Appendix of Crowd-SAM

Table of contents:

A Implementation Details of Crowd-SAM

Table I: Summarization of hyper-parameter setting in Crowd-SAM

Hyper-Parameter	Value	Meaning
K	500	Maximum of prompts used in each crop
G_s	192	Grid size of in generating prompts
N_c	2	Number of levels in multi-cropping
R_s	512/1500	Overlapping ratio of cropped boxes
T_n	0.7	NMS threshold in merging crops
T_c	0.3	Confidence threshold
T_m	0.65	NMS threshold in each crop
T_s	0.8	Stability score threshold
δ	1	Mask jittering amplititude

We summarize the hyper-parameters used in automatic mask generation in Tab. I.

Inference details. For the multi-cropping setting, we employ $N = 2$ which means two-level of scales are utilized and an overlapping ratio of $R_s = 512/1500$, to strike a balance between performance and efficiency. The pre-processing of the input image and the mask post-processing is the same as that in SAM [4]. In the foreground location, we initialize a 192×192 grid and use a threshold $t = 0.5$ for filtering positive points. For mask decoding, we sample 32-point prompts in each iteration and set an upbound of $K = 500$ for the total sampled prompts. For EPS, we set the confidence threshold $T = 0.5$ to only utilize highly confident samples. We also adopt the same post-processing as SAM does including mask quality evaluation and NMS.

Training details. For the training of Crowd-SAM, we first need to generate mask labels to supervise PWD-Net. This step can be done with the help of SAM itself, and we utilize SAM (ViT-L) with GT bounding boxes as inputs to predict the masks. In the training process, we randomly pick several points inside the masks as positive points and several points outside as negative points. For positive points (prompts), we compute the IoU between the point-prompted mask and the box-prompted mask as the target. For negative points (prompts), we set the target equal to zero. The positive and negative points are kept with a ratio of 1:3 for balance. We train the model on one GPU with 2,000 iterations.

Post-processing. **Post-processing**. For the multiple masks predicted by SAM, we use the refined confidence score to select only the best confident one. This step helps reduce the time of processing massive masks. We use a score filter with a confidence threshold $T_c = 0.2$ to filter the predictions associated with background. Also, we compute a stability score s as defined in [4] by jittering on the masks as follows:

$$s = \frac{area(M + \delta)}{area(M - \delta)}, \tag{1}$$

where $M \in \mathbb{R}^{256\times256}$ is a raw predicted mask and δ is a hyper-parameter to control the amplitude of jittering. The low-quality masks are filtered with a threshold $T_s = 0.8$. Then for each crop, we use NMS with $T_m = 0.65$ to filter redundant predictions. We also filter those incomplete objects whose boundaries are close to the cropping edge. Finally, the results of different levels are merged before another step of NMS, where the IoU threshold T_n is 0.7.

B Training-free Version of Crowd-SAM

We devise a training-free version of Crowd-SAM as a complementary method. In this version, no learnable parameter is introduced and all computation uses only the pre-trained foundation models. We find that it also shows promising performance.

Given a query image I_0, several supporting images $I_1, I_2, ...I_n$ and their masks $M_1, M_2, ..., M_n$, we first extract the features of supporting images with an image encoder and obtain $F_1, F_2, ..., F_n$. Then, we compute the masked feature F'_n by multiplying the mask M_n with F_n. This step filters out the background part of the extracted features. These masked features are aggregated into a single prototype P encoding the semantic of pedestrians. This prototype can be computed and cached offline.

During inference, we compute the cosine similarity between the embedding of query image $F_0 \in \mathbb{R}^{h\times w\times c}$ and $P \in \mathbb{R}^{1\times c}$, where h, w, c are the height, width, and channel of the feature map, respectively. The result is a heatmap $H \in \mathbb{R}^{h\times w}$. Then, the foreground mask is calculated as $H' = H > s_p$, where s_p is a similarity threshold set to 0.15 in our experiments. H' is converted to point prompts and then decoded with EPS.

For mask selection, we adopt a heuristics-based rule that selects the mask with a maximum area. We find that this rule performs better than the single-output mode and the maximum-IoU rule. We use a joint confidence score by (i) gathering the average similarity score of each mask and (ii) multiplying it with the predicted IoU score. The quantitative results are shown in Tab. II, where we can see that training-free Crowd-SAM still outperforms its counterpart, *i.e.* Matcher [5].

Table II: Comparison results (%) of training-free few-shot detectors. TF-Crowd-SAM means the training-free version of Crowd-SAM. Experiments are conducted on the CrowdHuman [9] *val* set. Multi-cropping is not used for fair comparison.

Method	#Shot	AP	Recall	MR
Matcher [5]	1	8.0	23.9	88.9
TF-Crowd-SAM	1	52.4	65.4	91.6

C Additional Experiments on Few-shot settings

Influence of #Shot. We conduct comprehensive analysis on the influence of supporting images, with results presented in Tab. IV. Generally, increasing the number of training images leads to improved results, although the performance tends to plateau when the number of supporting images exceeds 20. Notably, we observe that with very limited training images, such as 1 and 5 shots, there is a significant increase in the standard deviation of the average precision (AP), indicating an instability in its training. We assume that this phenomenon arises because our training procedure still relies on high-quality annotated images for effective training.

Table III: Comparative results (%) with other label-efficient methods on COCO *val*. nAP represents novel AP.

Methods	Backbone	AP	nAP
STAC [10]	ResNet50-FPN	9.8	-
UB-Teacherv2 [6]	ResNet50-FPN	21.3	-
Ours (v1)	ViT-L	**23.0**	33.0
TFA [13]	ResNet101-FPN	28.7	10.0
De-FRCN [8]	ResNet101	33.9	18.5
Ours (v2)	ViT-L	22.0	**25.0**

Comparison on COCO. We conduct simple comparative studies on COCO with two versions of Crowd-SAM. The first (v1) adopts a trainable head that utilizes 0.5 % percent of labeled data in the COCO *trainval* set and the other (v2) employs a prototype-based classification head that uses the prototypes extracted by De-ViT [17]. The results are shown in Tab. III. As it can be seen, our method derives competitive results in both the settings by leading Unbiased Teacher by 1.7% AP and De-FRCN by 6.5% nAP.

Table IV: Results (%) on CrowdHuman [9] with different numbers of supporting images. No multi-cropping is used.

#Shot	1	5	10	20	30
AP	51.3 ± 7.05	68.1 ± 2.6	69.1 ± 0.62	71.6 ± 0.67	72.5 ± 0.78

D Dicussion and Limitation

D.1 Discussion

Parameter-efficiency. Thanks to the exceptional representational capabilities of vision foundation models, we require only a minimal number of parameters to tailor it for pedestrian detection tasks. We highlight that our model introduces only 0.8M learnable parameters within the segmentation head and PWD-Net. This is remarkably lightweight compared to the 304M parameters of the ViT-L/14 backbone. Consequently, Crowd-SAM can be regarded as a parameter-efficient transfer learning approach, encompassing techniques such as LoRA [1], TIP-Adapter [15], and prompt-tuning [2]. This characteristic facilitates seamless transfer across diverse domains and datasets with a minimal overhead.

Table V: Comparison of efficiency to existing state-of-the-art methods on CrowdHuman *val*. All the experiments are conducted on a 3090Ti GPU. * represents applying multi-cropping. All the rows use ViT-L as base models except the ones with $^+$, where SAM (ViT-B) and DINOv2 (ViT-S) are employed

Methods	Foundation Models	*AP*	Secs/Img
Iter-SRCNN [18]	-	85.9	0.11
Matcher [5]	SAM [4] + DINOv2 [7]	8.0	22.0
Crowd-SAM	SAM [4] + DINOv2 [7]	71.4	1.7
Crowd-SAM	HQ-SAM [3] + DINOv2 [7]	64.6	2.4
Crowd-SAM$^+$	Mobile-SAM [14] + DINOv2 [7]	33.0	0.7
Crowd-SAM*	SAM [4] + DINOv2 [7]	78.4	8.1

Extensibility. Remarkably, our method exhibits robust performance in pedestrian detection. Since our method focuses on crowded scenes with occlusions appearing, it can also be employed for other tasks that involve similar issues like remote sensing, vehicle detection, fruit counting and *etc*. On the other hand, handling multi-class few-shot object detection poses greater challenges compared to binary classification tasks. Nevertheless, extending Crowd-SAM to accommodate multi-class classification is relatively straightforward. Approaches such as constructing class prototypes [5,17] or employing transfer learning techniques [11,13] can be effectively leveraged. Additionally, given the impressive k-NN classification performance demonstrated by DINOv2, a k-NN classifier may suffice for many scenarios. However, we defer the exploration of this topic to future work.

As for the SAM version, we apply our adaptation to multiple SAM-variants [3, 14] to validate its extensibility. The performance and efficiency (Secs/Img) are reported in Tab. V. It is noteworthy that HQ-SAM [3] trains only the best-fitting mask in a single granularity, *e.g.* the whole object, and cannot improve the overall quality of mask predictions in any granularity. Moreover, considering the data bias in HQ-SAM, mask predictions at a certain granularity could be even worse, *e.g.* the part-level predictions. Unfortunately, in our method, all mask predictions are re-evaluated regarding their IoU scores and the mask at any level of granularity could be chosen for outputs. Thus, this conflict of training goals incurs a decline in AP.

Human interaction. It is notable that Crowd-SAM not only supports automatic annotation but also allows humans to interact with it. For example, human annotators can verify the results produced by Crowd-SAM or interact with it by adjusting the prompts or adding new prompts. As we maintain the original SAM frozen, these adjustments can be made using the same model. We argue that this verification process is still much faster than annotating a new object.

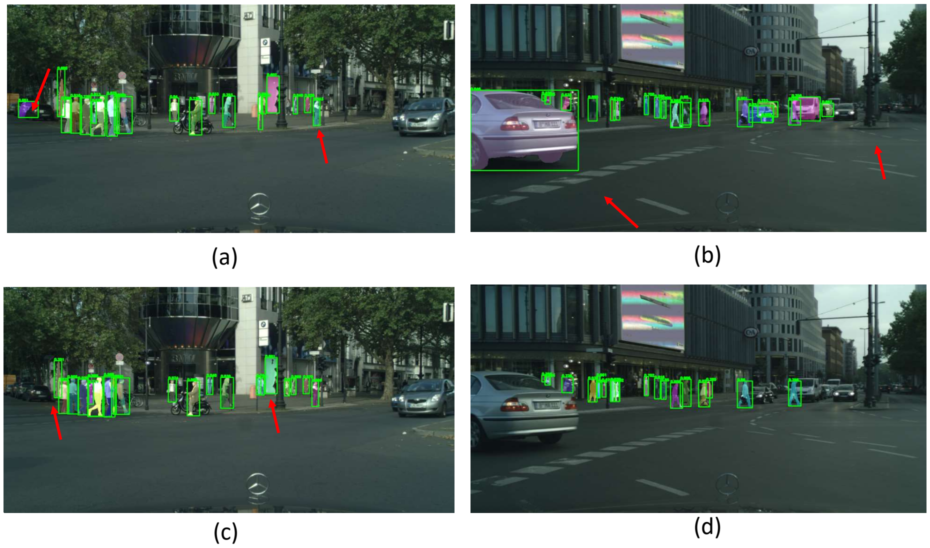

Fig. 1: Qualitative comparison of Crowd-SAM tuned on CrowdHuman (*a, b*) and CityPersons (*c, d*). False positives are pointed out with (*red arrows*).

D.2 Limitation

Semantic ambiguity. Despite its strong results, Crowd-SAM still faces a limitation in being confused by unseen backgrounds, which stems from insufficient representation of background samples. Occasionally, this leads to misclassification of objects from background categories as foreground. To illustrate this issue, we compare the visualization results of two models in Fig. 1. In panels (a) and (b), cars are assigned high confidence scores by the model trained on CrowdHuman, likely due to the rarity of cars in the CrowdHuman dataset. We propose a potential method to address this limitation by augmenting the dataset with more background samples, as outlined in Per-SAM [16]. This can be achieved by selecting potential backgrounds using a prototype or a fine-tuned classification model, thereby enhancing the robustness of the model. This method is in our scope for the next work.

Lack of efficiency. Another limitation of our model lies in its reliance on two foundation models, *i.e.* DINOv2 [7] and SAM [4], to achieve promising results, which introduces computational overhead in image encoding, as it necessitates forwarding each image twice. This incurs an extra 0.3-0.4s latency on a 3090Ti GPU card with a ViT-L/14 backbone. Unfortunately, there is currently no cost-effective solution to seamlessly integrate these two models without compromising their effectiveness. A closely related approach, SAM-CLIP [12], attempts to distill features from CLIP into SAM. However, their method relies on extensive data for training, which is not suitable for our objective of lightweight adaptation.

Furthermore, the scale variability of objects in real scenes continues to pose a challenge for SAM and SAM-based methodologies, including Crowd-SAM. At present, we tackle this issue by employing multi-cropping, albeit at the expense of a 4x or 5x increase in inference time. But we must highlight that our method without multi-cropping is still faster and more accurate than Matcher [5] which employs expensive computations in matching and clustering. Despite exploring alternative approaches such as feature-level cropping, none have yielded satisfactory results. Hopefully, our method without multi-cropping still retains strong and promising results. We will work for a faster and more efficient version of Crowd-SAM in the future by investigating additional prompt types, such as boxes.

References

1. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: Int. Conf. Learn. Represent. (2022)
2. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: Eur. Conf. Comput. Vis. pp. 709–727. Springer (2022)
3. 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)
4. 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)
5. 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)

6. 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)
7. 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)
8. 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)
9. 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)
10. Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y., Pfister, T.: A simple semi-supervised learning framework for object detection. In: arXiv:2005.04757 (2020)
11. 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)
12. Wang, H., Vasu, P.K.A., Faghri, F., Vemulapalli, R., Farajtabar, M., Mehta, S., Rastegari, M., Tuzel, O., Pouransari, H.: Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3635–3647 (2024)
13. Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly simple few-shot object detection. Int. Conf. Mach. Learn. (2020)
14. 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)
15. Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. In: Eur. Conf. Comput. Vis. (2022)
16. 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)
17. Zhang, X., Wang, Y., Boularias, A.: Detect every thing with few examples. arXiv preprint arXiv:2309.12969 (2023)
18. 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)