Supplementary Material About the Generation Process of Synthetic Scene Data

1 Data Collection and Synthetic Scene Generation

In this section, we give detailed descriptions of the dataset for dense descriptor learning. As mentioned in the paper, we synthetically create the scenes using raw data from publicly available datasets and the Internet without any human or robot labor. In total, we collected 130 distinct instances of 16 object classes as shown in Fig. 1. It is noteworthy that the corresponding object binary masks are also collected from the publicly available datasets.

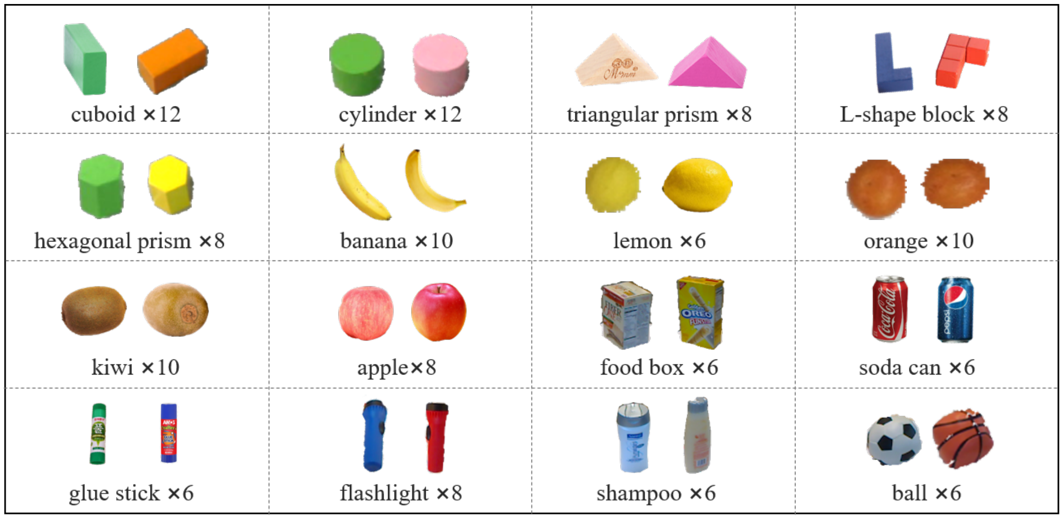

Fig. 1. Overview of the collected objects used for dense descriptor learning.

For synthetic data generation, we employ a simple yet reliable scheme as prior work, that is, overlaying object instance masks on randomly selected background images and establishing pixel correspondences accordingly. Details are described as below.

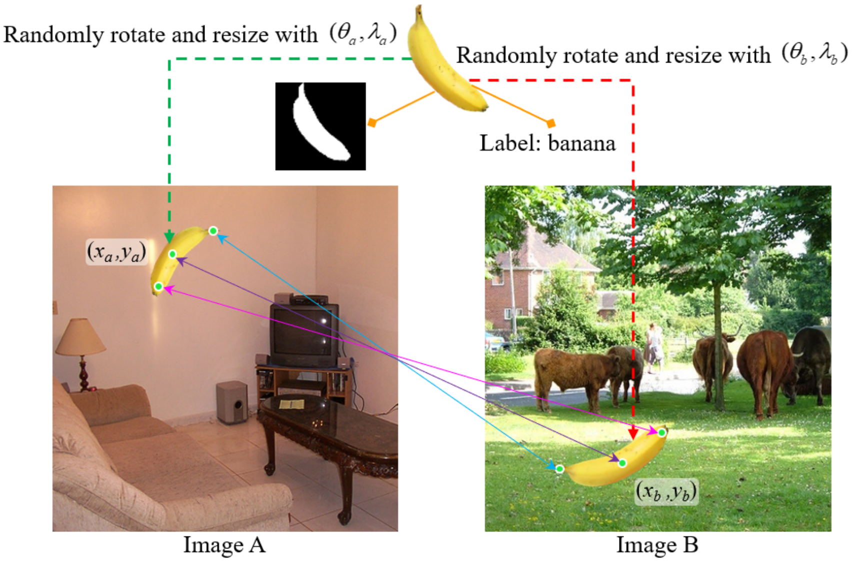
Randomly rotate and resize with (θ_a, λ_a)
Randomly rotate and resize with (θ_b, λ_b)
Label: banana
(x_a, y_a)
(x_b, y_b)
Image A
Image B

Fig. 2. Illustration of the synthetic scene generation.

As shown in Fig. 2, we first randomly select N objects of N classes from the collected raw data. Thus for each selected object, its category label and binary mask are known at the beginning. Then

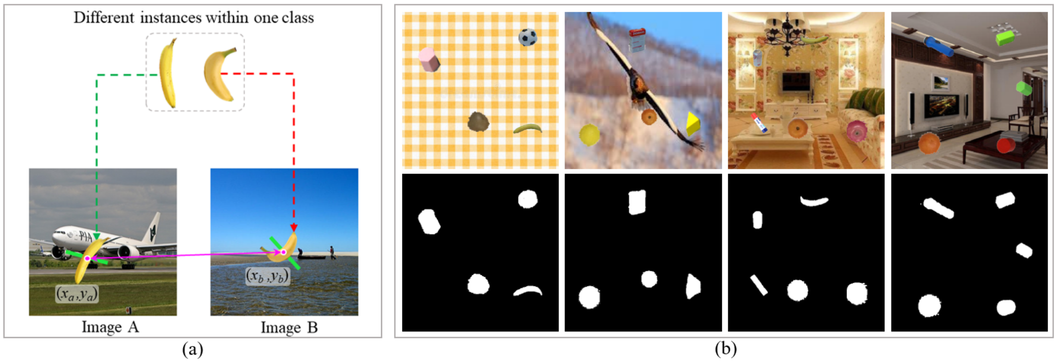

Fig. 3. (a) illustrates the generation of cross-instance synthetic scenes; (b) shows some samples of the created synthetic scene images and the corresponding binary masks.

the object is randomly rotated by θ_a and θ_b and resized by two scale factors λ_a and λ_b respectively. Afterwards, the two rotated and resized instances are overlaid on two randomly selected scene images I_a and I_b at the locations (x_a, y_a) and (x_b, y_b) respectively. Since both of the two instances are the regular variants of the same object, we can establish precise pixel correspondences between them without extra efforts. Besides, by overlaying the corresponding rotated and resized binary masks on two black images, the ground truth segmentation maps for the training of the segmentation branch are also available.

To further improve the matching performance of our multi-object dense descriptor, we also introduce cross-instance training data which are rarely considered in prior work. As illustrated in Fig. 3(a), two different instances within one class are overlaid on two background images with random rotation and scaling. Similarly, object category label and binary segmentation map could be easily obtained. However, it is hard to derive matched pixel pairs from cross-instance data in the aforementioned manner. Instead, we sample grasping-based matched pixel pairs from cross-instance data using the grasp affordance prediction model.