Supplementary Material:
Robust Region Feature Synthesizer for Zero-Shot Object Detection

Peiliang Huang[1], Junwei Han[1]*, De Cheng[2], Dingwen Zhang[1]*
[1]Brain and Artificial Intelligence Lab, School of Automation, Northwestern Polytechnical University
[2]State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University
{peilianghuang2017, junweihan2010, zhangdingwen2006yyy}@gmail.com, dcheng@xidian.edu.cn

1. Seen/unseen split of DIOR dataset

DIOR dataset. The DIOR [2] dataset is one of the largest, most discriminative, diverse, and publicly available object detection datasets for remote sensing imagery. It contains 23463 images with the size of 800 × 800 pixels and the spatial resolutions range from 0.5 to 30 m. All images are collected from Google Earth. The DIOR dataset is divided into train, test, and valuation sets where the train set and valuation set (i.e., 11725 images) are applied for training and the test set (i.e., 11738 images) is applied for testing. The full annotated dataset contains 20 classes of objects covering 192472 instances in total. These 20 classes include airplane (AI), airport (AR), baseball field (BF), basketball court (BC), bridge (BR), chimney (CH), dam (DA), expressway service area (ESA), expressway toll station (ETS), harbor (HA), golf field (GF), ground track field (GTF), overpass (OV), ship (SH), stadium (SA), storage tank (ST), tennis court (TC), train station (TS), vehicle (VE), and windmill (WI).

Seen/unseen split. Due to the lack of an existing zero-shot object detection protocol in remote sensing imagery, we first attempt to propose a challenging seen and unseen split for DIOR. We cluster the semantic word-vectors of all classes into different clusters using cosine similarity between the word-vectors as the metric. Fig 1 visualizes the clustering results. Classes with closer distances have stronger semantic relevance. We randomly select 4 classes (airport, basketball court, ground track field, and windmill) from different clusters as unseen classes and the rest of the 16 classes as seen classes for DIOR dataset. Inspired by the ZSD work [3] in natural images, this split is designed to follow the practical consideration: (a) unseen classes should be diverse, (b) the unseen classes should be semantically similar with at least some of the seen classes. In other words, the unseen classes are best from different clusters and each cluster should contain at least one seen class.

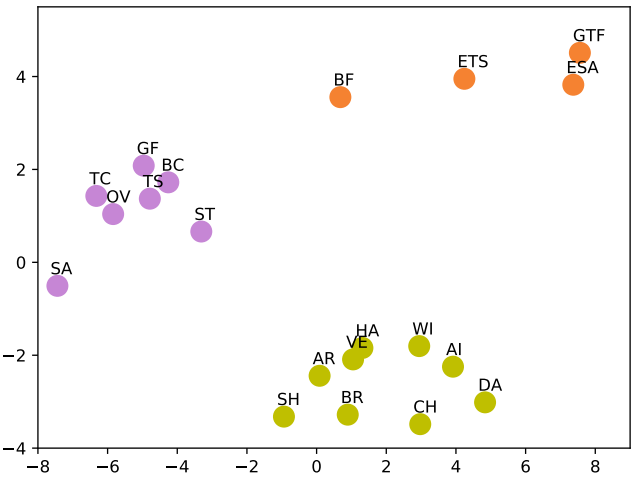

Figure 1. The relationship visualization of semantic word-vector on DIOR.

2. Class-wise mAP on DIOR

We report the class-wise AP and mAP performance of unseen classes in Table 1 for the ZSD setting and compare them with the second-best method [1]. We can observe that our method achieves 15.3%, 14.3%, 0.4%, and 7.6% improvement for "airport",

Table 1. Class-wise AP and mAP on unseen classes of DIOR dataset for ZSD setting.

Method	AR	BC	GTF	WI	mAP
SU [1]	13.1	5.6	22.9	0.5	10.5
Ours	**15.1**	**6.4**	**23.0**	**0.6**	**11.3**

*Corresponding author.

Table 2. Class-wise AP and mAP of 48/47 split on unseen classes of MS COCO dataset for ZSD setting.

48/17	airp	bus	cat	dog	cow	elep	umb	tie	snrd	skrd	cup	kife	cake	couch	kerd	sink	scrs	mAP
Ours	13.7	62.7	0.5	9.0	58.5	4.4	0.0	30.4	18.0	0.8	1.6	0.5	1.6	24.2	0.5	0.6	0.4	13.4

"basketball court", "ground track field" and "mAP", respectively. These comparing results demonstrate the effectiveness of our method.

3. Class-wise mAP on MS COCO for 48/17 split

Since other existing methods did not report their class-wise AP results on the 48/17 split, we only show our class-wise AP and mAP performance in Table 2. We can observe that our method achieves good detection performance on most classes.

References

[1] Nasir Hayat, Munawar Hayat, Shafin Rahman, Salman Khan, Syed Waqas Zamir, and Fahad Shahbaz Khan. Synthesizing the unseen for zero-shot object detection. In *ACCV*, 2020. 1

[2] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J PHOTOGRAMM*, 159:296–307, 2020. 1

[3] Shafin Rahman, Salman Khan, and Fatih Porikli. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *ACCV*, pages 547–563. Springer, 2018. 1