From Node to Graph: Joint Reasoning on Visual-Semantic Relational Graph for Zero-Shot Detection

Hui Nie[1,2], Ruiping Wang[1,2,3], Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]Beijing Academy of Artificial Intelligence, Beijing, 100084, China
hui.nie@vipl.ict.ac.cn, {wangruiping, xlchen}@ict.ac.cn

1. Details on Preliminary Studies (Sec. 3)

Dataset Statistics. For zero-shot recognition, the training data *zsr_train* and test data *zsr_test* contain 413282 and 62421 images respectively. For traditional image classification, the number of images in the test data *ic_test* is the same as *zsr_test*, while the training data *ic_train* consists of *zsr_train* and images of 15 unseen classes (up to 1000 images per category), resulting in 426161 images. The 15 unseen classes are *airplane, train, parking meter, cat, bear, suitcase, frisbee, snowboard, fork, sandwich, hot dog, toilet, mouse, toaster, hair drier*. For each of 15 unseen classes, the number of images is shown in Figure 1, some classes have less than 1000 images. Example images are shown in Figure 2.

2. Qualitative Analysis (Sec. 5.5)

More extra qualitative results of our method on the GZSD setting are shown in Figure 3. We show that our method obtains the correct detection for both seen and unseen classes, along with graph reasoning utilizing information of multiple objects. We also show some failure cases resulted from confusion with similar objects.

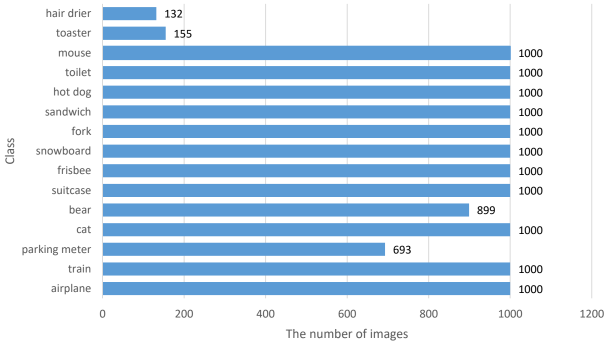

Figure 1. The number of images per class for 15 unseen classes.

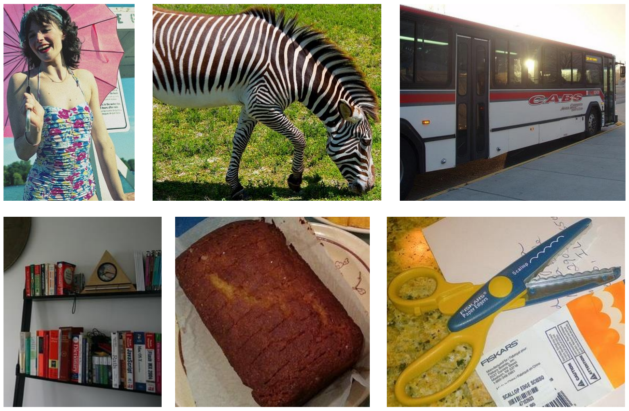

Figure 2. Example images from the *ic_train*.

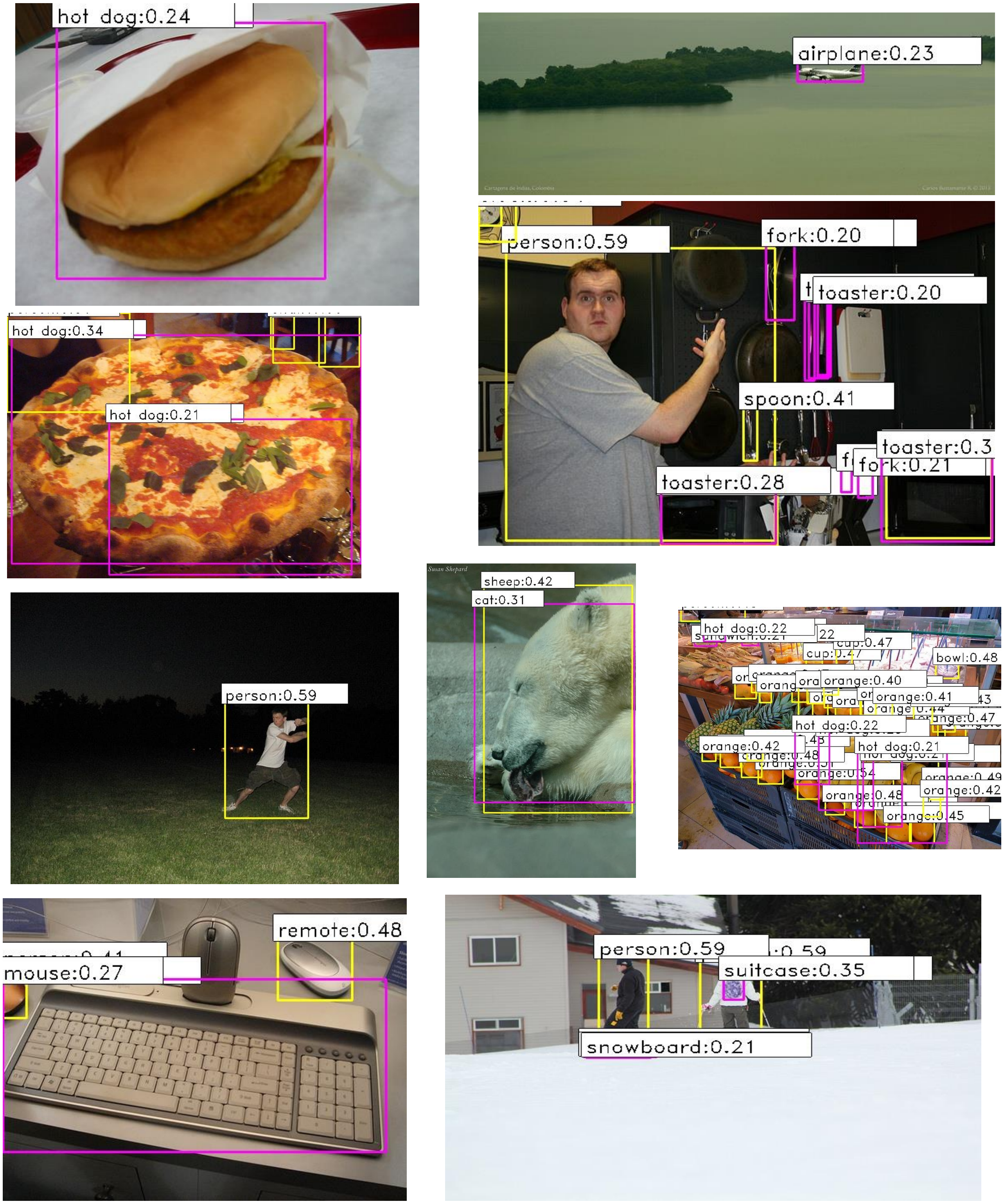

Figure 3. More qualitative results for the predictions of our method on GZSD. Yellow and pink boxes refer to predictions of seen and unseen classes respectively.