Hierarchical Scene Parsing by Weakly Supervised Learning with Image Descriptions

Ruimao Zhang, Liang Lin, Guangrun Wang, Meng Wang, and Wangmeng Zuo

SUPPLEMENTARY MATERIAL

A. Dataset

Sentence Annotation. We asked 5 annotators to provide one descriptive sentence for each image in the PASCAL VOC 2012 [1] segmentation training and validation set. Images from two sets are randomly partitioned into five subsets of equal size, each assigned to one annotator. We provided annotators with a list of possible entity categories, which is the 20 defined categories in PASCAL VOC 2012 segmentation dataset.

We ask annotator to describe the main entities and their relations in the images. We did not require them to describe all entities in images, as it would result in sentences being excessively long, complex and unnatural. Fig. 1 illustrates some pairs of images and annotated sentences in VOC 2012 *train* and *val* set. For most images, both the objects and their interaction relations can be described with one sentence. In particular, we summarize three significant annotation principles as follows:

- For the image with only an instance of some object category, e.g., the last image in the first row of Fig. 1, the sentence describes the relation between the object (i.e. airplane) and the background (i.e. runway);
- For the instances from the same category with the same state, we describe them as a whole. Such as the forth image in the seconde row of Fig. 1, the annotation sentences is "two motorbikes are parked beside the car".
- For the instances from the same category with the different state, the annotator may only describe the most significant one. As to the third image in the second row of Fig. 1, the annotator describe the people sitting on the chairs but ignore the baby sitting on the adult.

We did not prohibit describing entities that did not belong to the defined categories, because they are necessary for natural expression. But we will remove them in the process of generating semantic trees.

We annotate one sentence for each image because our method involves a language parser which produces one semantic tree for each sentence. At this point, we are unable to generate one tree structure from multiple sentences. Therefore, one sentence for each images is sufficient for our study. To give more details of the image descriptions, we provide our sentence annotations of entire dataset in "*train_sentences.txt*" and "*val_sentences.txt*" as supplementary materials.

As described in the main paper, we parse sentences and convert them into semantic trees which consist of entities, scene structure and relations between entities. Here we provide the list of 9 relation categories we defined: *beside, lie, hold, ride, behind, sit on, in front of, on* and *other*. The label *other* is assigned in the following two cases. (i)

Fig. 1. Some pairs of images and annotated descriptions in PASCAL VOC 2012 dataset. Images in the first row are sampled from training set, while the second row's images are collected from the validation set.

An entity has the relation with the background, which often happens at the last layer of the parsing structure. (ii) The *other* relation is used as placeholder for the relation not identified as any of the 8 other relations

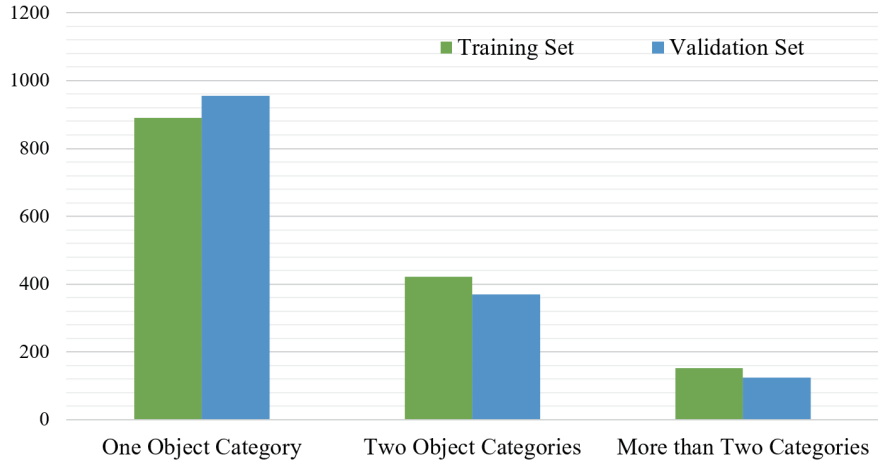

Fig. 2. The number of object category of each image in VOC *train* and *val* dataset. The abscissa indicates the number of object categories in the image. The ordinate indicates the number of images. In each image, the number of interaction relations usually increases with the number of objects growing.

Annotation Statistics. Since the sentence annotations are not a standard part of the PASCAL VOC dataset, we give some statistical analysis of images and annotations in Fig. 2 and Fig. 3 to incorporate more information about our parsing task. Fig. 2 shows the number of object category of each image in VOC *train* and *val* dataset. Obviously, for PASCAL VOC 2012 dataset, most images only contain one object category. In order to construct the tree structure, we combine the foreground object and the background, and assign "*other*" as their relationship. Another kind of images contain two or more object categories, and the number of relations in these images is

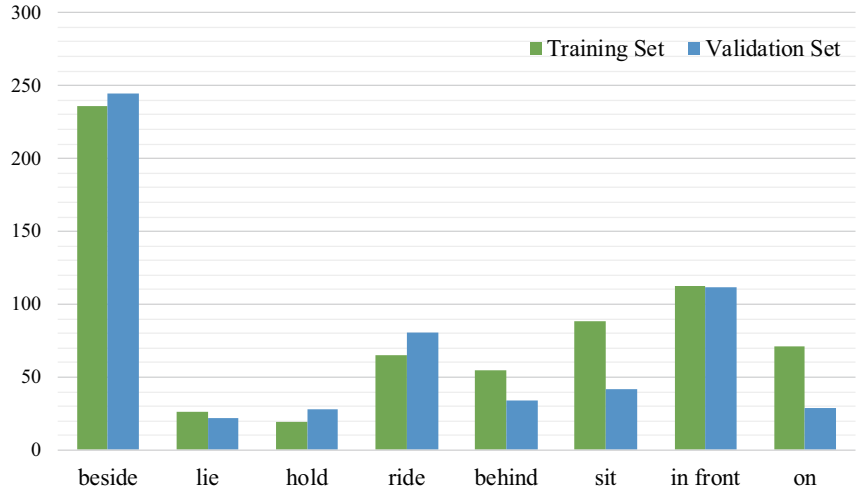

Fig. 3. The number of occurrences of each relation category in the *train* and *val* dataset. Note that each image may contain multiply relations.

greater than one. As stated above, we combine the merged foreground objects and the background with the relation "*other*" at the last layer of the semantic tree. According to the Fig. 2, the proportion of images with two or more object categories in the entire dataset is greater than $1/3$ (*i.e.* 39.21% for training set and 34.09% for validation set). Since the the number of interaction relations usually increases with the number of object growing, the total number of relations (except "*other*") in these images is more than 50% of the entire dataset based on our sentence annotations.

Fig. 3 reports the number of occurrences of each relation category in VOC *train* and *val* dataset. The most common relation label is "*beside*", and the number of its occurrences is 236 in training set and 245 in validation set. The label "*lie*" and "*hold*" are two least common labels, and occurrences times are around 20 in both training and validation set.

B. Experiment Results

Analysis on Relation Loss. We note that the RsNN model in previous works (e.g., Socher et al. [2]) only consider the structure supervision, but our model takes both structure and relation supervision during model training. To evaluate the performance of our method with and without relation supervision, we add some visualized results in Fig. 4. According to the figure, one can see that both of two methods learn the correct combination orders. However, our method can further predict the interaction relation between two merged object regions. More importantly, the relation loss can also regularize the training process of CNN, which makes the segmentation model more effective to discover the small objects and eliminate the ambiguity.

Analysis on Category Level Description. Instead of instance-level parsing, this work aims to learn a CNN-RsNN model for category-level scene parsing. When asking the annotator to describe the image, some guidelines are introduced in Sec.-A to avoid instance-level descriptive sentences. Under such circumstances, it is interesting to ask whether such annotation strategy are harmful to semantic labeling on images with multiple instances. To

Subset	(i)	(ii)	(ii)
Num. of Image	766	266	417
mean IoU	35.94%	33.19%	34.70%

TABLE I

RESULTS ON DIFFERENT SUBSET OF VOC 2012 *val* UNDER THE WEAKLY SUPERVISED LEARNING.

answer this, we divide the VOC 2012 *val* set into three subsets: (i) images with one instance from one object category, (ii) images with instances from multiple object categories, but only one instances from each category, and (iii) the others. The mean IoU of our model on these three subsets are reported in Table I. Although the number of object categories per image, the number of instances per category, and the number of images have the obvious difference among three subsets, the changes of mIoU remain in a small range. It demonstrates that our category-level descriptions have little negative effect on semantic labeling results of images with multiple instances.

Analysis on Parsing Results. To further investigate the performance of structure prediction, we provide some typical successful and failure cases of scene structure prediction in Fig. 5 and Fig. 6. All of them are generated under the weakly supervised setting as described in the main paper.

We first show some *successful* parsing results in Fig. 5. It is interesting to note that, our scene structure generation model is robust to small degree of semantic labeling error. As in the left image of the last row, even only a small part of the person is correctly labeled, both structure and relation prediction can be successfully predicted. The relation categories in these examples cover most of the defined relations in this article. Then, the *failure* cases are illustrated in Fig. 6. According to this figure, the failure predictions usually happen in the following three cases. (i) All of the structure and relation predictions are incorrect. Fig. 6-(a) and Fig. 6-(c) illustrate such situation. (ii) The structure is correct but the predicted relations are wrong. Fig. 6-(b) gives the example like this. (iii) Both the structure and relation predictions are partially correct. Fig. 6-(d) gives the example in such case.

According to the above discussion, one can see that the main cause of failure is the semantic labeling error, including seriously inaccurate labeling and complete failure in segmenting some object category. Moreover, when the semantic labeling is inaccurate, the relation tends to be wrongly predicted as others (see Fig. 6-(a)(b)(c)). When some object category is completely failed to be recognized, structure prediction is likely to be incorrect or partially incorrect (see Fig. 6-(a)(d)).

REFERENCES

[1] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

[2] R. Socher, C. C. Lin, A. Y. Ng, and C. D. Manning, "Parsing natural scenes and natural language with recursive neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 129–136.

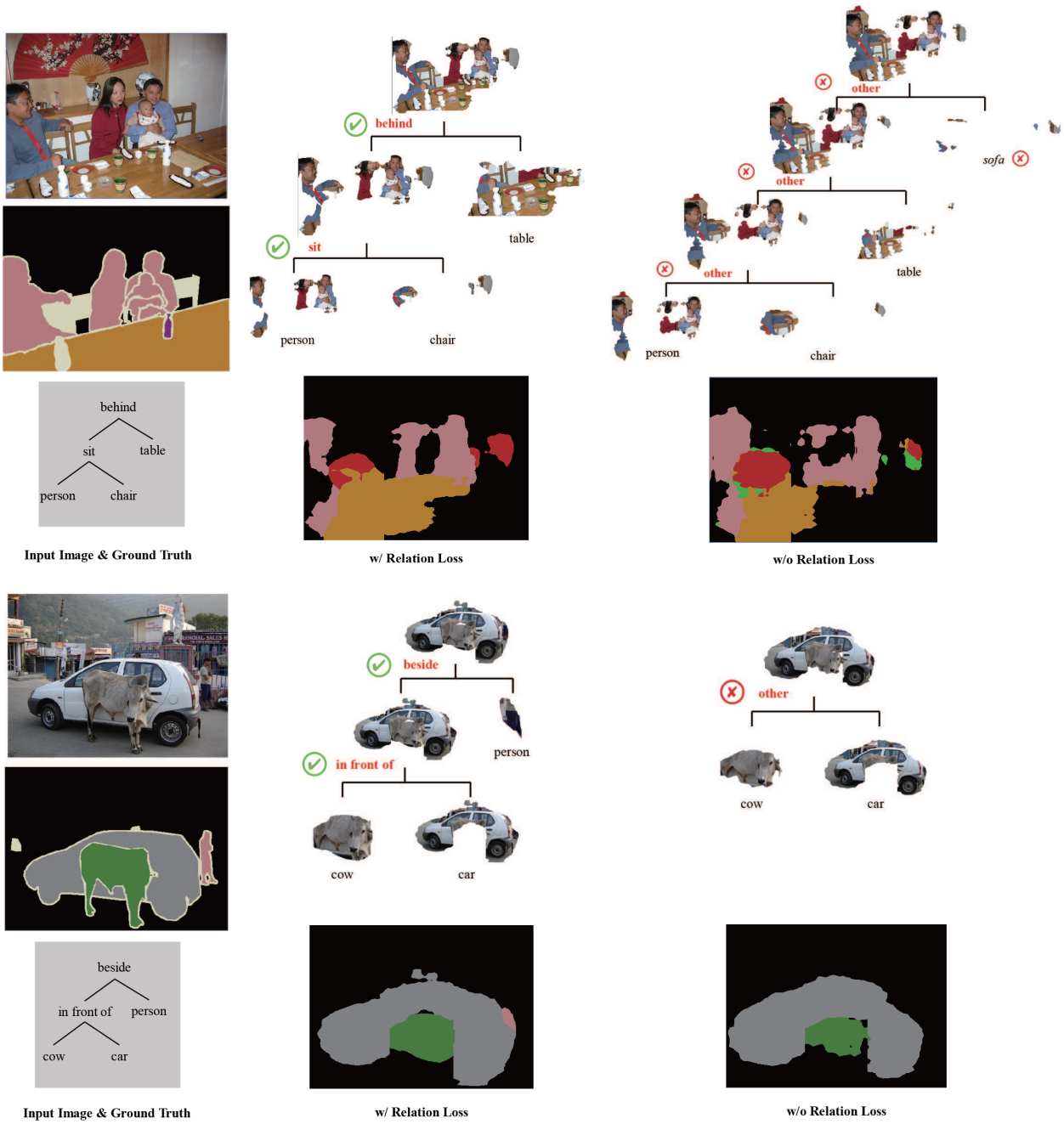

Fig. 4. Some visualized semantic segmentation and scene structure prediction results with and without relation loss on PASCAL VOC 2012 val dataset. The first column shows the input images, the ground truth of semantic labeling and semantic trees. The second column gives the segmentation and structure prediction results with the relation loss (our method). In contract, the results without the relation loss are illustrated in the last column (like Socher et al. method).

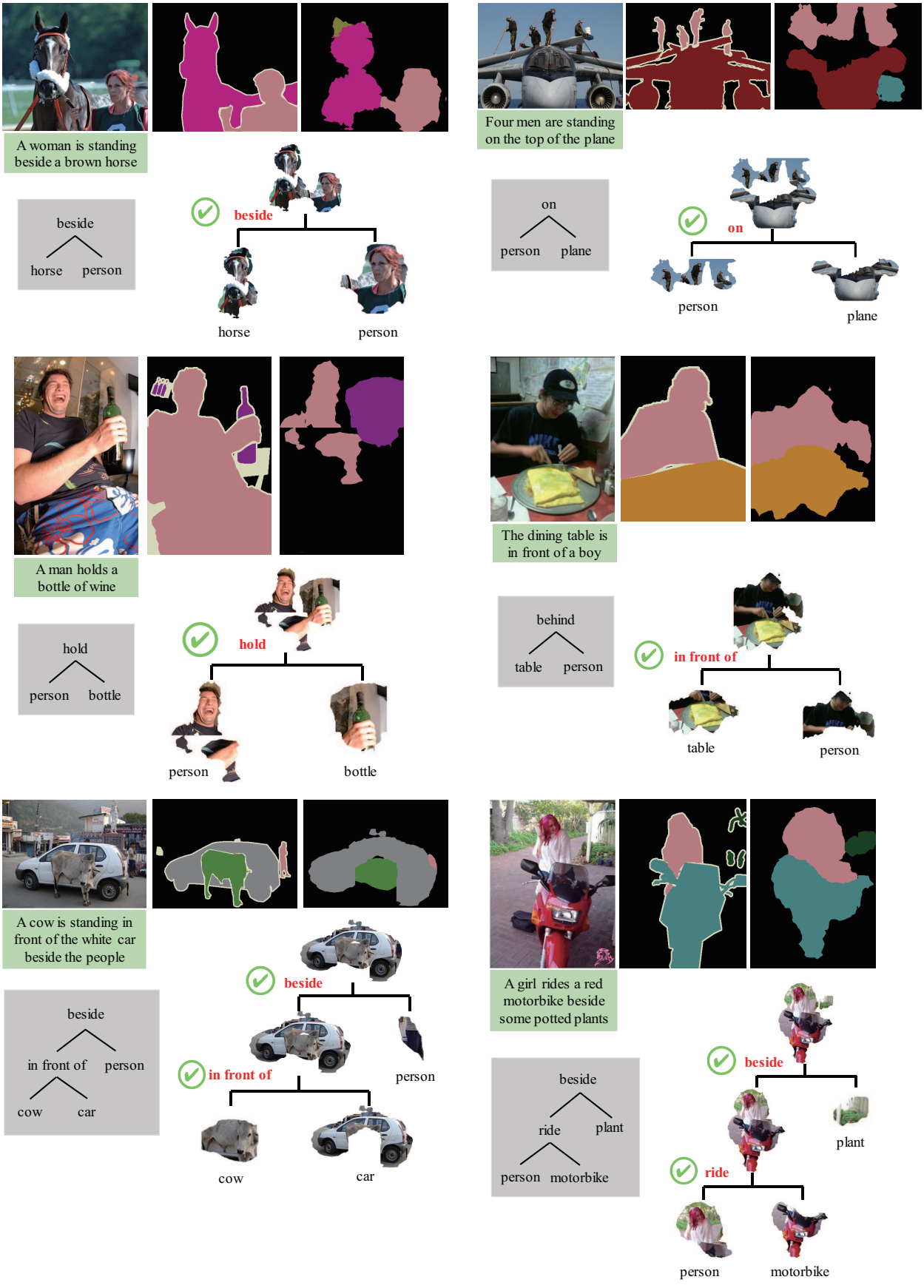

Fig. 5. The visualized successful scene parsing results in PASCAL VOC 2012 dataset under the weakly supervised setting.

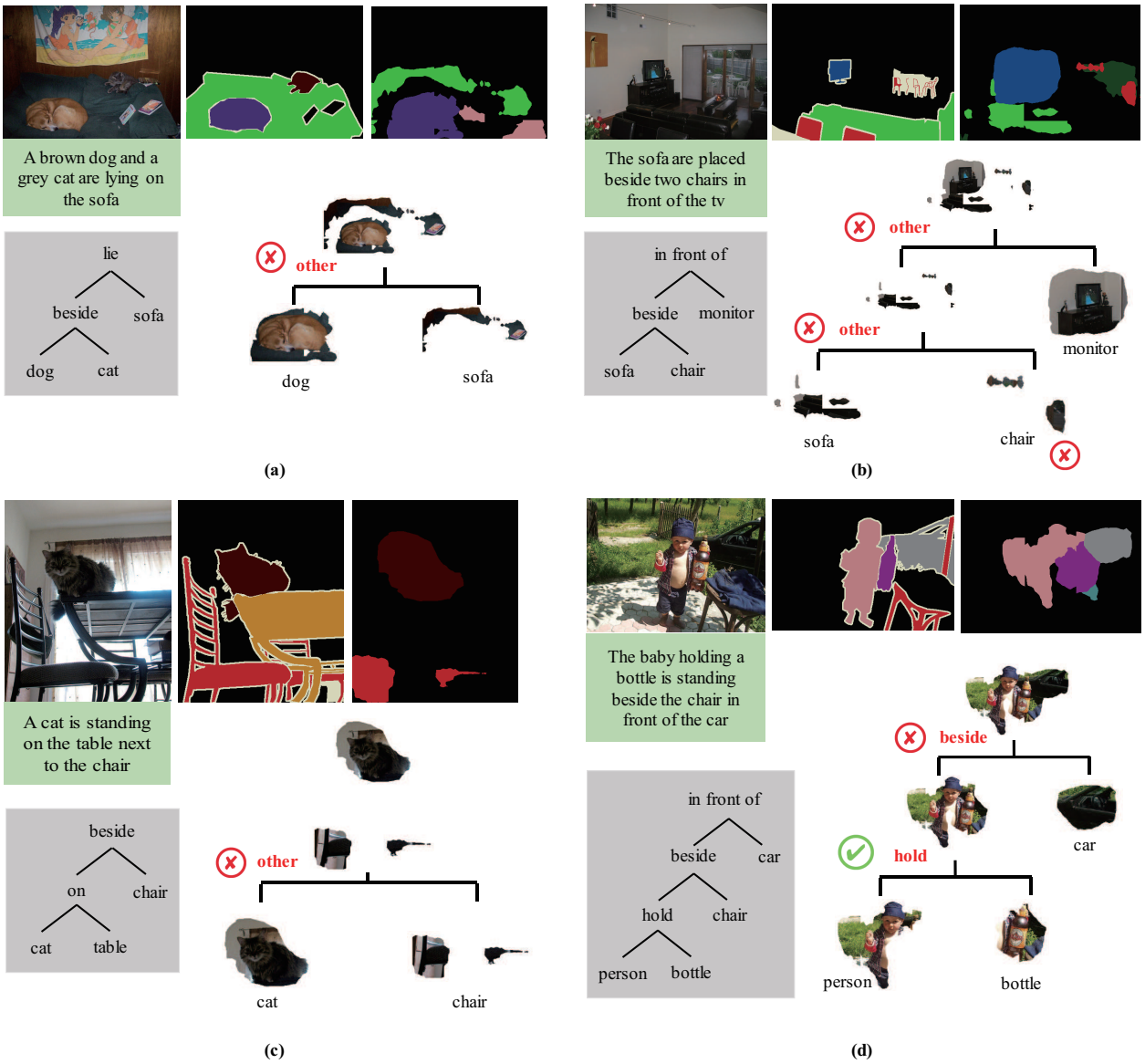

Fig. 6. The visualized failure scene parsing results in PASCAL VOC 2012 dataset under the weakly supervised setting.