Zero-shot Object Counting with Good Exemplars
-Supplementary Materials-

Anonymous ECCV 2024 Submission

Paper ID #812

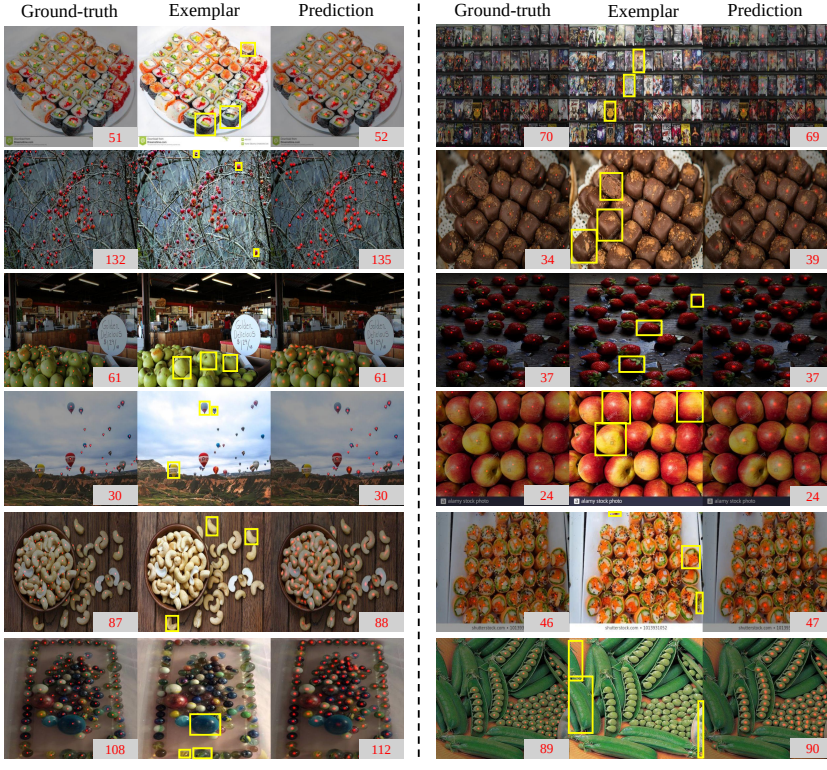

Fig. 1: Illustration of the found exemplars for images on the FSC-147 dataset, along with the density maps.

1 Overview

- Analysis of Density Maps (Sec. 2)
- Analysis of Negative Sample Density Maps (Sec. 3)
- Analysis of Positive and Negative Samples (Sec. 4)
- Ablation Study on IoU Threshold (Sec. 5)
- Ablation Study on Thresholds for Grounding DINO (Sec. 6)
- Limitation (Sec. 7)
- Conclusion (Sec. 8)

2 Analysis of Density Maps

Fig. 1 demonstrates the efficacy of VA-Count in generating density maps, where it is evident that our methodology yields estimations closely aligned with ground-truth densities across a spectrum of scenarios: handling of irregularly shaped objects (first and fifth rows), navigation through complex environmental backgrounds (images two, three, and four from the left), and accurate depiction of densely clustered objects (images two, three, and four from the right). The exemplars utilized are of exceptional quality. Notably, even in scenarios with significant object scale variability, as depicted in the lower left image, the algorithm successfully approximates true density values. Moreover, the robustness of VA-Count is highlighted in the rightmost sixth image, where despite the selection of exemplars with minor inaccuracies, the density map produced is of high fidelity. This illustrates VA-Count's advanced capability to learn and maintain the intrinsic correlation between the exemplars and the original images, ensuring that minor selection errors in exemplars have negligible impact on the overall density estimation accuracy.

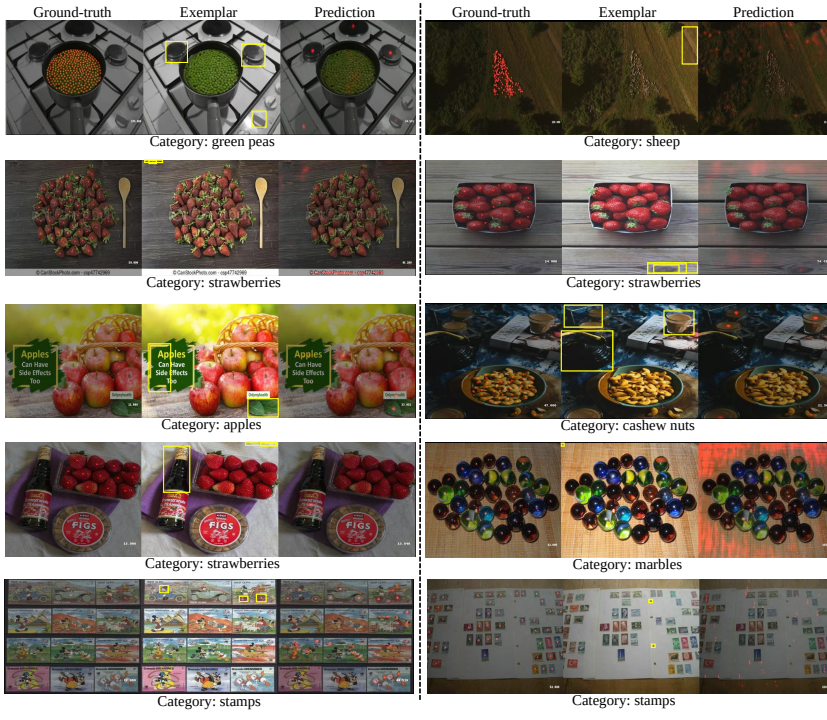

Fig. 2: Illustration of the final negative exemplars for images on the FSC-147 dataset, along with the density maps.

3 Analysis of Negative Sample Density Maps

Fig. 2 shows the negative exemplar and the corresponding density map display. The figure demonstrates that when the exemplar is not a sample of the corresponding category, it will not find the specified category, but instead will locate the area corresponding to the negative exemplar and generate a density map. When objects belonging to different categories are present within an image (as observed in positions left 1, left 4, left 5, and right 3), density maps specific to those categories are produced. Conversely, in scenarios devoid of distinguishable objects, where only the background is visible, the generated density maps correlate directly with the designated regions.

Fig. 3: Illustration of the positive (Pos.) and negative (Neg.) exemplars for images on FSC-147 dataset.

4 Analysis of Positive and Negative Samples

Fig. 3 illustrates the selection process for positive and negative samples. From the figure, it is evident that our method identifies positive samples as individual objects of the specified category, performing well not only for regular objects but also for items like nail polish, sunglasses, and stamps. In selecting negative samples, when objects of other categories are present in the image, our method can identify these objects as negative samples (as seen in left 2, left 3, right 2, right 3, and right 4). This demonstrates that VA-Count not only selects high-quality positive exemplars but also effectively avoids positive samples while selecting potentially confusing objects as negative samples.

5 Ablation Study on IoU Threshold

In this paper, the Intersection over Union (IoU) threshold plays a critical role in determining the quality of negative sample selection. Tab. 1 illustrates the influence of varying IoU thresholds on the accuracy of object counting, presenting data for the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) across both the validation and test datasets. Notably, the MAE demonstrates a non-linear trend, initially rising before diminishing, with the optimal performance observed at an IoU threshold of 0.5. In contrast, the RMSE experiences fluctuations, attributable to the varying quality of density maps influenced by the selection of negative samples. Such variations in density map quality introduce a stochastic element to the errors, thereby causing the observed fluctuations in RMSE.

Table 1: Ablation study on the contribution of the IoU threshold τ_{iou} for negative sample selection to the final results on the FSC-147 dataset. We present the MAE and RMSE across the validation and test sets for thresholds ranging from 0.1 to 0.9, as well as their average performance. The best results are highlighted in **bold**, and the second-best are underlined.

τ_{iou}	Val Set		Test Set		Avg	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
0.1	18.83	**72.26**	20.27	130.39	19.55	101.33
0.2	18.56	77.01	18.73	125.83	18.64	101.42
0.3	19.89	77.23	18.52	**125.41**	19.20	101.32
0.4	18.26	75.61	**17.54**	127.47	17.90	101.54
0.5	**17.87**	73.22	17.88	129.31	**17.87**	**101.26**
0.6	18.55	73.90	19.10	129.32	18.82	101.61
0.7	18.97	74.91	18.31	128.78	18.64	101.85
0.8	21.28	74.51	20.52	128.00	20.90	**101.26**
0.9	22.30	74.48	20.96	128.31	21.63	101.40

6 Ablation Study on Thresholds for Grounding DINO

In this study, the selection of logits thresholds for Grounding DINO is identified as a pivotal factor in curating exemplars. Excessively high thresholds hinder the selection of samples for more challenging categories, while excessively low thresholds not only escalate computational demands but also result in an abundance of superfluous samples. To address this, we conducted the experiments detailed in Tab. 2. At a threshold of 0.01, the inclusion of suboptimal exemplars significantly elevates the RMSE. Conversely, setting the threshold at 0.05 leads to a considerable overall error, as it precludes the selection of category-specific exemplars in certain images. The thresholds of 0.02, 0.03, and 0.04 exhibit comparatively lower MAE and RMSE values, with the optimal error minimization achieved at a threshold of 0.02. This nuanced approach underscores the importance of a balanced threshold setting in enhancing the efficacy of exemplar selection within the Grounding DINO framework.

Table 2: Ablation study on the contribution of the grounding DINO threshold for sample selection to the final results on the FSC-147 dataset. We present the MAE and RMSE across the validation and test sets for Logits thresholds τ_l ranging from 0.01 to 0.05, as well as their average performance. The best results are highlighted in **bold**, and the second-best are underlined.

τ_l	Val Set		Test Set		Avg	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
0.01	27.36	76.41	27.10	129.98	27.23	103.20
0.02	**17.87**	**73.22**	**17.88**	129.31	**17.87**	**101.26**
0.03	19.74	77.06	18.25	129.77	18.99	103.42
0.04	22.84	76.26	20.26	**128.69**	21.55	102.48
0.05	25.60	86.45	21.25	130.79	23.43	108.62

7 Limitation

To delve into the limitations of VA-Count, Fig. 4 showcases images with notable inaccuracies, highlighting three primary constraints in the algorithm's efficacy. Firstly, there is the challenge of background noise. Despite the strategic use of negative samples to mitigate errors from non-object classes, the algorithm remains excessively responsive to clear objects (first row). Secondly, the issue of density map numerical uncertainty is evident. As illustrated in the second row, despite both images having a mere count error of 1, the quality of their density maps is suboptimal. Specifically, the left image poorly locates a larger object in the foreground, while the right image incorrectly identifies two points of focus for a single pair of sunglasses, diverging from the ground-truth which associates one

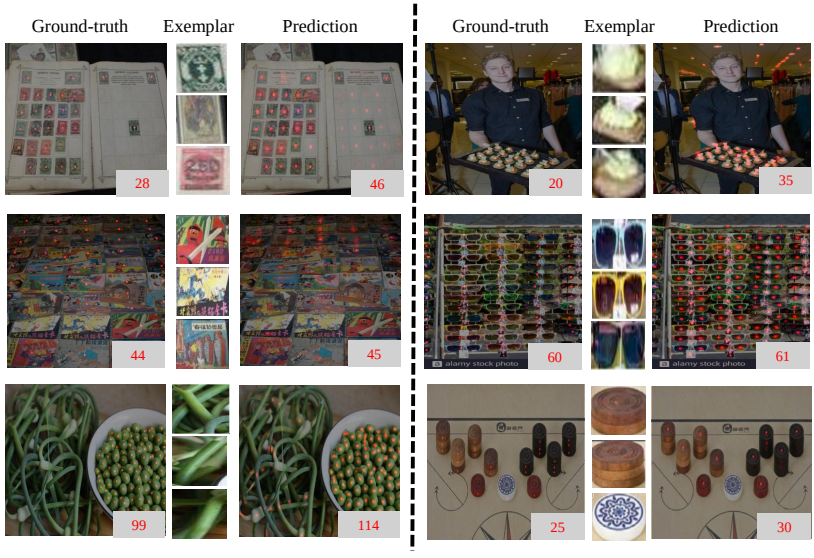

Fig. 4: Illustration of the error density map on the FSC-147 dataset.

focal point per pair of sunglasses. Lastly, exemplar inaccuracies persist. While our method achieves exemplar identification quality on par with annotated bounding boxes in most images, some discrepancies remain. For instance, as depicted on the left, entire strings of peas are mistakenly identified as exemplars, and on the right, stacked items, not individual objects due to their blurred edges, are erroneously treated as singular targets. These limitations represent key areas for our ongoing and future refinement efforts.

8 Conclusion

In Sec. 2 and Sec. 3, Fig. 1 and Fig. 2 showcase the density maps derived by integrating positive and negative samples with the original images, thereby substantiating the visual associations VA-Count has deciphered between exemplars and images. Sec. 4, elucidated through Fig. 3, delineates the meticulous process of selecting positive and negative samples, thereby underscoring the precision of VA-Count's sample selection mechanism. In Sec. 5 and Sec. 6, Tab. 1 and Tab. 2 provide empirical evidence supporting the judicious selection of threshold values within our methodology. Lastly, Sec. 7, as illustrated in Fig. 4, delves into the current constraints of VA-Count, pinpointing critical domains necessitating future advancements and methodological refinements.