Crowd-SAM: SAM as a Smart Annotator for Object Detection in Crowded Scenes

Zhi Cai[1,2], Yingjie Gao[1,2], Yaoyan Zheng[1,2]
Nan Zhou[1,2], and Di Huang[1,2⋆]

[1]SKLSDE, Beihang University, Beijing, China
[2]IRIP Lab, SCSE, Beihang University, Beijing, China
{caizhi97, gaoyingjie, yaoyanzheng, zhounan0431, dhuang}@buaa.edu.cn

Abstract. Object detection is an important task that finds its application in a wide range of scenarios. Generally, it requires extensive labels for training, which is quite time-consuming, especially in crowded scenes. Recently, Segment Anything Model (SAM) has emerged as a powerful zero-shot segmenter, offering a novel approach to instance segmentation. However, the accuracy and efficiency of SAM and its variants are often compromised when handling objects in crowded scenes where occlusions often appear. In this paper, we propose Crowd-SAM, a SAM-based framework designed to enhance the performance of SAM in crowded scenes with the cost of few learnable parameters and minimal labeled images. We introduce an efficient prompt sampler (EPS) and a part-whole discrimination network (PWD-Net), facilitating mask selection and contributing to an improvement in accuracy in crowded scenes. Despite its simplicity, Crowd-SAM rivals state-of-the-art fully-supervised object detection methods on several benchmarks including CrowdHuman and CityPersons. Our code is available at https://github.com/FelixCaae/CrowdSAM.

Keywords: Detection in crowded scenes · Few-shot learning · Segment anything model

1 Introduction

Object detection in crowded scenes is a fundamental task in areas such as autonomous driving and video surveillance. The primary focus lies in identifying and locating densely packed common objects like pedestrians and vehicles, where occlusions present significant challenges. Great progress has been made in recent years, including two-stage methods [45, 60] and query-based methods [8, 22, 62]. However, these methods mainly follow a supervised manner and necessitate extensive labeled training samples, incurring a considerable annotation cost of approximately 42.4 seconds per object [38] The density and complexity of crowded scenes further aggravate the annotation burden. [1]

[⋆] Corresponding author

[1] On average, an image contains approximately 22 objects in CrowdHuman [37] and 7 in MS COCO [24]

The high cost of collecting object annotations drives the exploration of alternatives such as few-shot learning [34, 39, 44], weakly supervised learning [41, 53], semi-supervised learning [29,30,42,49], and unsupervised learning [1,6,21,26,48]. The best-performing ones, *i.e.* semi-supervised object detection (SSOD), leverage both labeled and unlabeled data for training and achieve a big success on common benchmarks *e.g.* PASCAL VOC [15] and COCO [24]. Unfortunately, SSOD introduces extra complexity such as complicated augmentations and online pseudo-labeling.

Recently, prompt-based segmentation models have received increasing attention due to their flexibility and scalability. Particularly, Segment Anything Model (SAM) [20] show its high capability to effectively and accurately predict the masks of regions specified by prompts, in any form of points, boxes, masks, or text descriptions. Recognizing its exceptional potential, researchers have made many efforts to adapt it for various vision tasks such as medical image recognition [31], remote sensing analysis [4, 12], industrial defect detection [52], *etc.*

Despite the huge progress [18,46,54] following SAM, applying SAM for object detection in crowded scenes is seldom studied. In this paper, we investigate the potential of SAM in such cases with two motivations. First, SAM is pre-trained on a very large dataset *i.e.* SA-1B that contains most of the common objects and it is thus reasonable to utilize the knowledge to facilitate labeling massive data and training a brand-new detector. Second, SAM demonstrates a robust segmentation ability in handling complicated scenes characterized by clustered objects that are difficult for an object detector trained from scratch.

To this end, we propose Crowd-SAM, a smart annotator powered by SAM for object detection in crowded scenes. As depicted in Fig. 1, we introduce a self-promoting approach based on DINOv2 to alleviate the cost of human prompting. Our method employs dense grids equipped with an Efficient Prompt Sampler (EPS) to cover as many objects as possible at a moderate cost. To distinguish the masks from multiple outputs precisely in occluded scenes, we design a mask selection module, termed Part-Whole Discrimination Network (PWD-Net) that learns to differentiate the output with the highest quality in Intersection over Union (IoU) score. With a lightweight model design and fast training schedule, it delivers considerable performance on public benchmarks including CrowdHuman [37] and CityPersons [58].

Our contributions can be summarized as follows:

1. We propose Crowd-SAM, a self-prompted segmentation method, for labeling images containing clustered objects, producing accurate results with only a few examples.
2. We design two novel components of Crowd-SAM, *i.e.* EPS and PWD-Net, which effectively unleash the ability of SAM on crowded scenes.
3. We conduct comprehensive experiments on two benchmarks to demonstrate the effectiveness and generalizable nature of Crowd-SAM.

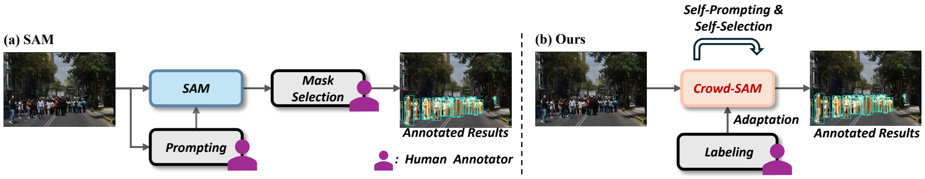


Fig. 1: Pipeline comparison between SAM and Crowd-SAM. Crowd-SAM only requires a few labeled images and can automatically recognize target objects.

2 Related Work

Object Detection. General object detection aims to identify and locate objects and is mainly divided into two categories: *i.e.* one-stage detectors and two-stage detectors. One-stage detectors predict bounding boxes and class scores by using image features [23, 27, 35], while two-stage detectors first generate region proposals and then classify and refine them [9, 10, 36]. Recently, end-to-end object detectors *e.g.* DETR [2, 55, 63] have replaced the hand-crafted modules such as Non-Maximum Suppression (NMS) by adopting one-to-one matching in the training phase, showing great potential in a wide variety of areas.

However, applying these detectors directly to pedestrian detection tends to incur performance degradation due to the fact that pedestrians are often crowded with occlusions appearing. Early work [32] proposes to integrate extra features into a pedestrian detector to explore low-level visual cues, while follow-up methods [5, 56] attempt to utilize the head areas for better representation learning. In [56], an anchor is associated with two targets, the whole body, and the head part, to achieve a more robust detector from joint training. AdaptiveNMS [25] adjusts the NMS threshold by predicting the density of pedestrians. Alternative methods focus on the design of loss functions to improve the training process. For example, RepLoss [45] encourages the prediction consistency of the same target while repels the ones from different targets. Recently, Zheng *et al.* [62] models the relation of queries to improve DETR-based detectors in crowded scenes and achieves remarkable success. Although these works have pushed the boundaries of object detection in crowded scenes to a new stage, they all rely on a large number of labeled samples for training, which is labor-intensive. This limitation inspires us to develop label-efficient detectors and automatic annotation tools, with the help of SAM.

Few-Shot Object Detection (FSOD). This task aims to detect objects of novel classes with limited training samples. FSOD methods can be roughly classified into meta-learning based [16, 51] and fine-tuning based ones [34, 39, 44]. Meta-RCNN [51] processes the query and support images in parallel via a siamese network. The Region of Interest (RoI) features of the query are fused with the class prototypes to effectively transfer knowledge learned from the support set. TFA [44] proposes a simple two-stage fine-tuning method that only fine-tunes the last layers of the network. FSCE [39] introduces a supervised contrastive loss in the fine-tuning stage to mitigate misclassification issues. De-FRCN [34] stops

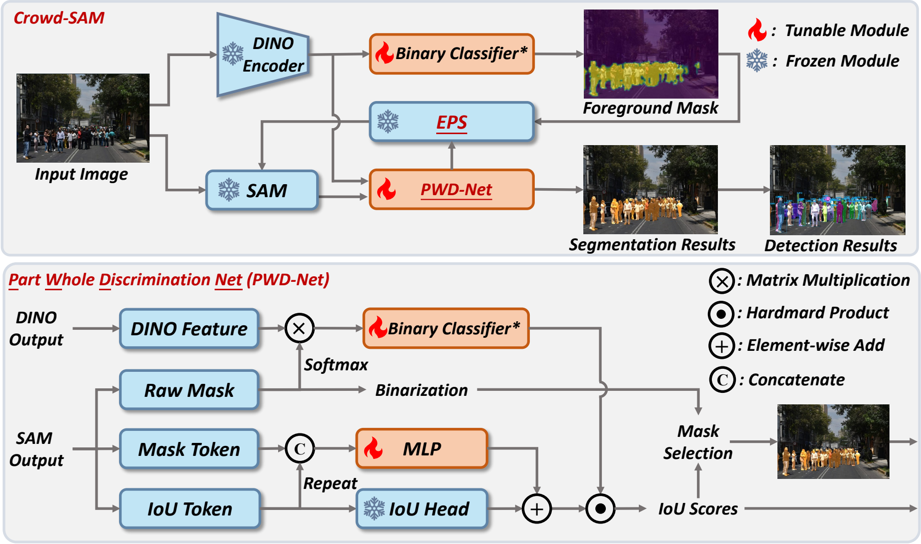


Fig. 2: The pipeline of Crowd-SAM shows the interaction between different modules. DINO encoder and SAM are frozen in the training process. * represents the parameters that are shared. For simplicity, the projection adapter of DINO is dismissed.

the gradient from the RPN and scales the gradient from R-CNN [36], followed by a prototypical calibration block to refine the classification scores.

Segment Anything Models. SAM [20], a visual foundation model for segmentation tasks, is trained on the SA-1B dataset using a semi-supervised learning paradigm. Its exposure to this vast repository of training samples renders it a highly capable class-agnostic model, effectively handling a wide range of objects in the world. Despite its impressive performance in solving segmentation tasks, it suffers from several issues like domain shift, inefficiency, class-agnostic design, *etc*. HQ-SAM [18] is proposed to improve its segmentation quality by learning a lightweight adapter. Fast-SAM [50] and Mobile-SAM [54] focus on fastening the inference speed of SAM by knowledge distillation. RSprompt [4] enables SAM to generate semantically distinct segmentation results for remote sensing images by generating appropriate prompts. Med-SA [47] presents a space-depth transpose method to adapt 2D SAM to 3D medical images and a hyper-prompting adapter to achieve prompt-conditioned adaptation. Unfortunately, these approaches necessitate a considerable amount of labeled data for effective adaptation, making them impractical for crowded scenes where annotation costs are prohibitive. Different from them, Per-SAM [57] and Matcher [28] teach SAM to recognize specified objects with only one or few instances by extracting training-free similarity priors. SAPNet [46] builds a weakly-supervised pipeline for instance segmentation. Although these approaches reduce data requirements, they still lag behind the demands of crowded scenes, such as pedestrian detection, particularly with occlusions.

3 Method

3.1 Preliminaries

SAM [20] is a powerful and promising segmentation model that comprises three main components: (**a**) an image encoder responsible for feature extraction; (**b**) a prompt encoder designed to encode the geometric prompts provided by users; and (**c**) a lightweight mask decoder that predicts the masks conditioned on the given prompts. Leveraging extensive training data, SAM demonstrates impressive zero-shot segmentation performance across various benchmarks. In particular, SAM makes use of points and boxes as prompts to specify interested regions.

DINO [3] represents a family of Vision Transformers (ViT) [7] learned in a self-supervised manner designed for general-purpose applications. During its training, DINO employs a self-distillation strategy akin to BYOL [11], fostering the learning of robust representations. DINOv2 [33] strengthens the foundation of DINO by integrating several additional pre-training tasks, improving its scalability and stability, especially for large models *e.g.* ViT-H (1 billion parameters). Thanks to its enhancement, DINOv2 shows a strong representation ability, in particular for the task of semantic segmentation.

3.2 Problem Definition and Overall Framework

Problem Definition. As shown in Fig. 1, our goal is to detect objects (*e.g.* pedestrians) in crowded scenes with few annotated data. We formulate this problem as a one-class few-shot detection task. A common few-shot pipeline is to divide data into the base split and the novel split. Differently, we directly use the data of the target class for model training, as the foundation models have already been trained on massive data. In particular, we employ segmentation masks as intermediate results which can be easily converted to bounding boxes. During the training and evaluation processes, only box annotations are provided.

Naive Study on SAM Auto-generator. The prompt number affects the performance of SAM and we analyze this issue for crowded scenes. In this case, we conduct several naive studies on CrowdHuman [37] with the auto generator of SAM, which utilizes grid points to search every region. Tab. 1 conveys three key observations: **(1)** dense grids are necessary for crowded scenes; **(2)** the ambiguity of distributions of point prompts and class-agnostic prompts incur many false positives (FPs); **(3)** the decoding time is a non-negligible burden when the grid size is large. Therefore, the **dense prompts** and **FP removal** are key aspects in designing SAM-based methods for detection/segmentation tasks in crowded scenes.

Overall Framework. Inspired by the studies above, we equip SAM [20] with several proper components to achieve an accurate and efficient annotation framework, as illustrated in Fig. 2. To accurately locate clustered objects, we employ the foundational model DINOv2 [33] to predict a semantic heat map, a task that can be formulated with a simple binary classifier. To discriminate

Table 1: Comparison in terms of recall, average FPs, and decoding time (T) of different grid sizes (N_G) adopted by the SAM generator on CrowdHuman [37]. The oracle model derives the prompts by computing the center of ground truth boxes. The decoding time is collected on a 3090 Ti GPU card.

N_G	16	32	64	128	Oracle
Recall	33.6	58.0	63.4	76.0	91.4
avg. FPs	51	112	227	485	-
T (s)	0.059	0.22	0.83	3.2	0.045

the output masks that is a mixture of correct masks, backgrounds, and part-level masks, we design Part-Whole Discrimination Network (PWD-Net) that takes as input both the learned tokens from SAM and the semantic-rich tokens from DINOv2 to re-evaluate all the outputs. Finally, to handle the redundancy brought by the use of dense grids, we propose an Efficient Prompt Sampler (EPS) to decode the masks at a moderate cost.

We introduce the details of our methods in the following sections.

3.3 Class-specific Prompt Generation

Generating a unique point prompt for each object (*e.g.* pedestrian) in crowded scenes is de facto a non-trivial problem. Thus, we take a step back and study how to detect objects with multiple prompts associated with one object and apply the proper post-processing techniques to remove duplicates. To this end, we adopt a heatmap-based prompt generation pipeline that initially classifies the regions and then generates prompts from the positive regions.

For the input image $I \in \mathbb{R}^{H\times W\times 3}$, we first use a pre-trained image encoder E_D to extract semantic rich features. To better transfer the pre-trained features to object segmentation, we add an MLP block after the final output layer and thus obtain the adapted features $F_{DINO} \in \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}\times C}$, where s is the patch size of DINOv2 [33] and C is the output channel. Then, we employ a segmentation head Head_{CLS} to classify F_{DINO} pixel by pixel, resulting in a heatmap $\hat{\mathbb{H}} \in [0,1]^{\frac{H}{s}\times\frac{W}{s}}$ that indicates the locations of objects as $\hat{\mathbb{H}} = \text{Head}_{CLS}(F_{DINO})$.

Given only the bounding box annotations $B \in [0,1]^{N_B\times 4}$, where N_B is the number of targets, this binary segmentation head can be optimized with box-level supervision. However, the coarse boundaries tend to incur considerable points scattering in background regions. To alleviate this issue, we use SAM [20] to generate high-quality mask-level pseudo labels, which is illustrated in Fig. 4, with ground truth (GT). The decoded masks are then merged into a single foreground mask $\mathbb{H} \in \{0,1\}^{256\times 256}$. We use the dice loss for training the adapter and segmentation head with the generated pseudo masks as follows:

$$\mathcal{L}_{fg} = dice(f(\hat{\mathbb{H}}), \mathbb{H}), \tag{1}$$

where f is an up-sampling function that resizes $\mathbb{H}$ to 256×256. During inference, we add a threshold t for mask binarization, which is simply set at 0.5 in our experiments. The binarized masks are mapped to point prompts P_G which only contain those in positive regions.

3.4 Semantic-guided Mask Prediction

Given the proposals generated in Sec. 3.3, our further aim is to efficiently decode the dense prompts and accurately discriminate the generated masks. As depicted in Fig. 3, each instance contains a set of prompts due to the density of grids. Supposing that only one in-position prompt is required for mask prediction in SAM, decoding all the prompts would lead to not only a waste of computation but also more FPs for some poorly located prompts.

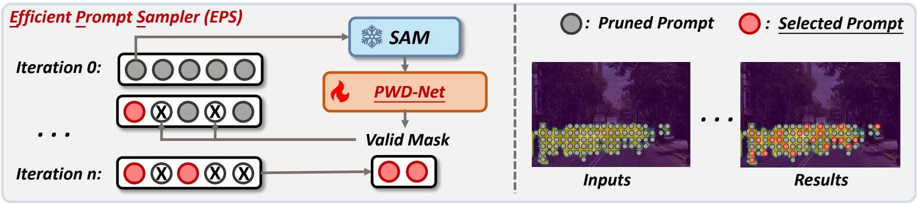


Fig. 3: Illustration of EPS. PWD-Net produces valid masks with a threshold. In each iteration, we prune prompts (*with a cross above*) that fall inside valid masks .

Efficient Prompt Sampler (EPS). To address this challenge, we introduce EPS, which dynamically prunes prompts based on the confidence of decoded masks. This method is elaborated in Algorithm 1. Beginning with the generated point prompts list P_G, EPS, in each iteration, samples a batch of prompts P_B from P_G using uniform random sampling. The sampled prompts P_B are then appended to the output points list P_S. Subsequently, we employ the SAM generator with the batched prompts P_B to produce masks M. Furthermore, we aggregate discriminant confidence scores S from PWD-Net, which are employed to select valid masks M' using a score threshold T, by using $M' = M[S > T]$. Refer to a detailed exposition of PWD-Net in the subsequent section. The valid masks represent regions that are believed to be well segmented, and we thus remove the points that have already been covered by any mask m in M' from the prompts list P_G. The iteration halts when P_G is empty. Additionally, EPS establishes a stopping criterion with a parameter K, terminating the sampling process once the total sample count reaches this limit. This parameter is instrumental in managing the overall decoding cost. In our experiments, K is empirically set to 500 to strike a balance between efficiency and recall.

Part-Whole Discrimination Network (PWD-Net). Given the raw masks predicted by SAM conditioned on the sampled batch of prompts P_B, we design an automatic selector for choosing the best-fitting mask. It is expected to optimize the outputs in two aspects: **(i)** refining the output IoU score according to

Algorithm 1: Process of Efficient Prompt Sampler (EPS).

```
Input: feature extracted from image $I$ by image encoder of SAM: $F_{SAM}$;
       generated prompt list: $P_G = \{p_1, p_2, ..., p_N\}$; SAM generator: $G$; mask
       confidence threshold: $T$; target sampled prompt list size: $K$
Output: sampled prompt list: $P_S$; valid mask list: $M_S$
$P_S \leftarrow \emptyset$;
while $|P_G| > 0$ and $|P_S| < K$ do
    Sample a batch of prompts $P_B \subseteq P_G$ with uniform random sampling;
    $P_S \leftarrow P_S \cup P_B$;
    $P_G \leftarrow P_G \setminus P_B$;
    Generate masks $M$ corresponding to $P_B$ by $G(F_{SAM}, P_B)$;
    Select valid masks $M'$ according to $T$ from $M$;
    $M_S \leftarrow M_S \cup M'$;
    for $p \in P_G$ do
        if $\exists m \in M'$ such that $p \in m$ then
            $P_G \leftarrow P_G \setminus \{p\}$
        end
    end
end
```

the quality of related masks if they are positive samples and **(ii)** suppressing the scores of samples that fall in background regions.

Illustrated in Fig. 2, for the masks generated corresponding to N prompts, we leverage the *Mask Tokens* and *IoU Tokens* within the mask decoder of SAM along with the sophisticated features extracted by the self-supervised pre-trained model DINOv2 [33]. $\mathcal{M}$ and $\mathcal{U}$ are responsible for mask decoding and IoU prediction in the SAM Mask Decoder, respectively. Thus, we suppose that they contain shape-aware information, which is helpful in discriminating the mask. These components enable us to compute a discriminant confidence score S for each specific prompt in a few-shot scenario. Initially, the refined IoU score S_{iou} is computed as follows:

$$S_{iou} = \text{Head}_{par}(\text{Concat}(\text{Repeat}(\mathcal{U}), \mathcal{M})) + \text{Head}_{IoU}(\mathcal{U}), \tag{2}$$

where Head_{par} is a parallel IoU head, consisting of an MLP block; $\mathcal{U} \in \mathbb{R}^{N \times 1 \times C}$ and $\mathcal{M} \in \mathbb{R}^{N \times 4 \times C}$ denote the *IoU Tokens* and *Mask Tokens* respectively. Notice in Eq. (2), S_{iou} is a sum of outputs from two individual heads, the parallel adapter head Head_{par} and the original Head_{IoU}. We freeze the parameters of Head_{IoU} to avoid overfitting in few-shot learning. Moreover, since these two tokens, $\mathcal{M}$ and $\mathcal{U}$ have different shapes, we repeat $\mathcal{U}$ by 4 times and concatenate these two tokens in the channel dimension. The refined score $S_{iou} \in \mathbb{R}^{N \times 4}$ encapsulates the quality of the generated masks by assessing the texture-aware feature from $\mathcal{M}$ and $\mathcal{U}$.

Further, by harnessing the semantic feature embedded in the self-supervised pre-trained model DINOv2, along with the mask data $\hat{M}$, we calculate the discriminant score S_{cls} :

$$S_{cls} = \sigma(\text{Head}_{CLS}(\mathcal{O})), \mathcal{O} = \text{Pool}(d(\text{Softmax}(\hat{M})) \circ F_{DINO}). \tag{3}$$

Here, $\mathcal{O} \in \mathbb{R}^{N \times C}$ denotes the extracted *Semantic Token*, and $\hat{M} \in \mathbb{R}^{N \times H \times W \times 4}$ and $F_{DINO} \in \mathbb{R}^{N \times h \times w \times C}$ represent the masks generated by SAM [31] and the features extracted by DINOv2, respectively. We denote d as a down-scale function that resizes the mask to $h \times w$, consistent with F_{DINO}. Pool is a global pooling function that conducts mean pooling on the x-axis and y-axis and $\circ$ is the Hadamard product. The discriminant score $S_{cls} \in \mathbb{R}^{N \times 4}$ predicts whether masks belong to the foreground or background. Head_{CLS} shares the same parameters with the binary classifier introduced in Sec. 3.3. Finally, we calculate a joint score of discrimination and estimate the quality for masks by simply multiplying the two scores: $S = S_{iou} \cdot S_{cls}$.

During training, prompts sampled from real masks are taken as input. Regarding the prompts within the foreground, the discriminant confidence score aims to accurately predict the IoU of the generated and real masks. Conversely, for the prompts within the background, the score is ideally 0. Hence, the loss function for this aspect is formulated as follows:

$$s^i_{target} = \begin{cases} IoU(m^i, m^i_{GT}), & m^i_{GT} \in M^{bg}_{GT}, \\ 0, & m^i_{GT} \in M^{fg}_{GT}, \end{cases} \tag{4}$$

$$\mathcal{L}_{iou} = \text{MSE}(S, S_{target}). \tag{5}$$

Here, s^i_{target} denotes the target score of the mask m^i generated by the i^{th} prompt, and $S_{target} = \{s^i_{target}\}^N_{i=1}; M = \{m^i\}^N_{i=1}; M_{GT} = \{m^i_{GT}\}^N_{i=1} = M^{bg}_{GT} \cup M^{fg}_{GT}$.

3.5 Training and Inference

The total training loss of the entire framework combines Eq. (1) and Eq. (5):

$$\mathcal{L} = \mathcal{L}_{fg} + \mathcal{L}_{iou} \tag{6}$$

At inference, we select the mask with the highest confidence score S among the 4 masks as the output of PWD-Net, which we denote as $M \in \mathbb{R}^{N \times H \times W}$ and $S_o \in \mathbb{R}^N$. We adopt a window cropping strategy to enhance the performance on small objects as [20]. This strategy slices the whole image into overlapping crops where each crop is individually processed. The final results are merged from the outputs of each crop and here we apply NMS to remove duplicate proposals.

4 Experiments

Datasets. Following [62], we adopt CrowdHuman [37] as the benchmark to conduct main experiments and ablation studies. CrowdHuman [37] collects and annotates images containing crowded persons from the Internet. It contains 15,000,

Table 2: Comparative results (%) on CrowdHuman [37] *val*. All the SAM-based methods adopt ViT-L [7] as the pre-trained backbone (SRCNN denotes Sparse R-CNN [40], a baseline in [62] and * represents using the multi-crop trick).

Method	Backbone	#Shot	AP	MR^{-2}	Recall	Secs/Img
Fully supervised object detectors						
ATSS [59]	ResNet-50 [14]	Full	80.3	59.7	86.1	0.051
FCOS [43]	ResNet-50 [14]	Full	76.3	65.5	82.6	0.045
Iter-SRCNN [62]	ResNet-50 [14]	Full	85.9	58.3	93.3	0.25
DINO [55]	ResNet-50 [14]	Full	86.7	57.6	94.5	0.27
Few-shot object detectors						
TFA [44]	ResNet-101 [14]	10	46.9	84.3	57.9	0.067
FSCE [39]	ResNet-101 [14]	10	43.0	84.7	50.0	0.072
De-FRCN [34]	ResNet-101 [14]	10	46.4	85.9	65.5	0.072
SAM-based approaches						
SAM [20]	ViT-L [7]	0	-	-	65.6	1.3
SAM* [20]	ViT-L [7]	0	-	-	79.6	6.7
Matcher [34]	ViT-L [7]	1	8.0	88.9	23.9	22.0
Crowd-SAM	ViT-L [7]	10	71.4	75.1	83.9	1.7
Crowd-SAM*	ViT-L [7]	10	78.4	74.8	85.6	8.1

4,370, and 5,000 images for training, validation, and testing, respectively. We also evaluate our method on CityPersons [58] for a realistic urban-scene scenario. Additionally, we utilize OCC-Human [61], which is specially reputed for occluded persons. For these pedestrian datasets, we use visible annotations (only including visible areas of an object) for training and evaluation. To validate the extensibility of Crowd-SAM, we further devise a multi-class version of Crowd-SAM by adding a multi-class classifier. We employ 0.1 % percent of the COCO [24] *trainval* set for training and the COCO *val* set for validation. Besides, we validate our method on a subset with occluded objects on COCO, *i.e.* COCO-OCC [17], extracted by selecting the images whose objects have a high overlapping ratio.

Implementation Details. We utilize SAM (ViT-L) [20] and DINOv2 (ViT-L) [33] as the base models for all experiments. In the fine-tuning stage, all their parameters are frozen to avoid over-fitting. Instead of real GT, we use the pseudo masks generated by SAM as S_{target} to supervise the learning of PWD-Net in Eq. (5). These generated pseudo labels are of high quality as shown in Fig. 4(c). We randomly pick the points from the pseudo masks as positive training samples and the ones from the background as negative training samples. In the training process, we use Adam [19] with a learning rate of 10^{-5}, a weight decay of 10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.99$ for optimization. We train our module for 2,000 iterations with a batch size of 1, which can be done on a single GTX 3090 Ti GPU in several minutes. For more details, please refer to the Appendix.

Evaluation Metrics. Following [62], we use AP with IoU threshold at 0.5, MR^{-2}, and Recall as metrics. Generally, a higher AP, Recall, and lower MR^{-2} value indicates better performance.

4.1 Experimental Results on Pedestrian Detection

For fair comparison, we re-implement the counterparts [34,43,55,59,62] with a 2× schedule using visible annotations in the CrowdHuman [37] and CityPersons [58] datasets.

Main Results. We compare Crowd-SAM with most related methods including *fully-supervised object detectors* [43, 55, 59, 62], *few shot object detectors* [34] and *SAM-based methods* [20,28]. Notice that we use visible annotations for which derive different results from those on full box annotations.

As Tab. 2 shows, with only 10 labeled images, Crowd-SAM achieves comparable performance with *full-supervised object detectors* whose best result is 86.7% AP, delivered by DINO [55]. Particularly, Crowd-SAM outperforms an advanced anchor-free detector FCOS [43] by 2.1% AP. These results indicate that by using proper adaptation techniques, SAM can reach very competitive performance on complex pedestrian detection datasets like CrowdHuman. On the other side, Crowd-SAM achieves SOTA performance on few-shot detection settings and outperforms all the *few-shot object detectors* by a significant margin. DeFRCN [34] is a well-established few-shot object detector equipped with ResNet-101 [14] that reports 46.4% in AP. Notably, our method exceeds it by 32%, indicating the superiority in the few-shot detection setting. The qualitative comparison between Crowd-SAM and De-FRCN is depicted in Fig. 4. For *SAM-based methods*, our Crowd-SAM largely leads the SAM baseline by 6% (with multi-crop) and 18.3% (w/o multi-crop) in Recall. Besides, Crowd-SAM is also superior to the other methods such as Matcher [28].

Results on OccHuman and CityPersons. To investigate the performance of Crowd-SAM in occluded scenes, we compare it with an advanced counterpart Pose2Seg [61], and the results are shown in Tab. 3. It is noteworthy that Pose2Seg is a fully-supervised detector while Crowd-SAM is a few-shot detector. As can be seen from the results, Crowd-SAM leads Pose2Seg by 9.2% AP which demonstrates its robustness in occluded scenes. Also, we apply our method to a not-so-crowded but more realistic urban dataset, *i.e.* CityPersons [58], as reported in Tab. 4. Crowd-SAM outperforms TFA [44] by 2.7% AP, FSCE [39] by 1.1% AP, and De-FRCN [34] by 7.8% AP. In conclusion, our method remains competitive compared to advanced few-shot object detectors, even though it is not specifically designed for sparse scenes.

These results illustrate that Crowd-SAM well unleashes the power of vision foundation models, *i.e.* SAM and DINO, in all of the crowded, occluded, and urban scenes.

4.2 Experimental Results on Multi-class Object Detection

To further explore the extensibility in a more popular setting, we devise a multi-class Crowd-SAM. The multi-class version of Crowd-SAM is slightly different by replacing the binary classifier with a multi-class one. We then validate multi-class Crowd-SAM on COCO [24], a widely adopted common object detection

Table 3: Comparative results (%) on OccHuman *val*, where AP_M and AP_H represent AP in moderate and hard cases according to occlusion ratios, respectively.

Method	Backbone	AP	AP_M	AP_H
Mask R-CNN [13]	ResNet50-FPN [14]	16.3	19.4	11.3
Pose2Seg [61]	ResNet50-FPN [14]	22.2	26.1	15.0
Crowd-SAM	ViT-L [7]	**31.4**	**26.5**	**17.7**

Table 4: Comparative results (%) on CityPersons *val*.

Method	# Shot	AP	MR^{-2}
FCOS [43]	Full	58.8	30.0
ATSS [59]	Full	54.1	27.8
Iter-SRCNN [62]	Full	57.9	31.0
TFA [44]	50	30.6	53.8
FSCE [39]	50	32.2	46.5
De-FRCN [34]	50	25.5	67.1
Crowd-SAM	50	33.3	31.7

benchmark, and COCO-OCC [17] which is a split of COCO that mainly consists of images with a high occlusion ratio.

We compare our method with two supervised detectors, *i.e.* Faster R-CNN [36] and BCNet [17], and report the results in Tab. 5. It can be seen that our Crowd-SAM is comparable to the supervised detectors on both datasets and drops only 1.4 AP% when comparing those of COCO and COCO-OCC. This minor drop indicates that our method is robust to occlusions.

4.3 Ablation Studies

Ablation on Modules. We conduct ablation studies on the key components of Crowd-SAM, including foreground location, EPS, and PWD-Net, to validate their effectiveness. In Tab. 6, the performance in AP significantly drops by 7.4% and recall by 8.5% when FG location is removed, indicating the importance of restricting foreground areas. As for EPS, once it is removed, the AP drops by 0.6%. We suppose that EPS not only accelerates the sampling process of dense prompts but also helps focus on the difficult part of the image, which conveys ambiguous semantics. We also compare EPS with some other batch iterators in Tab. 7. Besides, we find that PWD-Net is indispensable and when removed, the AP dramatically drops to 17.0%. Finally, we point out that multi-cropping is a strong trick to enhance the performance which contributes 7.0% AP to the final performance. Overall, these results prove that all the components are essential.

Ablation on EPS. We conduct a comprehensive study on EPS by comparing it with several variants. We forward each image only once to avoid the extra latency caused by multi-crop. We use a default sampler that iterates through all training samples and a random iterator with a halting threshold K as counterparts. As reported in Tab. 7, the AP and Recall increase with the grid size for

Table 5: Comparative results (%) on COCO *val* and COCO-OCC.

Methods	Backbone	COCO-OCC		COCO	
		AP	AP_{50}	AP	AP_{50}
Faster R-CNN [36]	ResNet50-FPN [14]	29.7	50.0	33.5	53.7
BCNet [17]	ResNet50-FPN [14]	31.7	51.1	34.6	54.4
Crowd-SAM	ViT-L [7]	20.6	31.5	22.0	33.7

Table 6: Ablation results (%) on the main components in Crowd-SAM. FG location represents the use of the binary classifier on DINOv2 features. The last line represents a SAM baseline with a standard 32 × 32 grid as inputs.

FG loc.	EPS	PWD-Net	Multi-Crop	AP	MR^{-2}	Recall
✓	✓	✓	✓	78.4	74.8	85.6
	✓	✓	✓	71.0	77.9	77.1
✓		✓	✓	77.8	71.8	83.9
✓	✓		✓	17.0	99.6	83.1
✓	✓	✓		71.4	75.1	83.9
			✓	8.7	99.8	70.1

all three samplers. However, the default sampler suffers an out-of-memory error when the grid size reaches 128, preventing it from being adopted in this setting. As for the random sampler, its performance is constrained by K and a grid size larger than 64 only leads to limited improvement, *e.g.* 0.1% AP. On the contrary, our EPS benefits from a much larger grid size and achieves a better result.

Ablation on PWD-Net. We compare PWD-Net to the variants that replace some tokens with a full-zero placeholder. We also consider two designs, directly tuning the IoU head or learning a parallel IoU head. As shown in Tab. 8, all the three tokens, *i.e.* Mask Token $\mathcal{M}$, IoU Token $\mathcal{U}$, and Semantic Token $\mathcal{O}$, contribute to the final result. Particularly, once $\mathcal{M}$ is removed, the AP drops by 40.0%, which is a catastrophic decline. This degradation suggests that the mask token contains the shape-aware feature that is essential for the part-whole discrimination task. Notably, the AP drops by 2.8% when we tune the pre-trained IoU Head, suggesting that it is prone to overfit the few labeled images. By freez-

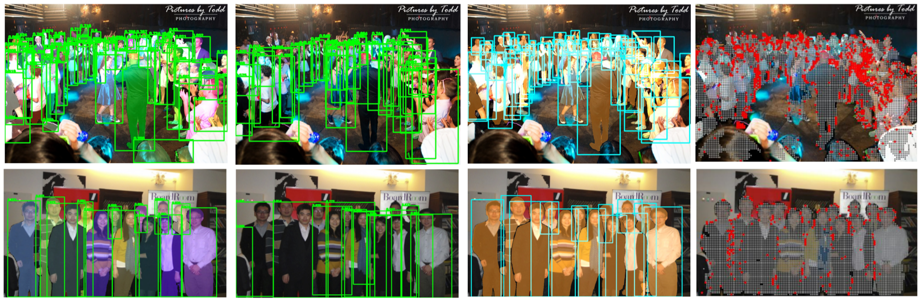
Table 7: Comparison (%) of different samplers on CrowdHuman [37] *val*. Full means using all prompts. OOM represents out-of-memory errors which occur when the GPU memory is all consumed.

		Full		Random		EPS	
Grid	K	AP	Recall	AP	Recall	AP	Recall
32 × 32	500	57.6	60.5	57.6	60.5	57.0	60.0
64 × 64	500	69.4	73.9	69.4	73.9	69.3	73.6
128 × 128	500	OOM		69.7	74.1	72.4	77.8
192 × 192	500	OOM		69.8	73.7	**73.2**	**78.2**
256 × 256	500	OOM		68.9	73.0	72.3	78.0

Table 8: Ablation results (%) on the design of PWD-Net. $\mathcal{M}$, $\mathcal{U}$, and $\mathcal{O}$ represent the mask token, IoU token, and semantic token, respectively. For the IoU head, F means freezing the original IoU head and training a parallel one, and T indicates tuning the original IoU head.

$\mathcal{M}$	$\mathcal{U}$	$\mathcal{O}$	IoU head	AP	MR^{-2}
✓	✓	✓	F	78.4	74.8
✓	✓	✓	T	75.6(-2.8)	80.4(+5.6)
✓	✓		F	77.3(-1.1)	73.5(-1.3)
✓		✓	F	76.3(-2.1)	74.8(-0.0)
	✓	✓	F	38.4(-40.0)	95.8(+21.0)

ing the IoU head of SAM, PWD-Net can benefit more from the shape-aware knowledge learned from massive segmentation data.


(a) Crowd-SAM

(b) De-FRCN

(c) Ground Truth

(d) Prompts

Fig. 4: Qualitative comparison between Crowd-SAM (a) and De-FRCN (b). Crowd-SAM predictions are more accurate especially in the boundaries of persons. We also plot the GT boxes (*blue rectangles*) and the generated masks (*yellow regions*), which are of high quality (c). In (d), we plot our prompt filtering results, where preserved prompts (*red points*) are much fewer than the removed ones (*gray points*). Zoom in for a better view.

5 Conclusion

This paper proposes Crowd-SAM, a SAM-based framework, for object detection and segmentation in crowded scenes, designed to streamline the annotation process. For each image, Crowd-SAM generates dense prompts for high recall and uses EPS to prune redundant prompts. To achieve accurate detection in occlusion cases, Crowd-SAM employs PWD-Net which leverages several informative tokens to select the masks that best fit. Combined with the proposed modules, Crowd-SAM achieves 78.4% AP on CrowdHuman, comparable to full-supervised detectors, validating that object detection in crowded scenes can benefit from foundation models like SAM with data efficiency.

Acknowledgements

This work is partly supported by the National Natural Science Foundation of China (No. 62022011), the Research Program of State Key Laboratory of Software Development Environment, and the Fundamental Research Funds for the Central Universities.

References

1. Bar, A., Wang, X., Kantorov, V., Reed, C.J., Herzig, R., Chechik, G., Rohrbach, A., Darrell, T., Globerson, A.: Detreg: Unsupervised pretraining with region priors for object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14605–14615 (2022)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Eur. Conf. Comput. Vis. pp. 213–229. Springer (2020)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Int. Conf. Comput. Vis. pp. 9650–9660 (2021)
4. Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., Shi, Z.: Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. IEEE Transactions on Geoscience and Remote Sensing (2024)
5. Chi, C., Zhang, S., Xing, J., Lei, Z., Li, S.Z., Zou, X.: Pedhunter: Occlusion robust pedestrian detector in crowded scenes. In: AAAI. vol. 34, pp. 10639–10646 (2020)
6. Dai, Z., Cai, B., Lin, Y., Chen, J.: Up-detr: Unsupervised pre-training for object detection with transformers. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1601–1610 (2021)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Int. Conf. Learn. Represent. (2021)
8. Gao, F., Leng, J., Gan, J., Gao, X.: Selecting learnable training samples is all detrs need in crowded pedestrian detection. In: ACM Int. Conf. Multimedia. pp. 2714–2722 (2023)
9. Girshick, R.: Fast r-cnn. In: Int. Conf. Comput. Vis. pp. 1440–1448 (2015)
10. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 580–587 (2014)
11. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Adv. Neural Inform. Process. Syst. **33**, 21271–21284 (2020)
12. Gui, S., Song, S., Qin, R., Tang, Y.: Remote sensing object detection in the deep learning era—a review. Remote Sensing **16**(2), 327 (2024)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Int. Conf. Comput. Vis. pp. 2961–2969 (2017)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 770–778 (2016)

15. Hoiem, D., Divvala, S.K., Hays, J.H.: Pascal voc 2008 challenge. World Literature Today **24**(1), 1–4 (2009)
16. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: Int. Conf. Comput. Vis. pp. 8420–8429 (2019)
17. Ke, L., Tai, Y.W., Tang, C.K.: Deep occlusion-aware instance segmentation with overlapping bilayers. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4019–4028 (2021)
18. Ke, L., Ye, M., Danelljan, M., Tai, Y.W., Tang, C.K., Yu, F., et al.: Segment anything in high quality. In: Adv. Neural Inform. Process. Syst. vol. 36 (2024)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Int. Conf. Learn. Represent. (2015)
20. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Int. Conf. Comput. Vis. pp. 4015–4026 (2023)
21. Li, M., Wu, J., Wang, X., Chen, C., Qin, J., Xiao, X., Wang, R., Zheng, M., Pan, X.: Aligndet: Aligning pre-training and fine-tuning in object detection. In: Int. Conf. Comput. Vis. pp. 6866–6876 (2023)
22. Lin, M., Li, C., Bu, X., Sun, M., Lin, C., Yan, J., Ouyang, W., Deng, Z.: Detr for crowd pedestrian detection. arXiv preprint arXiv:2012.06785 (2020)
23. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Int. Conf. Comput. Vis. pp. 2980–2988 (2017)
24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Eur. Conf. Comput. Vis. pp. 2906–2917 (2014)
25. Liu, S., Huang, D., Wang, Y.: Adaptive nms: Refining pedestrian detection in a crowd. In: IEEE Conf. Comput. Vis. Pattern Recog. (June 2019)
26. Liu, S., Li, Z., Sun, J.: Self-emd: Self-supervised object detection without imagenet. arXiv preprint arXiv:2011.13677 (2020)
27. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Eur. Conf. Comput. Vis. pp. 21–37. Springer (2016)
28. Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., Shen, C.: Matcher: Segment anything with one shot using all-purpose feature matching. In: Int. Conf. Learn. Represent. (2024)
29. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. In: Int. Conf. Learn. Represent. (2021)
30. Liu, Y.C., Ma, C.Y., Kira, Z.: Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9819–9828 (2022)
31. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
32. Mao, J., Xiao, T., Jiang, Y., Cao, Z.: What can help pedestrian detection? In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3127–3136 (2017)
33. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. Trans. Mach. Learn Res. (2024)
34. Qiao, L., Zhao, Y., Li, Z., Qiu, X., Wu, J., Zhang, C.: Defrcn: Decoupled faster r-cnn for few-shot object detection. In: Int. Conf. Comput. Vis. pp. 8681–8690 (2021)

35. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 779–788 (2016)
36. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Adv. Neural Inform. Process. Syst. vol. 28 (2015)
37. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123 (2018)
38. Su, H., Deng, J., Fei-Fei, L.: Crowdsourcing annotations for visual object detection. In: Workshops at the twenty-sixth AAAI conference on artificial intelligence (2012)
39. Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C.: Fsce: Few-shot object detection via contrastive proposal encoding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7352–7362 (2021)
40. Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al.: Sparse r-cnn: End-to-end object detection with learnable proposals. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14454–14463 (2021)
41. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2843–2851 (2017)
42. Tang, Y., Chen, W., Luo, Y., Zhang, Y.: Humble teachers teach better students for semi-supervised object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3132–3141 (2021)
43. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Int. Conf. Comput. Vis. pp. 9627–9636 (2019)
44. Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly simple few-shot object detection. Int. Conf. Mach. Learn. (2020)
45. Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., Shen, C.: Repulsion loss: Detecting pedestrians in a crowd. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7774–7783 (2018)
46. Wei, Z., Chen, P., Yu, X., Li, G., Jiao, J., Han, Z.: Semantic-aware sam for point-prompted instance segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3585–3594 (2024)
47. Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y., Arbel, T.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
48. Xie, E., Ding, J., Wang, W., Zhan, X., Xu, H., Sun, P., Li, Z., Luo, P.: Detco: Unsupervised contrastive learning for object detection. In: Int. Conf. Comput. Vis. pp. 8392–8401 (2021)
49. Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z.: End-to-end semi-supervised object detection with soft teacher. In: Int. Conf. Comput. Vis. pp. 3060–3069 (2021)
50. Xu, Z., Wenchao, D., Yongqi, A., Yinglong, D., Tao, Y., Min, L., Ming, T., Jinqiao, W.: Fast segment anything. arXiv preprint arXiv:2306.12156 (2023)
51. Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L.: Meta r-cnn: Towards general solver for instance-level low-shot learning. In: Int. Conf. Comput. Vis. pp. 9577–9586 (2019)
52. Ye, Z., Lovell, L., Faramarzi, A., Ninic, J.: Sam-based instance segmentation models for the automation of masonry crack detection. arXiv preprint arXiv:2401.15266 (2024)

53. Zeng, Z., Liu, B., Fu, J., Chao, H., Zhang, L.: Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In: Int. Conf. Comput. Vis. (2019)
54. Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S.H., Lee, S., Hong, C.S.: Faster segment anything: Towards lightweight sam for mobile applications. arXiv preprint arXiv:2306.14289 (2023)
55. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.Y.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: Int. Conf. Learn. Represent. (2023)
56. Zhang, K., Xiong, F., Sun, P., Hu, L., Li, B., Yu, G.: Double anchor r-cnn for human detection in a crowd. arXiv preprint arXiv:1909.09998 (2019)
57. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Dong, H., Qiao, Y., Gao, P., Li, H.: Personalize segment anything model with one shot. In: Int. Conf. Learn. Represent. (2024)
58. Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3213–3221 (2017)
59. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9759–9768 (2020)
60. Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z.: Occlusion-aware r-cnn: Detecting pedestrians in a crowd. In: Eur. Conf. Comput. Vis. pp. 637–653 (2018)
61. Zhang, S.H., Li, R., Dong, X., Rosin, P., Cai, Z., Han, X., Yang, D., Huang, H., Hu, S.M.: Pose2seg: Detection free human instance segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 889–898 (2019)
62. Zheng, A., Zhang, Y., Zhang, X., Qi, X., Sun, J.: Progressive end-to-end object detection in crowded scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 857–866 (2022)
63. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: Int. Conf. Learn. Represent. (2021)