# Point, Segment and Count: A Generalized Framework for Object Counting

## Supplementary Material

This supplementary material provides the following extra content:

1. Visual illustration of proposed hierarchical knowledge distillation in Fig. 6;
2. Results of object counting in terms of NAE and SRE in Tab. 6;
3. Failure cases of proposed PseCo in Fig. 7.
4. Ablation on the computation costs in Sec. 6.

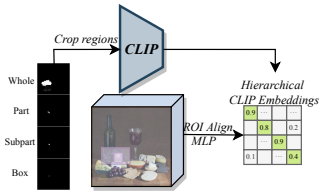


Figure 6. Illustration of the proposed hierarchical knowledge distillation. The hierarchical mask proposals from SAM of the same points are cropped from the input image, which are then fed into CLIP. The CLIP embeddings are used as classification weights to discriminate the classifier, which encourages discriminative classification among hierarchical proposals.

| | Val set | | Test set | |
|---|---|---|---|---|
| | NAE$_\downarrow$ | SRE$_\downarrow$ | NAE$_\downarrow$ | SRE$_\downarrow$ |
| SAM-Free [34] | - | - | 0.29 | 3.80 |
| C-DETR [28] | - | - | 0.19 | 5.23 |
| **Ours** | **0.22** | **2.99** | **0.19** | **3.12** |
| ZSOC [38] | 0.36 | **4.26** | 0.34 | 3.74 |
| SAM-Free [34] | - | - | 0.37 | 4.52 |
| **Ours** | **0.32** | 4.31 | **0.25** | **3.46** |

Table 6. Results on the few-shot (first part) and zero-shot (second part) object counting. We reported the results on FSC-147 in terms of Normalized Relative Error (NAE) and Squared Relative Error (SRE). NAE $= \frac{1}{N}\sum_{j=1}^{N}\frac{|\hat{y}_i - y_i|}{y_i}$ and SRE $= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{(\hat{y}_i - y_i)^2}{y_i}}$, where $y_i$ and $\hat{y}_i$ are GT and predicted counts. ZSOC belongs to density-based methods. Our proposed method achieves state-of-the-art methods on the two metrics.

## 6. Ablation on the computation costs.

It is acknowledged that our method is slower than traditional two-stage object detection or object counting methods if using the same backbone. The computation costs of our method mainly lie in the frequent inference of the mask decoder of SAM. However, compared to vanilla SAM in [34] that employs $32 \times 32$ grid point prompts, the proposed class-agnostic localization significantly reduces the computation costs. Specifically, there are only an average of 378 and 388 candidate points for each image in the FSC-147 test and val sets. It is worthwhile to note that these points are selected from $256 \times 256$ heatmaps, 64 times than $32 \times 32$ grid points. In addition, the point decoder shares the same architecture as the mask decoder of SAM and only needs 1 inference for each image.

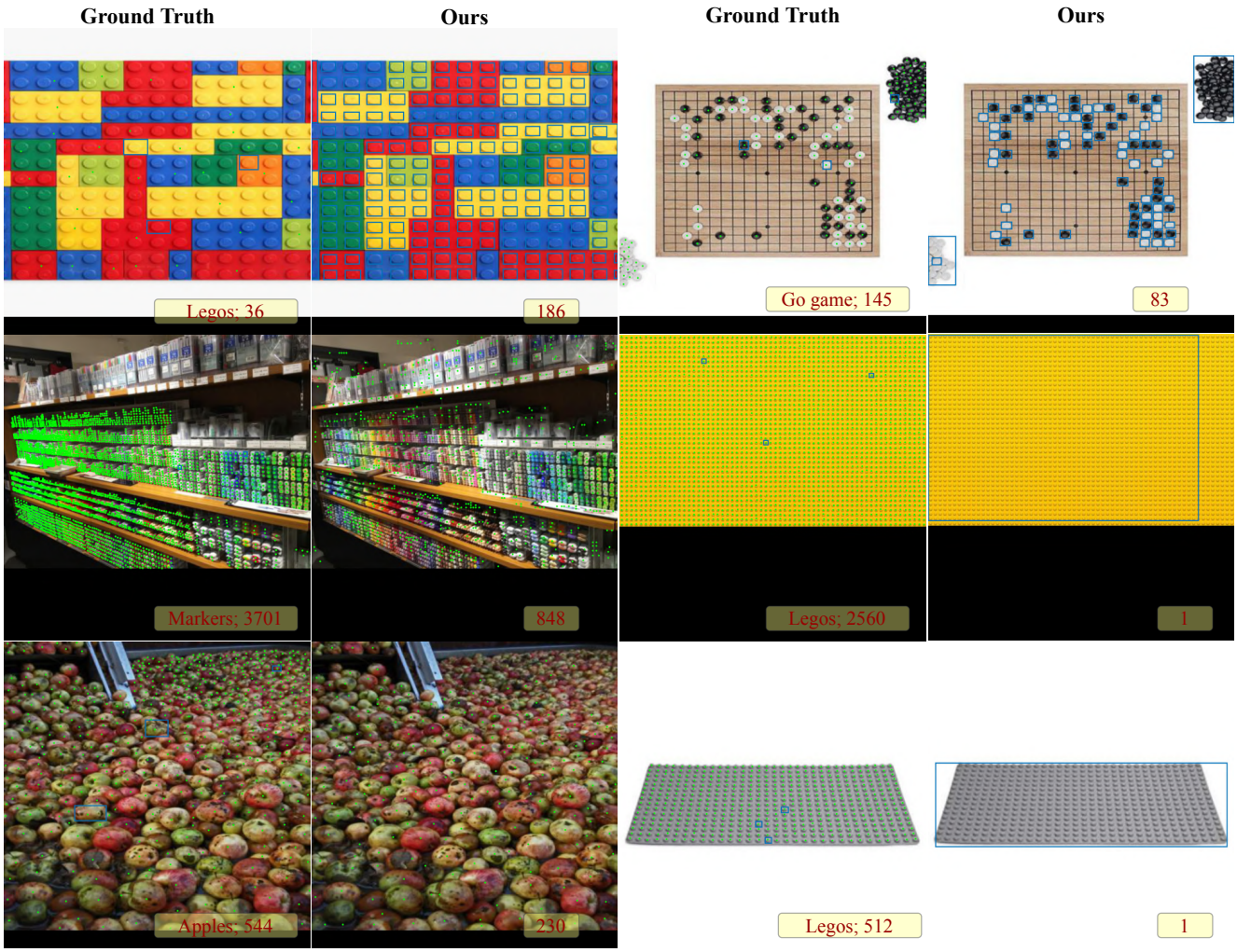


Figure 7. Failure cases for (a) few-shot and (b) zero-shot object counting and detection. Under the few-shot setting, PseCo may fail to identify target objects given wrong example images, and cannot detect objects on extremely crowded scenes. A similar phenomenon has been observed under the zero-shot setting. The imprecise text prompts, *e.g.*, 'go game' and 'legos' cannot be used to distinguish the objects.