the outstanding generalization ability of the foundation model. The two stages are iterated multiple times to improve the quality of predictions progressively. Experimental results on multiple datasets demonstrate that P2P achieves SOTA performance in both detection and segmentation tasks, further narrowing the performance gap with fully-supervised methods. The source code and supplementary material can be found at https://github.com/guangqian-guo/P2P.

1 Introduction

The accurate detection and segmentation of objects in diverse scenarios stand as important tasks, serving as the foundation for various high-dimensional perception domains, like robotic perception, autonomous driving, *etc*.

In recent years, weakly supervised methods have gained widespread attention as an approach to reduce the dependence on annotations in fully supervised methods.

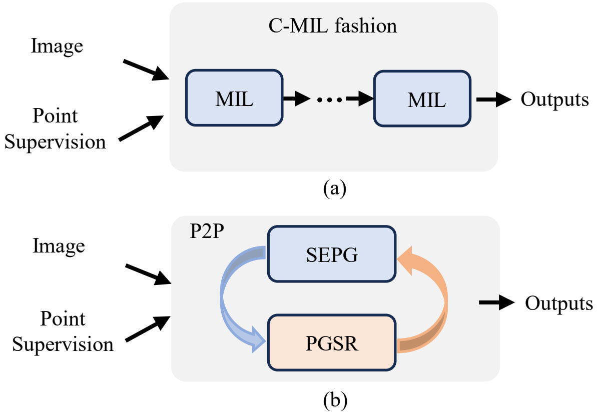

Figure 1: Training paradigms with two different PSOD frameworks: (a) Basic PSOD framework, generally using a cascaded MIL fashion. (b) Our P2P framework.

Typically, the weak supervision includes image-level [Bilen and Vedaldi, 2016; Wan *et al.*, 2018; Xu *et al.*, 2022], point-level [Chen *et al.*, 2022; Liao *et al.*, 2023] and scribble-level [Zhang *et al.*, 2020], *etc*. Among them, object detection and instance segmentation with point-level supervision (shortened to PSOD and PSIS, respectively) have attracted growing attention recently thanks to the low annotation burden and distinctive location information of points.

However, the performance of existing point-supervised methods is still far from satisfactory, about 59% and 55% of that of fully supervised detection and segmentation baselines. This is contrary to the core of weakly supervised methods, *i.e.*, ***to release the annotation burden while still achieving decent performance.*** PSOD is used as an example to illustrate the limitations of point-supervised vision tasks. The current PSOD methods generally follow the paradigm of Multiple Instance Learning (MIL) or Cascaded MIL (C-MIL) fashion. They first use proposal generation methods (*e.g.*, Selective Search [Uijlings *et al.*, 2013], MCG [Arbeláez *et al.*, 2014] or neighbor proposal sampling [Chen *et al.*, 2022]) to construct proposal bags. After that, top-k proposals with high scores are selected from hundreds of independent proposals as the final result by MIL. Due to the lack of object size and edge information, proposal generation is more random and

[*] Corresponding author

low-quality, and the results are limited by inaccurate proposals. Additionally, multiple independent proposals bring high randomness in the selection process, and it is easy to converge to the sub-optimal solution and focus on the discriminative part rather than the entirety of the object.

Thanks to the substantial progress in the visual foundation models, such as Segment Anything Model (SAM) [Kirillov *et al.*, 2023], many downstream tasks have witnessed significant breakthroughs [Wang *et al.*, 2023]. ***We hold the perspective that, rather than directly designing large foundation models, it is more meaningful to leverage them for specific tasks in resource-constrained situations.*** Notably, some efforts have been made to adapt SAM for weakly supervised tasks, *e.g.*, weakly supervised semantic segmentation [Chen *et al.*, 2023b]. However, these studies have employed SAM as a supplementary tool in a simplistic way and have not attempted to explore how to better guide SAM by enhancing the semantic representation capability of prompts. Compared with bounding boxes or masks, point annotations inherently possess limited semantic representation. When points are directly used as prompts for SAM, only 40% of the masks cover more than 70% of foreground pixels, significantly lower than the results (about 80%) obtained when using boxes as prompts. This highlights the crucial importance of semantic-explicit prompts.

In this paper, we propose a novel framework, referred to as **P**oint-to-**P**rompt (P2P) for point-supervised detection and segmentation, by transforming the point supervision into visual prompt learning. We are the first to attempt to switch point-supervised tasks into visual prompt learning based on foundation model. An overview of the contrast between the existing PSOD framework and our framework is presented in Fig. 1. P2P comprises two integral processes: the Semantic-Explicit Prompt Generation (SEPG) stage and the Prompt Guided Spatial Refinement (PGSR) stage. Specifically, the SEPG stage is designed to generate semantic explicit pseudo boxes as prompts under the guidance of semantic confidence. The PGSR stage further refines the target regions covered in the semantic-explicit prompts by leveraging the outstanding generalization ability of the foundation model SAM. It operates through an iterative process, involving multiple iterations between SEPG and PGSR, ultimately resulting in the generation of precise pseudo-labels. Utilizing SAM, our method can output precise pseudo-masks, so it can be applied as both a point-supervised detection and segmentation method that transforms points into accurate pseudo masks and boxes. Experiments on the challenging MS COCO 2017 and PASCAL VOC 2007 datasets are conducted to validate both the detection and the segmentation performance. Given point supervision, P2P further closes the gap with the fully supervised model and achieves 84% and 75% of the performance of fully supervised on the COCO dataset, respectively. Our main contributions are as follows:

- We design a novel Point-to-Prompt (P2P) method for point-supervised object detection and segmentation, which transforms the point supervision into prompting to predict precise pseudo-labels.
- We propose an iterative learning framework using visual foundation model to achieve a semantic-explicit output, including a semantic-explicit prompt generation stage and a prompt guided spatial refinement stage.
- Our P2P method achieves state-of-the-art performance on detection and segmentation, significantly narrows the performance gap with fully supervised methods, and provides new insights for point supervision tasks.

2 Related Work

Weakly Supervised Detection. For image-supervised detection, the difficulty lies in how to mine the location of each instance with only semantic information. Existing methods [Tang *et al.*, 2018; Wan *et al.*, 2018; Gao *et al.*, 2019] generally build image-level proposal bags containing hundreds of proposals, and then mine instance-level supervision through MIL, unsupervised clustering, and contrastive learning. Due to the lack of location information, the performance is still poor for some complex datasets, such as COCO [Lin *et al.*, 2014]. Point-supervised methods benefit from additional point-level annotations, providing coarse location information. [Ren *et al.*, 2020b] designs a unified network compatible with various supervision forms. P2BNet [Chen *et al.*, 2022] specifies that low-quality proposals limit the performance of this task and proposes to generate proposals through a neighbor sampling policy and designs a cascade MIL framework.

Weakly Supervised Instance Segmentation. Weakly supervised instance segmentation is mainly performed by estimating instance-level pseudo masks and refining the estimated masks by training a segmentation model. To obtain the pseudo mask for each instance with only point-level information, previous approaches have either used off-the-shelf proposal methods [Zhou *et al.*, 2019] or generated instance-level localization maps [Kim *et al.*, 2022; Liao *et al.*, 2023] by refining the attention maps of CAM [Zhou *et al.*, 2016] or ViT [Dosovitskiy *et al.*, 2020]. The performance of the current methods is significantly constrained by the quality of the attention map. [Liao *et al.*, 2023] analyzes that only about 30% of the ViT attention maps can cover more than 50% of the foreground objects, greatly limiting the performance.

Prompt engineering and Foundation Models. Prompt engineering refers to the process of designing prompts that enable the foundation model to adapt and generalize to different tasks. Segment Anything Model (SAM) [Kirillov *et al.*, 2023], as a representative prompt-based foundation model, is designed for image segmentation and has brought a new trend in solving other downstream tasks. The research community has been actively engaged in exploring and pushing the capability boundaries of SAM and applying it to various tasks, *e.g.*, Remote Sensing [Chen *et al.*, 2023a], medical image analysis [He *et al.*, 2023], and weakly supervised semantic segmentation [Chen *et al.*, 2023b]. Inspired by these approaches, we apply foundation model to point-supervised detection and segmentation and significantly facilitate the performance of point-supervised tasks.

3 Method

3.1 Overview

Problem. A point annotation p can be represented as $p = (p_x, p_y, c)$, where (p_x, p_y) and c represent point location and

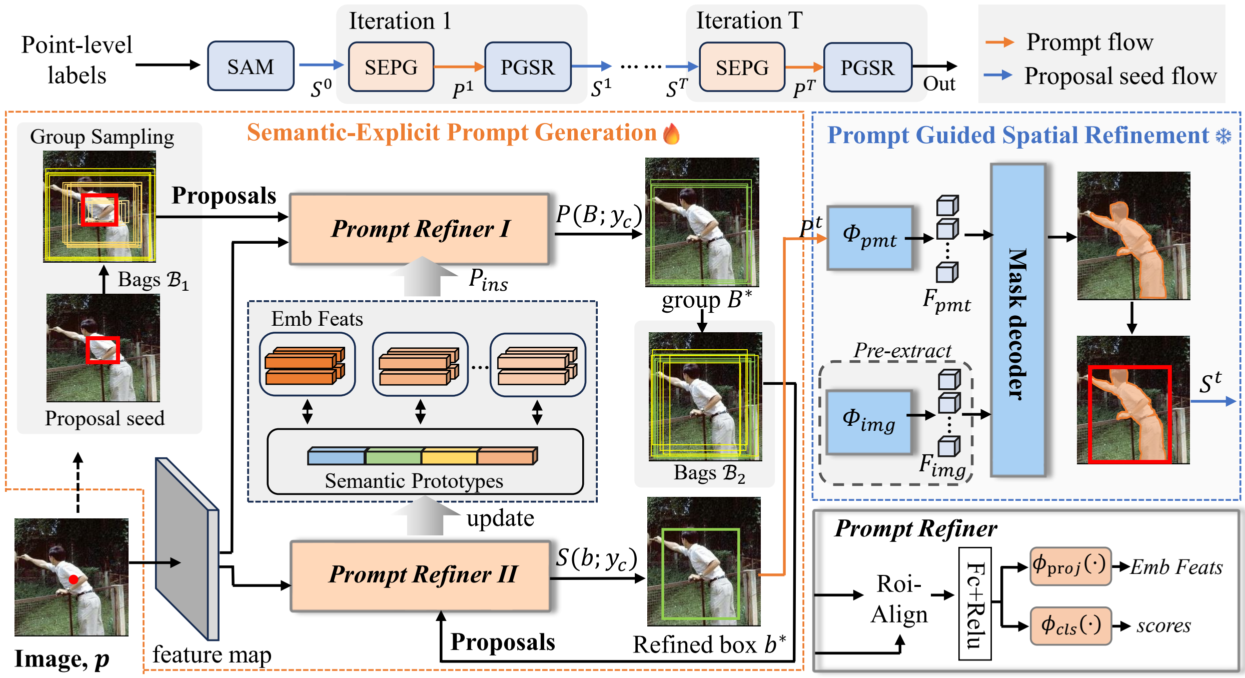

Figure 2: Framework of P2P, which performs SEPG and PGSR iteratively to generate more accurate pseudo labels. Specifically, given an image and point annotation, we first use SAM to generate proposal seeds as the initial input to SEPG. In SEPG, we use a group sampling strategy and two joint confidence-based refiners to get the refined box b^*. Guided by b^*, PGSR further spatially refines the object region to get more accurate masks and boxes under the given semantic and updates the proposal seeds.

object category, respectively. Point-supervised tasks aim to train a point-to-label regressor using point annotations to predict pseudo-labels. Subsequently, a task-related sub-network (*e.g.*, a detector) is retrained in a fully-supervised manner for inference. Thus, the core of this task lies in designing an accurate point-to-label regressor, denoted as $\Phi_{reg}(\cdot)$, which transforms the point annotation into precise pseudo annotations. To design a well-formed regressor, we introduce the P2P framework, proposing to first transform point supervision into explicit visual prompts and then obtain pseudo-labels guided by these prompts.

Framework. We structure the P2P as an iterative refinement process of the Semantic-Explicit Prompt Generation (SEPG) stage and the Prompt Guided Spatial Refinement (PGSR) stage. The two stages respectively serve the roles of "point to prompt" and "prompt to pseudo-mask". In P2P, the point annotation p is first viewed as the prompt of SAM to generate an initial mask for each object. The outer rectangle of the mask is used as the proposal seed S^0. Taking S^0 as input, P2P initiates the first round of iteration. While proposal seed may not be entirely accurate, it can still provide valuable prior information on object size. Then, taking the initial proposal seed S^0 as input, the refined box is generated by two prompt refiners under semantic supervision. Compared to S^0, the refined box covers the main semantic part of the object and can be used as the subsequent prompt, referred to as P^1. After, in PGSR, SAM is used to refine the spatial regions guided by the semantic-explicit prompt P^1 and generates the next round of proposal seed S^1. Improved proposal seed leads to better prompt, and better prompt, in turn, contributes to better proposal seed. The two modules iterate T times, ultimately yielding predicted pseudo-masks and pseudo-boxes through PGSR. The overall pipeline of P2P is depicted in Fig. 2.

3.2 Semantic-explicit Prompt Generation

We design a semantic prompt generator that takes semantic-agnostic seeds as input and produces semantic-explicit prompts. Our approach adopts a *group-then-individual* strategy on two refiners, as illustrated in Fig. 2. Initially, feature maps are extracted by ResNet-50 [He *et al.*, 2016] backbone. In *Prompt Refiner I*, we obtain a semantically accurate proposal group B^*, followed by *Prompt Refiner II*, we further refine the proposal group to obtain refined proposal b^*.

Seeds-based Group Sampling. Previous methods usually use neighbor sampling to build proposal bags that contain hundreds of individual proposals that usually suffer from low quality and lack of good priority. To mitigate that, we introduce a group sampling strategy based on the initial proposal seed S^0. For the first phase, we create the proposal bag $\mathcal{B}_1$ by progressively sampling n proposal groups $\{B_i\}_{i=1}^n$ for each instance based on the proposal seed, *i.e.*, $\mathcal{B}_1 = \{B_i\}_{i=1}^n$. These distinct proposal groups are generated by scaling the proposal seed at various scales. Each proposal group comprises m proposals with strong spatial correlation, denoted as $B_i = \{b_{i,j}\}_{j=1}^m$, where $b_{i,j}$ denotes the jth proposal of the ith group and m signifies the number of proposals in B_i. The motivation behind designing proposal groups is to reduce the solution space by ***selecting proposal group instead of individual proposals***. For the second phase, we construct the proposal bag $\mathcal{B}_2$ by augmenting the group of proposals produced in the first phase. We adopt a "proposal jittering" strategy [Liao *et al.*, 2022] to generate randomly jittered proposals in four orientations.

Proposal-to-Prompt Semantic Lifting. Different from classical MIL frameworks, we adopt a *group-then-individual* strategy, *i.e.*, selecting a group of proposals with strong spatial correlation first and then further refining the proposals according to the group in the second phase. We employ a more stable feature prototype for computing group-based semantic distribution. The refiners in P2P comprise a classification head and an embedding head, which calculate classification scores and feature embeddings, respectively.

In *Prompt Refiner I*, the problem lies in identifying a semantic-accurate ***proposal group***, which determines the direction of model optimization. The basic MIL head or classification head is commonly employed as the refiner, but the inherent instability in its training process easily lead the model towards sub-optimal solutions. To remedy the bias of predicted probabilities, we use a prototype representation to obtain stable group-based semantic distributions of proposals.

A memory buffer is established to keep a set of prototypes $\mathcal{V} = \{\mathcal{V}_c\}_{c=1}^{C}$ for each category, which preserves the semantic-explicit features. These prototypes are updated using the selected high-quality embedding features from *Prompt Refiner II* in each iteration , via the Exponential Moving Average (EMA) algorithm. After that, the ***group-based instance-level probability*** distribution $\mathcal{P}_{ins}(B_i; y_c)$ of each proposal group B_i can be measured by the similarity between the feature embeddings of the proposal groups and their corresponding semantic prototypes.

$$\mathcal{P}_{ins}(B_i; y_c) = \frac{\exp(sim(Z_i, \mathcal{V}_c))}{\sum_i \exp(sim(Z_i, \mathcal{V}_c))}, \tag{1}$$

where Z_i indicates the feature embedding of the proposal group B_i. It is calculated by averaging the feature embeddings of all the proposals in the group, as $Z_i = \frac{1}{|B_i|}\sum_j z_{i,j}$, where $z_{i,j}$ indicates jth proposal in B_i, and $|B_i|$ is the number of proposals. $sim(\cdot, \cdot)$ denotes the cosine similarity metric, utilized to quantify the similarity between the embedding features and the semantic prototypes. Furthermore, we calculate the ***group-based semantic-level probability*** $\mathcal{P}_{sem}(B_i; y_c)$ for each proposal group B_i. It is computed by the score of the proposal group, as

$$\mathcal{P}_{sem}(B_i; y_c) = \frac{\exp\{\frac{1}{|B_i|}\sum_{b_{i,j}\in B_i} O(b_{i,j}; y)\}}{\sum_i \sum_y \exp\{\frac{1}{|B_i|}\sum_{b_{i,j}\in B_i} O(b_{i,j}; y_c)\}}, \tag{2}$$

where $O(b_{i,j}; y_c)$ denotes the score of jth proposal in B_i. Finally, we define a ***group-based joint probability*** distribution that combines the semantic level and the instance level, termed $\mathcal{P}(B_i; y_c) = \mathcal{P}_{ins}(B_i; y_c) * \mathcal{P}_{sem}(B_i; y_c)$, which indicates the semantic probability of proposal group B_i for a given semantic label $y_c \in \{y_1, y_2, ..., y_C\}$.

In the learning procedure, based on the above definition, $\mathcal{P}(B_i; y_c)$ is applied to the refinement process, the corresponding loss function of the first phase (termed $\mathcal{L}_1$) is defined as:

$$\mathcal{L}_1 = -\sum_{c=1}^{C} y_c \text{log} \sum_i \mathcal{P}(B_i; y_c) + (1 - y_c)\text{log}(1 - \sum_i \mathcal{P}_{sem}(B_i; y_c)). \tag{3}$$

The proposal group that contains multiple semantic-explicit proposals with the highest semantic confidence is selected, termed B^*.

Prompt Refiner II performs further ***proposal refinement*** as well as ***prototype update*** with a similar structure as the first phase, *i.e.*, comprising a classification head and an embedding head. For further refinement, based on the proposal group B^*, the proposal bag $\mathcal{B}_2$ is constructed as the input of this phase. The proposals in $\mathcal{B}_2$ maintain strong spatial correlation and have identified the main semantic regions. So in this phase, we only employ the general classification head with *Sigmoid* function to compute proposal score $s(b; y_c)$ and adopt the focal loss for further refinement. (More details are in the supplementary material.) The proposals with the top-k highest scores are weighted to obtain the final refined box b^*.

For prototype update, we treat the proposal score of this phase as an indicator. For example, for the proposal $b \in \mathcal{B}_2$, the corresponding embedding feature and score are denoted as z and s, respectively. During each training iteration, the embedded features whose corresponding score s exceeds a certain threshold τ are selected, as

$$v_c = \begin{cases} z, & s(b; y_c) \geq \tau, \\ 0, & otherwise, \end{cases} \tag{4}$$

where v_c denotes the local prototype of the current iteration for category c. And then the global semantic prototypes are updated with local prototypes via the EMA algorithm, as $\mathcal{V}_c = \alpha * \mathcal{V}_c + (1 - \alpha) * v_c$, where $\mathcal{V}_c$ denotes the semantic prototype of category c, and α is momentum parameter and empirically set to be 0.99. Consequently, we obtain a set of prototypes $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}_C\}$ for all categories, which are continuously updated during the training process.

Discussion. The design of prototypes allows the two refiners to mutually reinforce each other. We use the scores from *Prompt Refiner II* as an indicator to update the semantic prototypes with high-quality embeddings. High-quality prototypes are then utilized in *Prompt Refiner I* to compute instance probabilities, yielding high-quality proposals. When these proposals are inputed into *Prompt Refiner II*, the quality of the semantic prototypes is further enhanced, leading to a mutually improving process.

3.3 Prompt Guided Spatial Refinement

In this stage, we leverage SAM to further refine the spatial regions of objects. We use the pre-trained weights of SAM, without modifying the original structures and fine-tuning its parameters. Since SAM does not have the ability of semantic understanding, the ability to get the desired output relies on the accuracy of the semantic prompts. A single point is

Method	**Backbone**	**Sup.**	**AP**	**AP**$_{50}$	**AP**$_{75}$	**AP**$_{s}$	**AP**$_{m}$	**AP**$_{l}$
Faster R-CNN [Ren *et al.*, 2015]	ResNet-50	$\mathcal{F}$	37.4	58.1	40.4	21.2	41.0	48.1
RetinaNet [Lin *et al.*, 2017]	ResNet-50	$\mathcal{F}$	36.5	55.4	39.1	20.4	40.3	48.1
Sparse R-CNN [Sun *et al.*, 2021]	ResNet-50	$\mathcal{F}$	37.9	56.0	40.5	20.7	40.0	53.5
DINO [Zhang *et al.*, 2022]	ResNet-50	$\mathcal{F}$	49.0	66.4	53.3	31.4	52.2	64.0
DINO [Zhang *et al.*, 2022]	Swin-L	$\mathcal{F}$	57.0	75.7	62.7	40.5	61.1	73.5
ICMWSD [Ren *et al.*, 2020a]	ResNet-50	$\mathcal{I}$	12.6	26.1	-	-	-	-
CASD [Huang *et al.*, 2020]	ResNet-50	$\mathcal{I}$	13.9	27.8	-	-	-	-
SPE [Liao *et al.*, 2022]	CaiT	$\mathcal{I}$	7.2	18.2	4.8	-	-	-
WSCL [Seo *et al.*, 2022]	ResNet-101	$\mathcal{I}$	14.4	28.7	12.6	5.4	17.9	25.5
JLWSOD [Qi, 2023]	ResNet-50	$\mathcal{I}$	14.9	29.8	-	-	-	-
UFO^2 [Ren *et al.*, 2020b]	ResNet-50	$\mathcal{P}$	13.2	28.9	-	-	-	-
P2BNet-FR† [Chen *et al.*, 2022]	ResNet-50	$\mathcal{P}$	22.1	47.3	-	-	-	-
SAM-FR† [Kirillov *et al.*, 2023]	ResNet-50	$\mathcal{P}$	27.3	45.3	28.5	18.0	30.7	34.4
P2P-FR† (Ours)	ResNet-50	$\mathcal{P}$	31.6	53.8	32.7	20.5	35.7	38.5
P2P-DINO† (Ours)	ResNet-50	$\mathcal{P}$	38.2	57.2	40.9	25.6	42.6	46.9
P2P-DINO† (Ours)	Swin-L	$\mathcal{P}$	**45.1**	**66.1**	**48.9**	**33.5**	**50.3**	**53.8**

Table 1: Comparison of our P2P with other SOTA methods under different forms of supervision on COCO 2017 *val* set. Specifically, $\mathcal{F}$, $\mathcal{I}$, and $\mathcal{P}$ indicate full, image-level, and point-level supervision respectively. FR indicates Faster RCNN. † denotes fully-supervised refinement.

semantic-ambiguous because there are semantic biases between part and the global of an object (*e.g.*, *clothes* and *person*). In contrast, refined boxes generated in the SEPG stage cover the main semantic regions of objects and can be used as semantic explicit prompts to guide SAM in generating spatially refined regions.

SAM consists of three main components, a heavy image encoder (Φ_{img}), a light mask decoder (Φ_{mask}), and a prompt encoder (Φ_{pmt}). In our approach, we input the predicted boxes and the original point annotations as prompts into the prompt encoder to extract prompt embeddings. We extract the image embeddings in advance with the image encoder, and then during model training, the prompt embeddings and the off-the-shelf image embeddings are fed into the lightweight mask decoder to get the refined mask $\mathcal{M}$. The pre-extraction operation eliminates the computation of the heavyweight image encoder and greatly reduces the time and computation consumption. The overall process can be illustrated as follows:

$$\begin{aligned} F_{\text{img}} &= \Phi_{\text{img}}(\mathcal{I}), \qquad F_{\text{pmt}} = \Phi_{\text{pmt}}(\{P_{\text{box}}, P_{\text{point}}\}), \\ \mathcal{M} &= \Phi_{\text{mask}}(F_{\text{img}}, F_{\text{pmt}}), \end{aligned} \tag{5}$$

where $\mathcal{I}$ indicates the original image, F_{img} represents the image features encoded by Φ_{img}, F_{pmt} denotes prompt features encoded by Φ_{pmt}, and $\{P_{\text{box}}, P_{\text{point}}\}$ indicates the box and point prompts, respectively. With semantic-explicit prompts, SAM can further refine the target region to cover more complete objects and generate accurate masks.

3.4 Training and Evaluation

We follow the standard pipeline [Chen *et al.*, 2022] of point supervised tasks. **During training**, (i) a point-to-label regressor (P2P) is trained to obtain pseudo-labels (boxes or masks). (ii) After that, we retrain a task-related sub-network (*e.g.,* a detector or segmentation network) with the full supervision of pseudo-labels. **For evaluation**, we only use the retrained sub-network to obtain detection or segmentation results and evaluate the detection or segmentation performance.

Method	**Set**	**Backbone**	**Sup.**	**AP**$_{50}$
Faster R-CNN* [2015]	07	ResNet-50	$\mathcal{F}$	71.5
WSDDN [2016]	07	ResNet-50	$\mathcal{I}$	39.3
OICR [2017]	07	ResNet-50	$\mathcal{I}$	42.0
PCL [2018]	07	ResNet-50	$\mathcal{I}$	45.8
MELM [2018]	07	ResNet-50	$\mathcal{I}$	47.1
W2F† [2018]	07	ResNet-50	$\mathcal{I}$	52.4
CASD [2020]	07	ResNet-50	$\mathcal{I}$	56.8
SPE [2022]	0712	CaiT [2021]	$\mathcal{I}$	51.0
P2BNet-FR*† [2022]	07	ResNet-50	$\mathcal{P}$	48.3
SAM-FR† [2023]	07	ResNet-50	$\mathcal{P}$	47.9
P2P-FR† (Ours)	07	ResNet-50	$\mathcal{P}$	**61.9**

Table 2: The performance comparison of fully-supervised ($\mathcal{F}$), image-supervised ($\mathcal{I}$), and point-supervised ($\mathcal{P}$) detectors on Pascal VOC dataset. * indicates our re-implemented results and † denotes fully-supervised refinement.

4 Experiment

4.1 Experiment Settings

Datasets. We evaluate the proposed method on two benchmarks: MS COCO 2017 [Lin *et al.*, 2014] and Pascal VOC 2007 [Everingham *et al.*, 2015]. **MS COCO 2017** is a widely used large-scale dataset that contains 115K images in the *train* set and 5K images in *val* set, with 80 object categories collected in natural scenes. In **Pascal VOC 2007**, there are 2501 and 2510 images in training and validation sets, respectively, with 20 categories under common scenarios.

Evaluation Metrics. We use AP for MS COCO and VOC to measure the performance of detection and segmentation. And we report AP, AP$_{50}$, AP$_{75}$ for MS COCO and AP$_{50}$ for VOC. The mIoU and Correct Localization (CL) are also calculated to directly measure the quality of the pseudo boxes. Specifically, mIoU is calculated by the mean IoU between predicted pseudo boxes and their corresponding ground-truth

Method	Backbone	Sched.	Sup.	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Mask R-CNN [He *et al.*, 2017]	ResNet-50	1x	$\mathcal{F}$	34.7	55.7	37.2	18.3	37.4	47.2
Mask R-CNN [He *et al.*, 2017]	ViT-S	1x	$\mathcal{F}$	38.8	61.2	41.3	-	-	-
Mask2Former [Cheng *et al.*, 2022]	Swin-S	50e	$\mathcal{F}$	46.1	69.4	49.8	25.4	49.7	68.5
IRNet [Zhou *et al.*, 2019]	ResNet-50	1x	$\mathcal{I}$	6.1	11.7	5.5	-	-	-
BESTIE [Kim *et al.*, 2022]	HRNet-48	1x	$\mathcal{I}$	14.3	28.0	13.2	-	-	-
BoxInst [Tian *et al.*, 2021]	ResNet-101	3x	$\mathcal{B}$	33.2	56.5	33.6	16.2	35.3	45.1
DiscoBox [Lan *et al.*, 2021]	ResNet-50	3x	$\mathcal{B}$	32.0	53.6	32.6	11.7	33.7	48.4
WISE-Net [Laradji *et al.*, 2020]	ResNet-50	1x	$\mathcal{P}$	7.8	18.2	8.8	-	-	-
BESTIE † [Kim *et al.*, 2022]	HRNet-48	1x	$\mathcal{P}$	17.7	34.0	16.4	-	-	-
AttnShift† [Liao *et al.*, 2023]	ViT-S	1x	$\mathcal{P}$	21.2	42.0	19.4	-	-	-
SAM-MR† [Kirillov *et al.*, 2023]	ResNet-50	1x	$\mathcal{P}$	24.3	43.8	24.3	12.7	28.3	33.6
P2P-MR† (Ours)	ResNet-50	1x	$\mathcal{P}$	26.4	48.6	26.2	13.6	30.6	36.6
P2P-MF† (Ours)	Swin-S	50e	$\mathcal{P}$	**34.9**	**58.9**	**36.1**	**19.9**	**40.7**	**52.3**

Table 3: The segmentation performance of fully supervised $\mathcal{F}$, box-supervised $\mathcal{B}$, image-supervised $\mathcal{I}$, and point-supervised $\mathcal{P}$ methods on MS COCO 2017 *val* set. MR and MF indicate Mask RCNN and Mask2Former, respectively. † denotes fully-supervised refinement.

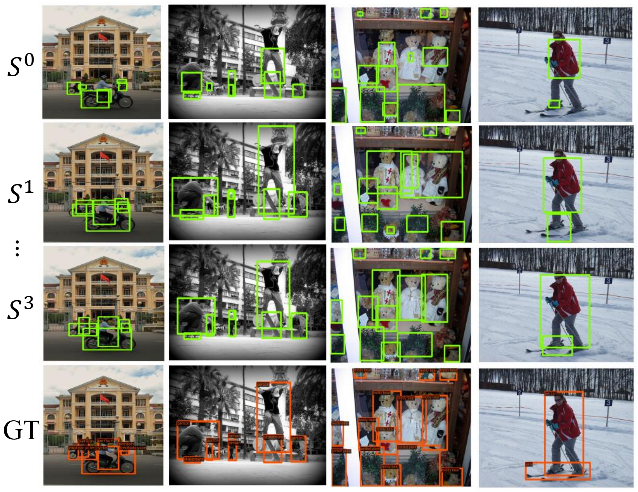

Figure 3: Visualization of the pseudo bounding boxes from different iterations of P2P and Ground Truths (GT). (Best viewed in color.)

bounding boxes of all objects in the training set. CL, denoting the correct localization rate, is computed as the ratio of IoU between the prediction and the ground truth exceeding a certain threshold. We report CL at thresholds of 0.5, 0.7, and 0.9 (termed CL@0.5, CL@0.7, and CL@0.9, respectively) to assess the quality of pseudo boxes.

Implementation Details. For SEPG, it is trained with SGD [Robbins and Monro, 1951] optimizer with batch size 16. The learning rate is initialized as 0.02. SEPG is trained for 12 epochs at the first round of iteration then 6 epochs in each subsequent round of iteration. For PGSR, we use the pre-trained ViT-H version of SAM [Kirillov *et al.*, 2023] and freeze its weights during the training stage. Before training starts, we pre-extract the image embeddings offline. So only the lightweight prompt encoder and mask decoder are involved in the PGSR, which greatly reduces the time and computational consumption of the training process.

Method	SEPG	PGSR	Iter	mIoU	AP	AP_{50}	AP_{75}
Baseline	-	-	-	57.5	22.1	47.3	-
Ours	✓			60.3	23.1	49.6	18.1
	✓	✓		68.1	30.4	52.7	31.2
	✓	✓	✓	**69.7**	**31.5**	**53.1**	**32.9**

Table 4: Effect of each component in P2P.

4.2 Performance and Comparison

P2P serves as a pseudo-boxes and -masks generator. By retraining Faster RCNN (FR) and Mask RCNN (MR) with pseudo-boxes and -masks in a fully supervised manner, we report the detection and segmentation performance. Furthermore, we conduct comparative analyses by benchmarking our method against fully supervised methods and with various forms of weakly supervised methods.

Detection Performance. We conduct comparisons between our method and fully, image-, and point-supervised detection methods using COCO and VOC datasets, as presented in Tab. 1 and 2. On the COCO dataset, our method outperforms the SOTA P2BNet [Chen *et al.*, 2022] by 9.5% (31.6% vs 22.1%) and 6.5% (53.8% vs 47.3%) in terms of AP and AP_{50}, respectively, and achieves 84% of the fully supervised Faster RCNN (FR). On the VOC dataset, our method achieves 86% of the performance of fully supervised FR. From Tab. 1, we find that the image-supervised methods perform poorly on the challenging COCO dataset, achieving only about 36% of the fully-supervised baseline. This is notably lower than the performance of the point-supervised method, highlighting the favorable trade-off between labeling burden and performance offered by the point-supervised approach. With the SOTA DINO [Zhang *et al.*, 2022] equipped with Swin-L [Liu *et al.*, 2021] backbone as the retrained detector, P2P-DINO even exceeds the fully supervised baseline methods (*e.g.,* Sparse RCNN), further demonstrating the potential of our method.

Segmentation Performance. Tab. 3 gives the performance of the segmentation methods on the COCO dataset with different forms of supervision. Our P2P-MR reports a

Method	**mIoU**	**AP**	**AP$_{50}$**	**AP$_{75}$**
SEPG-base	56.6	21.1	44.9	17.1
SEPG-G	55.2	21.5	46.6	16.9
SEPG-G-SC	**60.3**	**23.1**	**49.6**	**18.1**

Table 5: Effect of different techniques in SEPG stage. SEPG-base stands for seeds-based sampling, SEPG-G stands for group selection, and SEPG-G-SC stands for semantic confidence.

T	**mIoU**	**CL@0.5**	**CL@0.7**	**CL@0.9**
1	68.07	76.77	60.27	23.10
2	69.43	78.74	62.13	24.09
3	**69.70**	**79.05**	**62.40**	24.07
4	69.61	78.00	61.91	**24.35**

Table 6: Effect of different iteration T.

Method	**mIoU**	**CL@0.5**	**CL@0.7**	**CL@0.9**
P2BNet	57.5	65.9	35.7	9.3
SAM	58.25	60.32	47.32	21.27
P2P (Ours)	**69.70**	**79.05**	**62.40**	**24.07**

Table 7: Detailed comparison of the quality of pseudo labels.

Method	**Backbone**	**Epoch**	**Params (MB)**	
			Total	**Learnable**
P2BNet	ResNet-50	12	41.6	41.6
SEPG (Ours)	ResNet-50	$12 + 6 \times (T - 1)$	47.6	47.6
PGSR (Ours)	-	-	4.1	0
P2P (Total)			51.7	47.6

Table 8: Comparison of memory cost.

significant performance improvement of 5.2% AP over the AttnShift [Liao *et al.*, 2023] approach and outperforms the image-supervised BESTIE [Kim *et al.*, 2022] (with HRNet-48 [Wang *et al.*, 2020] as the backbone) by a large margin. Additionally, it also outperforms SAM-MR (where SAM is used as a pseudo-masks generator) by 2.1% AP and 4.8% AP$_{50}$. Furthermore, our method with only point supervision achieves nearly 80% of the performance exhibited by both baseline (Mask RCNN) and SOTA (Mask2Former [Cheng *et al.*, 2022]) fully supervised methods.

4.3 Ablation Study

We conduct analyses of the impact of key components of our method on the COCO dataset.

Effect of each component in P2P. As shown in Tab. 4, P2BNet is used as the baseline, and the key components include (i) **SEPG**: The group sampling reduces the solution space, and more accurate pseudo-labels are obtained with the two semantic confidence-guided refiners. It contributes to a mIoU improvement by more than 3 points. (ii) **PGSR**: SAM leverages semantic-explicit prompts for spatial refinement, leading to a substantial enhancement in the quality of pseudo-labels, with an improvement of about 8 points. (iii) Employing an **Iter**ative strategy where the two models mutually enhance each other, we observe a notable 12% improvement in performance compared to the baseline model.

Effect of different techniques in SEPG. We validate different techniques in SEPG and the results are presented in Tab. 5. The baseline model is designed by adopting the seeds-based sampling strategy followed by cascaded MIL refinement, referred to "SEPG-base", achieving 21.3 AP and 44.9 AP$_{50}$. Then, we adopt group refinement, *i.e.*, selecting groups first and then individual proposals. When using the MIL refinement head, referred to as "SEPG-G", AP reaches 21.5, as shown in the second row of Tab. 5. Further improvement is achieved with our designed semantic confidence refinement head, denoted as "SEPG-G-SC", resulting in a mIoU of 60.3 and an AP of 23.1, as shown in the fourth row of Table 5.

Effect of iterative training. We examine the impact of training iterations T. Tab. 6 reports the performance of P2P with different iteration numbers. We observe that the value of mIoU is consistently improved as T increases, reaching a peak when $T = 3$, followed by a stabilization around the peak. Notably, the highest performance surpasses that of $T = 1$ by nearly 2% across all metrics, showing the effectiveness of iterative training in elevating the quality of predictions. Some representative visualizations of different iterations are presented in Fig. 3.

Comparison of the quality of pseudo labels. As shown in Tab. 7, we conducted a more detailed comparison of the quality of pseudo-labels generated by P2BNet and single-point prompted SAM. Our approach achieved significant improvements, outperforming SAM by 11.45 and 18.73 points on mIoU and CL@0.5, respectively. This indicates that our method effectively enhances the quality of pseudo labels.

Comparison of memory cost. Tab. 8 illustrates the comparison of memory consumption between P2BNet and our P2P in the training stage. Thanks to our pre-extraction operation, even with the large foundation model, the number of parameters in the actual training process is still comparable, just about $10M$ higher than P2BNet.

5 Conclusion

In this paper, we introduce a point-supervised object detection and segmentation framework, called P2P, which transforms weak point-level annotations into explicit visual prompts, and guides the foundation model to produce the desired output by improving the semantic representation of the prompts. P2P performs as an iterative procedure that includes two stages: SEPG and PGSR. The SEPG performs semantic confidence lifting of the proposals through group sampling and semantic prototypes. The PGSR stage leverages SAM to output refined masks, which in turn are transformed into new proposal seeds. Finally, accurate pseudo masks and boxes are obtained in the iteration of SEPG and PGSR. Extensive experiments validate the effectiveness of our method.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62372382 and Grant 62306239.

References

[Arbeláez *et al.*, 2014] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 328–335, 2014.

[Bilen and Vedaldi, 2016] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2846–2854, 2016.

[Chen *et al.*, 2022] Pengfei Chen, Xuehui Yu, Xumeng Han, Najmul Hassan, Kai Wang, Jiachen Li, Jian Zhao, Humphrey Shi, Zhenjun Han, and Qixiang Ye. Point-to-box network for accurate object detection via single point supervision. In *European Conference on Computer Vision*, pages 51–67. Springer, 2022.

[Chen *et al.*, 2023a] Keyan Chen, Chenyang Liu, Hao Chen, Haotian Zhang, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *arXiv preprint arXiv:2306.16269*, 2023.

[Chen *et al.*, 2023b] Tianle Chen, Zheda Mai, Ruiwen Li, and Wei-lun Chao. Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation. *arXiv preprint arXiv:2305.05803*, 2023.

[Cheng *et al.*, 2022] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Everingham *et al.*, 2015] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015.

[Gao *et al.*, 2019] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9834–9843, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[He *et al.*, 2023] Sheng He, Rina Bao, Jingpeng Li, P Ellen Grant, and Yangming Ou. Accuracy of segment-anything model (sam) in medical image segmentation tasks. *arXiv preprint arXiv:2304.09324*, 2023.

[Huang *et al.*, 2020] Zeyi Huang, Yang Zou, BVK Kumar, and Dong Huang. Comprehensive attention self-distillation for weakly-supervised object detection. *Advances in neural information processing systems*, 33:16797–16807, 2020.

[Kim *et al.*, 2022] Beomyoung Kim, Youngjoon Yoo, Chae Eun Rhee, and Junmo Kim. Beyond semantic to instance segmentation: Weakly-supervised instance segmentation via semantic knowledge transfer and self-refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4278–4287, 2022.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[Lan *et al.*, 2021] Shiyi Lan, Zhiding Yu, Christopher Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S Davis, and Anima Anandkumar. Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3406–3416, 2021.

[Laradji *et al.*, 2020] Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. Proposal-based instance segmentation with point supervision. In *2020 IEEE International Conference on Image Processing*, pages 2126–2130. IEEE, 2020.

[Liao *et al.*, 2022] Mingxiang Liao, Fang Wan, Yuan Yao, Zhenjun Han, Jialing Zou, Yuze Wang, Bailan Feng, Peng Yuan, and Qixiang Ye. End-to-end weakly supervised object detection with sparse proposal evolution. In *European Conference on Computer Vision*, pages 210–226. Springer, 2022.

[Liao *et al.*, 2023] Mingxiang Liao, Zonghao Guo, Yuze Wang, Peng Yuan, Bailan Feng, and Fang Wan. Attentionshift: Iteratively estimated part-based attention map for pointly supervised instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 19519–19528, 2023.

[Lin *et al.*, 2014] 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 *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE international conference on computer vision*, pages 10012–10022, 2021.

[Qi, 2023] Lai Qi. Towards precise weakly supervised object detection via interactive contrastive learning of context information. *arXiv preprint arXiv:2304.14114*, 2023.

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

[Ren *et al.*, 2020a] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10598–10607, 2020.

[Ren *et al.*, 2020b] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Alexander G Schwing, and Jan Kautz. Ufo 2: A unified framework towards omni-supervised object detection. In *European conference on computer vision*, pages 288–313. Springer, 2020.

[Robbins and Monro, 1951] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[Seo *et al.*, 2022] Jinhwan Seo, Wonho Bae, Danica J Sutherland, Junhyug Noh, and Daijin Kim. Object discovery via contrastive learning for weakly supervised object detection. In *European Conference on Computer Vision*, pages 312–329. Springer, 2022.

[Sun *et al.*, 2021] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 14454–14463, 2021.

[Tang *et al.*, 2017] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2843–2851, 2017.

[Tang *et al.*, 2018] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):176–191, 2018.

[Tian *et al.*, 2021] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5443–5452, 2021.

[Touvron *et al.*, 2021] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE international conference on computer vision*, pages 32–42, 2021.

[Uijlings *et al.*, 2013] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104:154–171, 2013.

[Wan *et al.*, 2018] Fang Wan, Pengxu Wei, Jianbin Jiao, Zhenjun Han, and Qixiang Ye. Min-entropy latent model for weakly supervised object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1297–1306, 2018.

[Wang *et al.*, 2020] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

[Wang *et al.*, 2023] Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Qiushi Yang, Yiheng Liu, Songyao Zhang, et al. Review of large vision models and visual prompt engineering. *arXiv preprint arXiv:2307.00855*, 2023.

[Xu *et al.*, 2022] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4310–4319, 2022.

[Zhang *et al.*, 2018] Yongqiang Zhang, Yancheng Bai, Mingli Ding, Yongqiang Li, and Bernard Ghanem. W2f: A weakly-supervised to fully-supervised framework for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 928–936, 2018.

[Zhang *et al.*, 2020] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 12546–12555, 2020.

[Zhang *et al.*, 2022] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[Zhou *et al.*, 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[Zhou *et al.*, 2019] Yanning Zhou, Hao Chen, Jiaqi Xu, Qi Dou, and Pheng-Ann Heng. Irnet: Instance relation network for overlapping cervical cell segmentation. In *Medical Image Computing and Computer Assisted Intervention*, pages 640–648. Springer, 2019.