Enhancing Zero-shot Object Counting via Text-guided Local Ranking and Number-evoked Global Attention

Supplementary Material

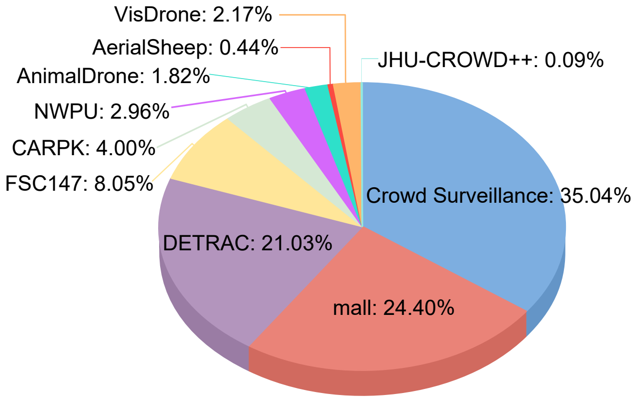

Figure 7. Proportions of each dataset in our ZSC-8K.

	#Images	#Classes	Avg	Total
SHA	482	1	501	241677
JHU	4372	1	346	1515005
NWPU	5109	1	418	2133238
MALL	2000	1	31	62325
CARPK	1448	1	62	89777
DETRAC	14000	1	86	1210000
ZSC-8K	8361	114	46	386498

Table 7. Comparisons with class-specific datasets.

	Val set		Test set	
	MAE	RMSE	MAE	RMSE
all patch features	12.31	62.81	11.60	107.31
avgp	**11.09**	**60.48**	**10.74**	**106.5**

Table 8. Ablation study on the global feature."avgp" means average pooling.

7. Dataset statistics

Images of ZSC-8K dataset are collected from multiple existing datasets: FSC-147 [38], JHU-Crowd [43], NWPU-Crowd [46], VisDrone [57], DETRAC [47], CARPK [15], MALL [7] and Crowd Surveillance [50]. The proportions of each dataset are depicted in Fig. 7.

Besides, in our proposed ZSC-8K dataset, the number of point annotations in an image ranges from 1 to 1,766. The histogram illustrated in Fig. 8 shows that the number of images which have 20 to 50 points annotations is the largest. And the ZSC-8K dataset is compared to other datasets in Tab. 7.

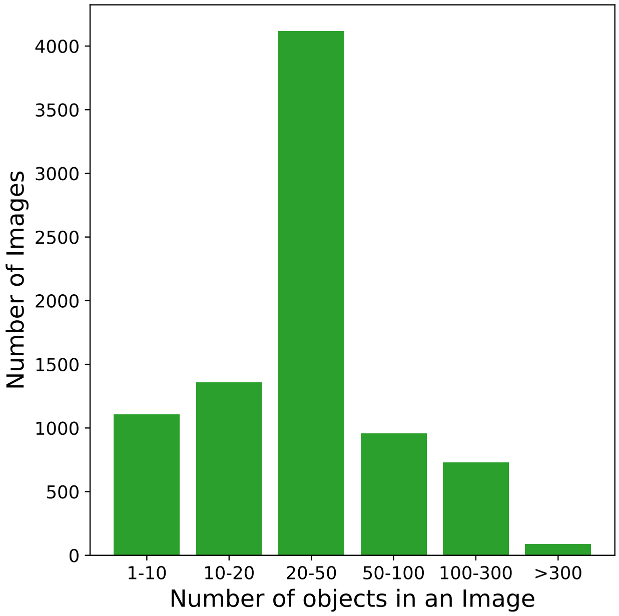

Figure 8. Number of images in several ranges of object count.

Splits	Val set		Test set	
	MAE	RMSE	MAE	RMSE
6 coarse→ 6 fine	18.21	60.21	16.46	101.35
5 coarse→ 5 fine	**17.52**	**58.39**	**16.42**	**100.63**
4 coarse→ 4 fine	18.35	60.70	16.66	101.05

Table 9. Performance of using different number ranges.

8. Ablation Study

The global feature in GDino-based methods. Because the selected number-evoked text prompt represents global information, global image-level feature is used to make cross-attention with it. Since the Swin-T used in GDino does not have a token can represent global feature, we try two ways to construct a global feature. Firstly, we use all patch-level features directly to make cross-attention with number-evoked text prompt. Then, we use all patch-level features to construct a global feature by average pooling. Tab. 8 shows that the global feature generated by average pooling achieves better performance. Since the number-evoked text prompt does not match each of the patch-level features.

Number ranges. The number ranges (5 coarse → 5 fine splits) are used for all datasets. Moreover, an ablation study is conducted on the number ranges with FSC in Tab. 9. It

Method	<10		10-100		100-1000		>1000	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
ClipCount	7.14	13.67	10.88	46.15	33.94	52.20	1718.42	2294.33
ClipCount+Ours	5.89	10.17	9.60	37.72	32.54	51.02	1706.89	2268.71

Table 10. The performance across different count magnitudes.

Method	Training→ Testing	MAE	RMSE
RCC [13]	FSC→CARPK	21.38	26.15
ClipCount [18]	FSC→CARPK	11.96	16.61
VLCounter [20]	FSC→CARPK	8.68	10.37
CountGD* [3]	FSC→CARPK	3.94	5.23
ClipCount+Ours	FSC→CARPK	10.13	13.28
VLCounter+Ours	FSC→CARPK	8.08	9.64
CountGD+Ours	FSC→CARPK	**3.71**	**4.96**

Table 11. Cross-dataset evaluation on CARPK dataset."FSC" means FSC-147 dataset. * means the results are reproduced by open-source codes.

shows that our design of number ranges (5 coarse → 5 fine splits) is optimal. Indeed, both 6 splits and 4 splits still work, which verifies the efficiency of the NGA.

9. Analysis

Performance across different count magnitudes. Experiments are conducted on the test set of FSC dataset to analyze the performance across different count magnitudes in Tab. 10. Our method outperforms the baseline on all 4 splits, which verifies that our proposed NGA can enhance the counting ability of VLMs. Besides, the experimental results also show that counting models face challenges when dealing with high-count images (>1000).

The generalizing ability on CARPK dataset. We make a comparison between our method and other ZSC methods on CARPK [15] in Tab. 11. Specifically, we apply our strategy to three baselines: ClipCount [18], VLCounter [20] and CounGD [3]. The results show that our strategy largely improve the baselines. On ClipCount, we gain 1.83 MAE and 3.33 RMSE improvement. On VLCounter, we obtain 0.6 and 0.73 points improvement on MAE and RMSE respectively. Furthermore, we achieve the best performance with CountGD which shows 3.71 MAE and 4.96 RMSE.

10. Details of cross-attention.

We use the global image-level feature as query and the number-evoked text prompt with highest similarity score as key and value. It can be roughly represented as:

$$F_{ca} = \mathrm{proj\_drop}(\mathrm{proj}(\mathrm{attn\_drop}(\mathrm{softmax}(\frac{q \cdot k^T}{\sqrt{C/H}}) \cdot v))), \tag{8}$$

where $q = w_q(F_g)$,$k = w_k(F^i_{ct})$,$v = w_v(F^i_{ct})$, w_q, w_k and w_v are learnable parameters. F_{ca} is the output features. And by following the above steps, we can guarantee that the cross-attention in our work is useful.

11. Visualizations

To further demonstrate the effectiveness of our proposed strategy. We visualize some results on the ZSC-8K dataset in Fig. 9 and some results on FSC-147 dataset in Fig. 10. As the two figures show, the results in the bottom rows of the two figures are better than the results in the top rows. The visualizations further verify that our strategy is effective. Besides, in Fig. 11, we present some failure cases. It demonstrates that our method may face challenges when dealing with scenarios involving high counts.

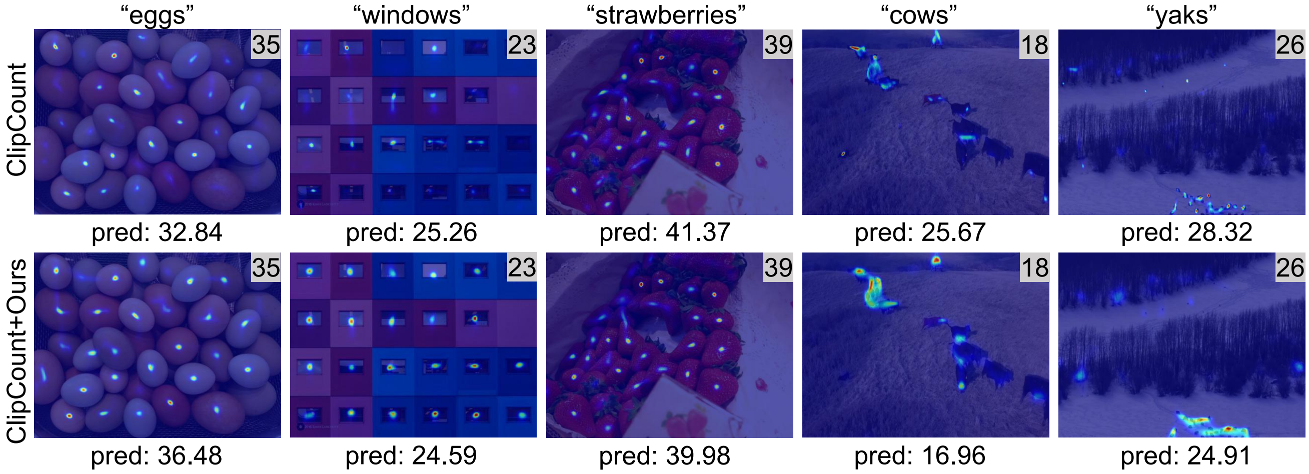

Figure 9. Qualitative results on our ZSC-8K Dataset. Ground-truth numbers are shown at the top-right corner.

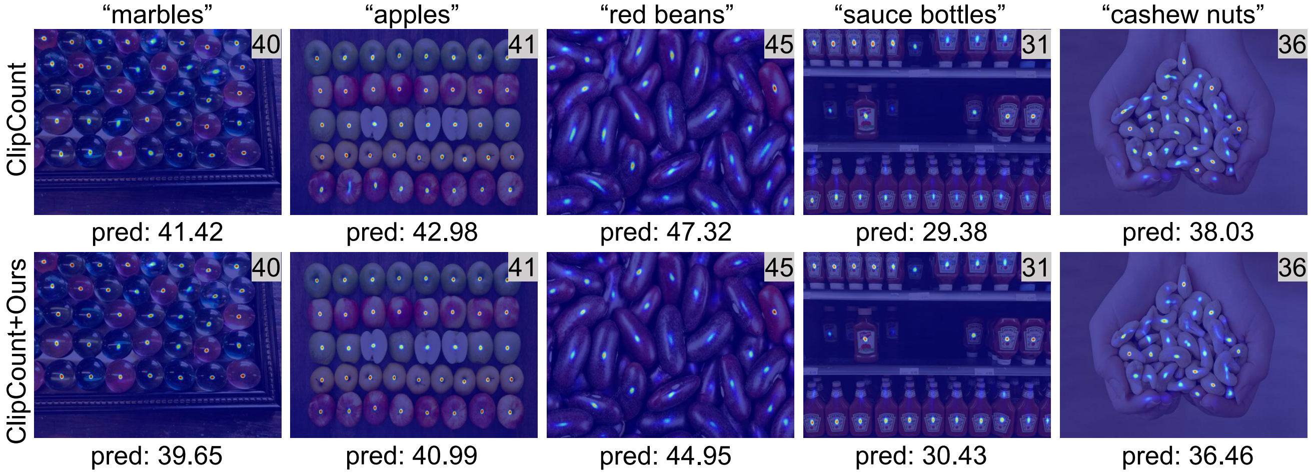

Figure 10. Qualitative results on FSC-147 Dataset. Ground-truth numbers are shown at the top-right corner.

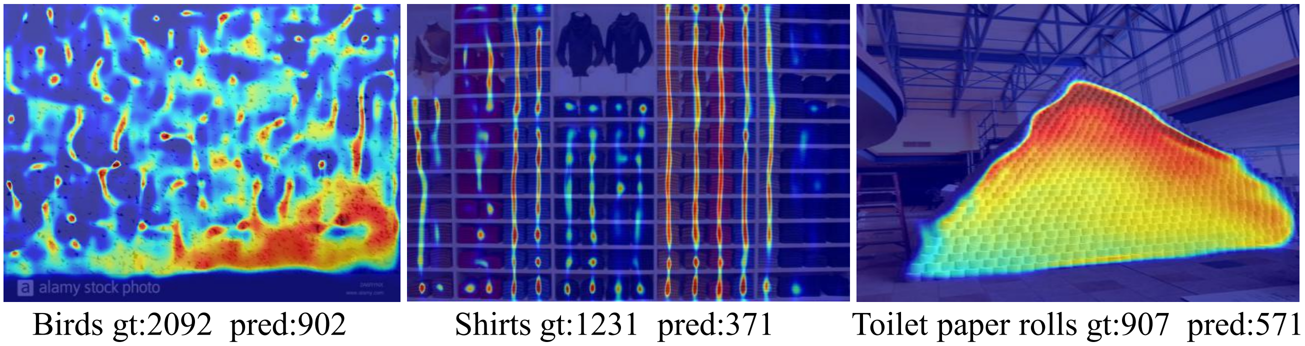

Figure 11. Failure cases.