Referring Expression Counting

Supplementary Material

1. REC-8K Dataset

Our dataset is collected from various sources as described in the dataset section. The major considerations for the dataset are:

- The dataset should cover both person and object categories with various attributes;
- The dataset should contain images that have diverse scenes and objects and for various application scenarios such as transportation, retail, warehousing, etc.;
- The dataset should contain images from various camera perspectives: surveillance camera, drone, egocentric, etc.;
- The dataset should pose challenges for referring expression counting, i.e. containing similar objects with different attributes;

We define referring expressions for each image and use Amazon Mechanical Turk (AMT) [1] to annotate the images. In order to ensure the quality of the annotations, we select annotators with high approval rates. We also conduct own quality control by converting the annotations to the CVAT [4] format and manually check and revise the annotations on the CVAT platform.

1.1. Data Sources and Splits

In REC-8K, the numbers of images from different data sources are: Crowd Surveillance [10]: 2803, mall [2]: 1952, DETRAC [9]: 1698, FSC147 [6]: 644, CARPK [3]: 320, NWPU [8]: 237, internet: 176, VisDrone [11]: 174, JHU [7]: 7.

Since we have collected the dataset from various sources, we use different split strategies for different sources.

FSC-147 [6], internet, JHU Crowd [7], NWPU [8]. Though the selected images from these sources are not related to each other, our samples are not in terms of images but in terms of Image-RE pairs. So our split strategy is to maximize the novel referring expressions in the validation and test sets, while ensuring there is no data leakage between the training and validation/test sets. We first consolidate Image-RE pairs for each attribute type (e.g. color, action, etc.) and then for each attribute type, we first fill validation and test sets with randomly selected REs and all images with that RE. Then we fill the training set with the remaining REs and images. This way we ensure for each attribute type, there are similar ratios of REs and images in the training, validation, and test sets.

VisDrone [11] contains street views from drones. We select images with street views and closer to the ground and consider attribute types of the person category, including action and location. Since the street scenes are different, we separate each scene into a different split. For example, with the different street layouts, we take different locations of target persons as novel referring expressions for the validation and test sets.

DETRAC [9] contains images from traffic surveillance cameras. We take the object category (e.g. vehicle) as the target category and choose images with a high number of vehicles. We consider the action & direction attributes for this source. Similar to VisDrone, we separate each surveillance camera into a different split. For example, for the same driving direction, we use different road scenes for the validation and test sets.

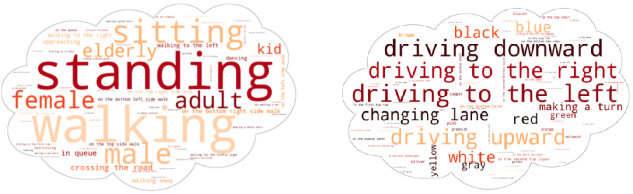
Carpk [3] contains parking lot images from drone perspectives. We also take the object category (e.g. vehicle) as the target and consider the color attribute. We separate each parking lot into a different split.

Mall [2] contains images from surveillance camera in a mall. We take the person category as the target and consider the gender and age attribute types. We randomly split the images into training, validation, and test sets.

Crowd Surveillance [10] contains images from surveillance cameras in various indoor and outdoor locations. We consider the person category as the target and the action attribute type. For each unique action, such as walking, standing, sitting, etc. we randomly select images for the validation and test sets.

1.2. Attributes illustration

We show the attribute frequencies separately for both the person and object categories in REC-8K in Fig. 1. For the person category, the most frequent attributes are coming from the gender, age and action attribute types. Action wise, the most frequent actions are walking, standing, and sitting. For the object category, the most frequent attributes are coming from the color, action, and location attribute types. Highly frequent actions are driving directions for cars. The attribute frequencies show that REC-8K is a diverse dataset with a wide range of attribute types.



(a) The Word Cloud of frequent attributes for the **person** category.

(b) The Word Cloud of frequent attributes for the **object** category.

Figure 1. Word clouds for the most frequent attributes in REC-8K.

2. Results

2.1. Results by attribute type

We show the results by attribute type for the person category in Tab. 1 and for the object category in Tab. 2. We provide the average count for each referring expression in the test set as a reference. It can be seen that for the person category, we have the highest performance for the gender attribute type, following by the action, age, and location attribute types. The lowest performance is for the orientation, clothing and accessory attribute types, which is expected since the orientation type requires more fine-grained localization, and the clothing and accessory types require more detailed object detection.

For the object category, we have the highest performance for the material/color attribute type, following by the orientation and size attribute types. The lowest performance is for the location and variety attribute types. Location is a relational attribute type, which is more challenging to predict. The reason why we have a relatively high performance of the location type for the person category is that many locations are in terms of limited road layouts, which are easier to predict. However, for the object category, there is a large number of unique values for the location attribute type and many more complex location descriptions for object category such as ”book in the top row or leftmost column”. For the variety attribute types, we have a relatively low performance because the number of samples for this attribute type is relatively small.

Table 1. Results by attribute type for **person** category.

type	avg. count	mae	rmse	precision	recall	f1
accessory	45.2	15.9	27.33	0.43	0.43	0.43
action	24.17	7.89	15.16	0.73	0.73	0.73
age	12.05	3.02	4.28	0.67	0.78	0.72
clothing	84.81	40.62	132.8	0.79	0.5	0.61
gender	13.59	2.66	3.47	0.73	0.8	0.76
location	85.88	43.62	91.15	0.63	0.67	0.65
orientation	17.82	16	23.6	0.43	0.77	0.55

Table 2. Results by attribute type for **object** category.

type	avg. count	mae	rmse	precision	recall	f1
color	16.99	5.79	12.14	0.72	0.77	0.74
location	17.76	11.97	17.76	0.42	0.28	0.34
material	32.38	3.75	4.74	0.77	0.81	0.79
orientation	7.83	3.35	5.78	0.59	0.81	0.68
size	10.5	1.0	1.0	0.62	0.62	0.62
variety	20.8	32.0	39.14	0.31	0.6	0.41

2.2. Qualitative results compared to base model

We show some qualitative results in Fig. 2 to illustrate the performance of our model compared to the base GroundingDino [5] model. Because of our contrastive learning module, our model is able to better distinguish different colors in (a), (b) and (f); better tell different actions in (d); and achieve fewer false positives for gender in (h). With our global-local feature fusion, our model is able to better localize the target object in (c), (e) and (g). These results show that our method makes improvements in terms of distinguishing different attributes of the same-class object and relational attribute types.

References

[1] Amazon Web Services, Inc. Amazon mechanical turk, 2023. [Online; accessed 1-April-2023]. 1

[2] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy. Cumulative attribute space for age and crowd density estimation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2467–2474, 2013. 1

[3] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H. Hsu. Drone-based object counting by spatially regularized regional proposal networks. In *The IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017. 1

[4] Intel. Cvat (computer vision annotation tool), 2023. [Online; accessed 1-April-2023]. 1

[5] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2

[6] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3394–3403, 2021. 1

[7] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *Technical Report*, 2020. 1

[8] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

[9] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding*, 2020. 1

[10] Zhaoyi Yan, Yuchen Yuan, Wangmeng Zuo, Xiao Tan, Yezhen Wang, Shilei Wen, and Errui Ding. Perspective-guided convolution networks for crowd counting. 2019. 1

[11] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2021. 1

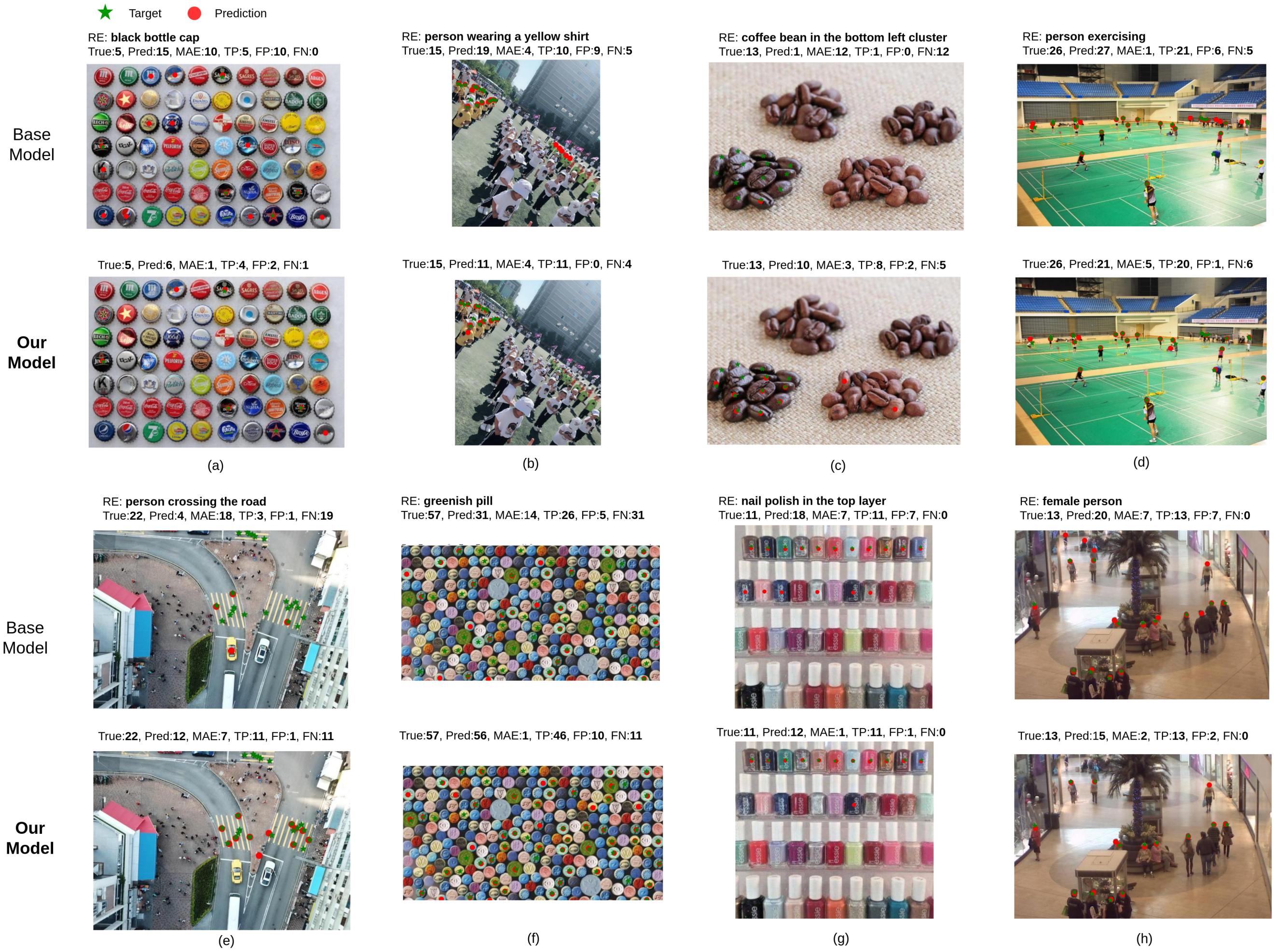

★ Target
● Prediction
RE: black bottle cap
True:5, Pred:15, MAE:10, TP:5, FP:10, FN:0
RE: person wearing a yellow shirt
True:15, Pred:19, MAE:4, TP:10, FP:9, FN:5
RE: coffee bean in the bottom left cluster
True:13, Pred:1, MAE:12, TP:1, FP:0, FN:12
RE: person exercising
True:26, Pred:27, MAE:1, TP:21, FP:6, FN:5
Base Model
True:5, Pred:6, MAE:1, TP:4, FP:2, FN:1
True:15, Pred:11, MAE:4, TP:11, FP:0, FN:4
True:13, Pred:10, MAE:3, TP:8, FP:2, FN:5
True:26, Pred:21, MAE:5, TP:20, FP:1, FN:6
Our Model
(a)
(b)
(c)
(d)
RE: person crossing the road
True:22, Pred:4, MAE:18, TP:3, FP:1, FN:19
RE: greenish pill
True:57, Pred:31, MAE:14, TP:26, FP:5, FN:31
RE: nail polish in the top layer
True:11, Pred:18, MAE:7, TP:11, FP:7, FN:0
RE: female person
True:13, Pred:20, MAE:7, TP:13, FP:7, FN:0
Base Model
True:22, Pred:12, MAE:7, TP:11, FP:1, FN:11
True:57, Pred:56, MAE:1, TP:46, FP:10, FN:11
True:11, Pred:12, MAE:1, TP:11, FP:1, FN:0
True:13, Pred:15, MAE:2, TP:13, FP:2, FN:0
Our Model
(e)
(f)
(g)
(h)

Figure 2. Results of our model compared to the base model (better viewed in enlarged version).