Zero-Shot Object Counting
Supplementary Material

Jingyi Xu[1], Hieu Le[2], Vu Nguyen[1], Viresh Ranjan[*3], and Dimitris Samaras[1]

[1]Stony Brook University [2]EPFL [3]Amazon

1. Overview

In this document, we provide additional experiments and analyses. In particular:

- Section 2 provides additional visualizations of our selected patches in multi-class cases.
- Section 3 provides the performance of our method when using different sets of candidate patches to select exemplars.
- Section 4 provides the results of selecting exemplars from the class-relevant patches based on the objectness score.
- Section 5 provides qualitative comparisons when using RPN proposals as counting exemplars.
- Section 6 compares our proposed patch selection method with directly using the generated prototype to perform correlation matching to get the similarity map.

2. Multi-class Zero-shot Counting

Figure 1 provides additional visualizations of the selected patches in multi-class cases. As can be seen from the figure, our proposed method can select counting exemplars according to the given class name and count instances from that specific class in the input image.

3. Different Methods to Acquire Candidate Patches

In our main experiments, the candidate patches for selection are randomly sampled from the input image. In this section, we conduct an ablation study on how to obtain the candidate patches. Instead of using randomly sampled patches, we take the proposals generated by RPN as the candidate patches and apply our patch selection method. We further combine the random patches and RPN generated proposals together as the candidate patches and evaluate the performance. Results are summarized in Table 1. As can be seen from the table, our proposed patch selection method can bring consistent performance improvements for all the three set of candidate patches.

4. Comparing Predicted Errors with Objectness Scores

In our proposed method, after obtaining the class-relevant patches, we select among them the final counting exemplars via an error predictor. To further validate the effectiveness of the error predictor, we compare the performance of our method with a baseline approach which simply uses the objectness score from the RPN to select counting exemplars from the class-relevant patches. Specifically, after obtaining the class-relevant patches, we rank them according to the objectness score and select the patches with the top-3 highest objectness scores as counting exemplars. As shown in Table 2, our method outperforms the baseline using the objectness score in most cases, which shows the effectiveness of our proposed error predictor.

[*]Work done prior to joining Amazon

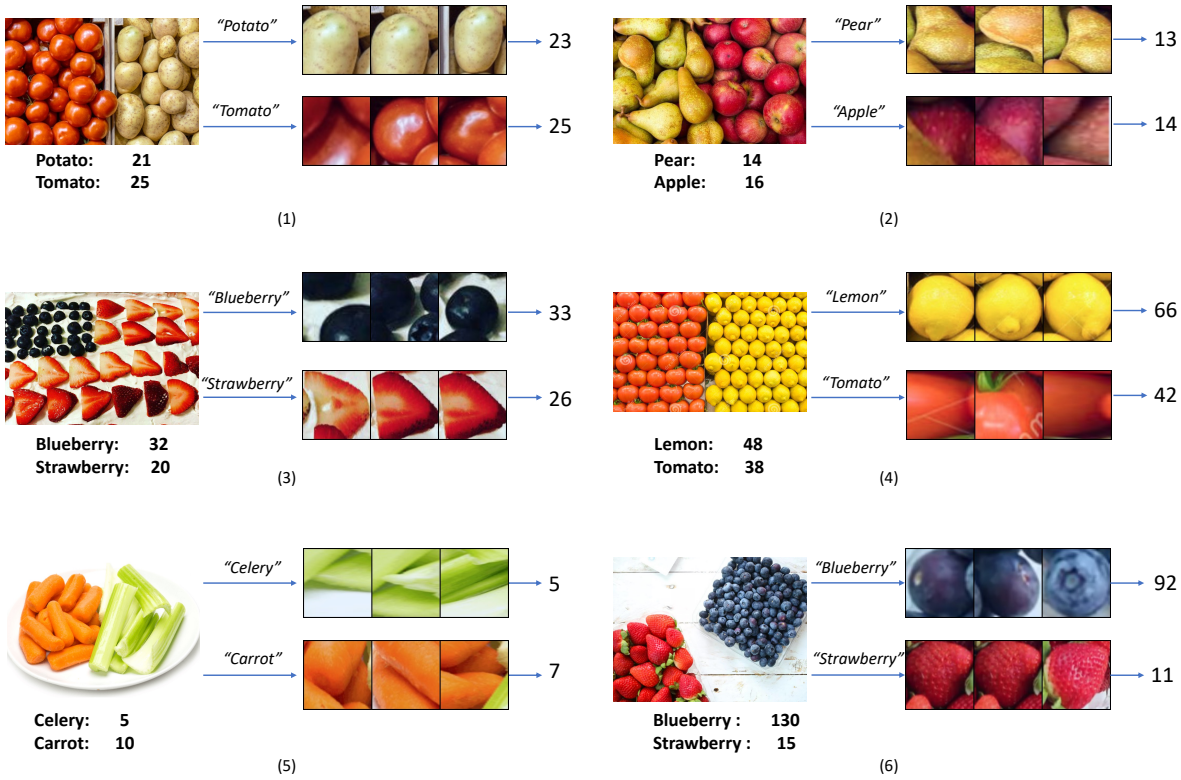

Figure 1. Visualizations of the selected patches: There are two classes with multiple object instances in a single image. To specify what to count, we provide the class name at test time. Our proposed method selects counting exemplars according to the given class name and count instances from the specific class.

Candidate Patches	Patch Selection	Val Set MAE	Val Set RMSE	Val Set NAE	Val Set SRE	Test Set MAE	Test Set RMSE	Test Set NAE	Test Set SRE
RPN	✗	29.64	94.93	0.43	4.96	25.04	126.02	0.36	4.18
	✓	26.76	87.41	**0.34**	4.36	23.52	126.68	0.33	3.94
Random	✗	35.20	106.70	0.61	6.68	31.37	134.98	0.52	5.92
	✓	26.93	88.63	0.36	**4.26**	22.09	**115.17**	0.34	3.74
Random+RPN	✗	29.97	91.61	0.44	5.17	24.91	126.35	0.36	4.45
	✓	**26.58**	**86.69**	0.35	4.28	**22.03**	116.42	**0.33**	**3.65**

Table 1. Performance on FSC-147 dataset when using different sets of candidate patches to do patch selection. Our proposed method brings consistent improvement in the performance.

	Val Set MAE	Val Set RMSE	Val Set NAE	Val Set SRE	Test Set MAE	Test Set RMSE	Test Set NAE	Test Set SRE
Obj Score	28.47	94.87	**0.34**	4.65	24.11	117.76	0.35	4.00
Pred Error	**26.93**	**88.63**	0.36	**4.26**	**22.09**	**115.17**	**0.34**	**3.74**

Table 2. Comparison between using the predicted counting error and the objectness score from RPN to select among class-relevant patches.

5. Qualitative Comparison with RPN

In Figure 2, we visualize some images from the FSC-147 dataset and the corresponding patches selected by our proposed method and RPN, respectively. The RPN-selected patches are the top-3 proposals with the highest objectness scores. As can be seen from the figure, our proposed method can accurately localize image patches according to the given class name. These selected patches can then be used as counting exemplars and yield reasonable counting results. In comparison, the patches selected by RPN might contain objects which are not relevant to the provided class name or contain multiple instances. These patches are not suitable to be used as counting exemplars and will lead to inaccurate counting results. This suggests that choosing counting exemplars based on objectness score is not reliable.

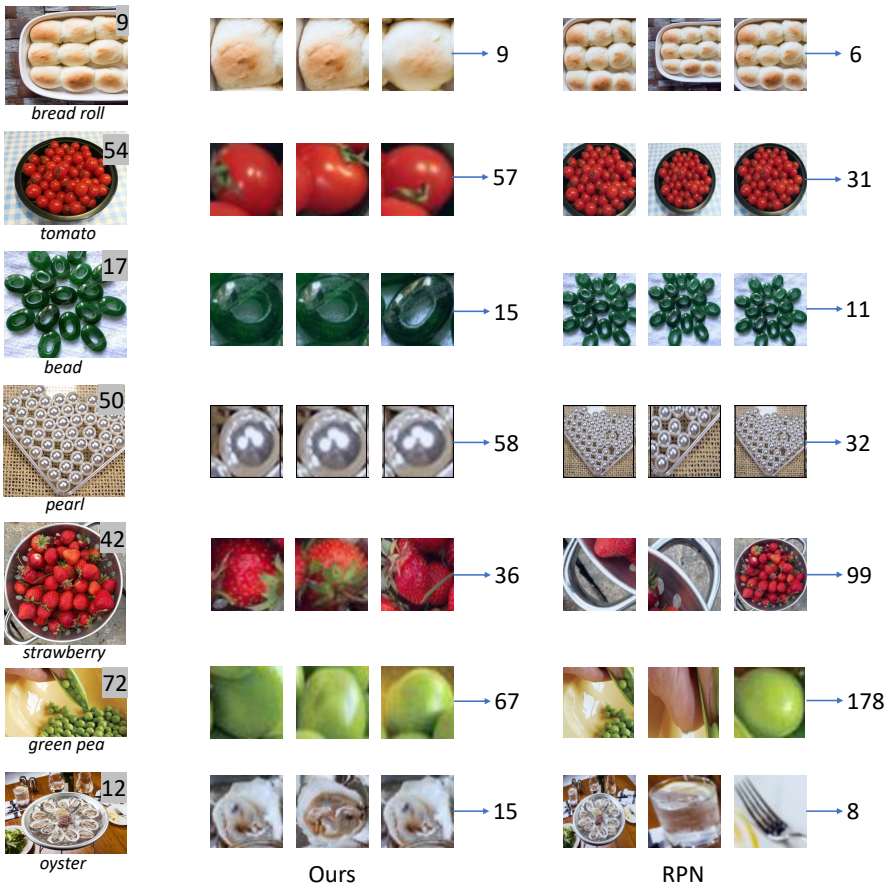

Figure 2. Qualitative comparison with RPN. Our proposed method can select patches suitable for counting while RPN-selected patches contain non-relevant objects or multiple instances.

6. Comparing with Correlation Matching via Prototype

Our strategy for zero-shot counting is to select patches across the query image and use them as exemplars for counting. The patches are selected via a generated class prototype and an error predictor. An alternative way is to use the generated prototype to do correlation matching directly instead of selecting patches from the input image. In this section, we compare the performance of these two strategies and show the advantage of our proposed one. Specifically, we use the generated prototype to do correlation matching with the features of input images to get the similarity map, which will then be given as input to the counter to get the density map and final count. The VAE used to generate the prototype in our main experiments is trained on the MS-COCO detection set. We train another VAE using the FSC-147 training set and report the performance with both VAEs in Table 3. As can be seen, our proposed patch selection method achieves better results. Directly using the generated prototype to perform correlation matching provides a simple solution for zero-shot counting. However, it is not optimal since the same prototype is applied to all the objects from different images. These objects typically exhibit large variability. Our patch selection method, in comparison, selects different exemplars dynamically according to the input image.

Patch selection	Training data for VAE	Val Set MAE	RMSE	NAE	SRE	Test Set MAE	RMSE	NAE	SRE
✗	MS-COCO	48.56	127.93	0.65	6.37	41.33	147.43	0.52	5.53
✗	FSC-147	30.51	101.39	0.41	4.66	28.03	132.34	0.37	4.42
✓	MS-COCO	**27.00**	**87.90**	**0.35**	**4.29**	**22.09**	**115.17**	**0.34**	**3.74**

Table 3. Comparison between our proposed method and the baseline approach of directly using the generated class prototype to do correlation matching.

7. Qualitative Analysis

In Figure 3, we show a few input images and the corresponding patches with the top-3 lowest and highest predicted counting errors. As can be seen from the figure, the patches with the smallest predicted errors are suitable to serve as counting exemplars and output meaningful density maps and accurate counting results. In comparison, the density maps produced by patches with the highest predicted errors fail to highlight the relevant image regions and lead to inaccurate counting results. This suggests that the predicted counting error can effectively indicate the goodness of the counting exemplars.

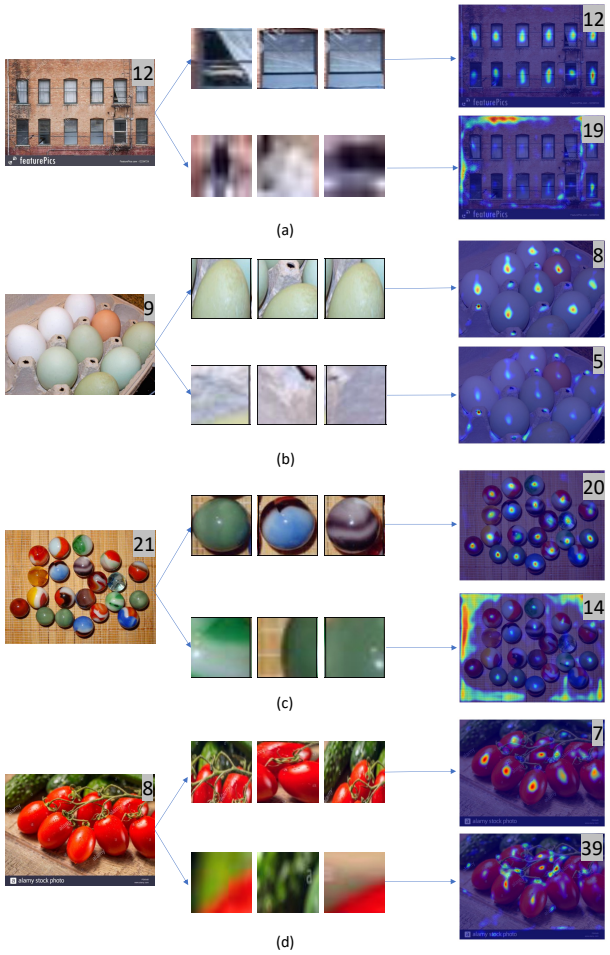

Figure 3. Visualizations of the patches with top-3 lowest and highest predicted counting errors.