Supplementary Material for Scaling up instance annotation via label propagation

Dim P. Papadopoulos*
MIT CSAIL
dimpapa@mit.edu

Ethan Weber*
MIT CSAIL
ejweber@mit.edu

Antonio Torralba
MIT CSAIL
torralba@mit.edu

This supplementary material provides additional information about the experiments and the dataset described in the main paper. We further discuss hyperparameter choices, the dataset composition with examples, qualitative clustering examples, and crowd-sourcing implementation details.

1. Simulated experiments

In this section, we provide extra details on the simulated experiments presented in Sec. 5.2 of the main paper about the procedure of tuning the hyperparameters of the annotation and the propagation steps.

Annotation and propagation. In Sec. 5.2 of the main paper we describe how we tune three hyperparameters of our method: the number of verified samples per cluster N_s, the cluster quality threshold K_a and the mask IoU threshold between the obtained and the ground-truth masks K_{iou}. Following the human annotation consistency in manually annotated instance segmentation datasets, a high quality dataset has a segmentation quality SQ between 0.8 andd 0.85. We want to find the optimal values for N_s, K_a and K_{iou}, while keeping $SQ \geq 0.85$. Fig. 1(a) shows the resulting SQ using different K_a and K_{iou} values and assumes that $N_s = \infty$ (i.e., the estimated quality of a cluster is the real one). We only show the SQ values that are above 0.85. Fig. 1(b) shows the SQ for $N_s = 15$ and different K_a and K_{iou} values. Note that $N_s = 15$ is the minimum number of verified samples that lead to $SQ \geq 0.85$. From all the possible solutions for the pair K_a and K_{iou} in Fig. 1(b), we keep the one that leads in the largest number of obtained annotations (Fig. 1(c)). This results in $N_s = 15$, $K_a = 0.85$ and $K_{iou} = 0.75$ (highlighted in a black circle in Fig. 1(b), (c)).

2. Large-scale experiments on Places

In this section, we provide extra statistics about the obtained annotations from our large-scale experiment and we present extra annotation and clustering examples.

*Indicates equal contribution

Category distribution. In Fig. 2, we show the number of the obtained annotations per object category. We also illustrate the per-class differences among the number of annotations for our annotations and ADE20K. For most object categories, we obtain one to two orders of magnitude more object annotations than ADE20K with a budget of less than $1,000. Note that a higher annotation budget can lead to more annotations. We only show the 69 object categories for which we obtain annotations on Places. There are another 31 categories where we did not obtain any new annotations due to early rejection of all the clusters in T. This is because the majority of the predictions made by the segmentation model are wrong leading to exclusively low quality clusters close to the root of T. Obtaining annotations for these categories requires to slightly increase the annotation budget and prevent the rejection of all low quality clusters.

Crowd-sourcing experiments. We provide here more details about the crowd-sourcing protocol used to obtain our high quality annotations on Amazon Mechanical Turk (AMT)[1]. In Fig. 3(a), we show the interface for the binary verification task. The annotators are shown a cropped image with a mask outline and target class and are instructed to respond positively if the mask outlines the target object correctly (IoU ≥ 0.75), and negatively otherwise.

To ensure good quality, the annotators first read a simple set of instructions with several examples (Fig. 3(b)). Then, they go through a simple qualification test, at the end of which we provide detailed feedback on how well they performed. Annotators who successfully pass this test can proceed to the annotation stage. In case of failure, they can repeat the test until they succeed. In the annotation stage, annotators are presented small batches of 56 consecutive masks. For increased efficiency, the batches consist of a single object category. During this stage, we control the quality by hiding 6 quality control examples inside each batch for which we have ground-truth annotation masks. We monitor annotators accuracy on these examples and prevent them to

[1] https://mturk.com/

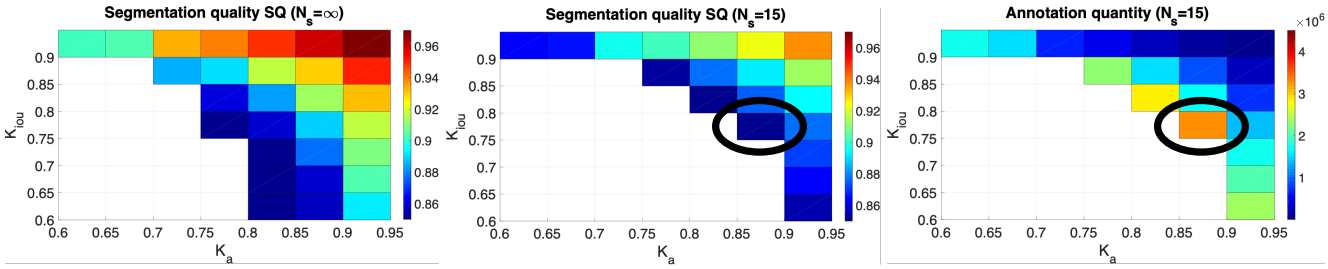

Figure 1. **Annotation and propagation hyperparameters.** (a) The segmentation quality SQ of the obtained annotations for $N_s = \infty$ using different K_a and K_{iou} values. (b) The segmentation quality SQ of the obtained annotations for $N_s = 15$ using different K_a and K_{iou} values. (c) The number of obtained annotations for $N_s = 15$ using different K_a and K_{iou} values.

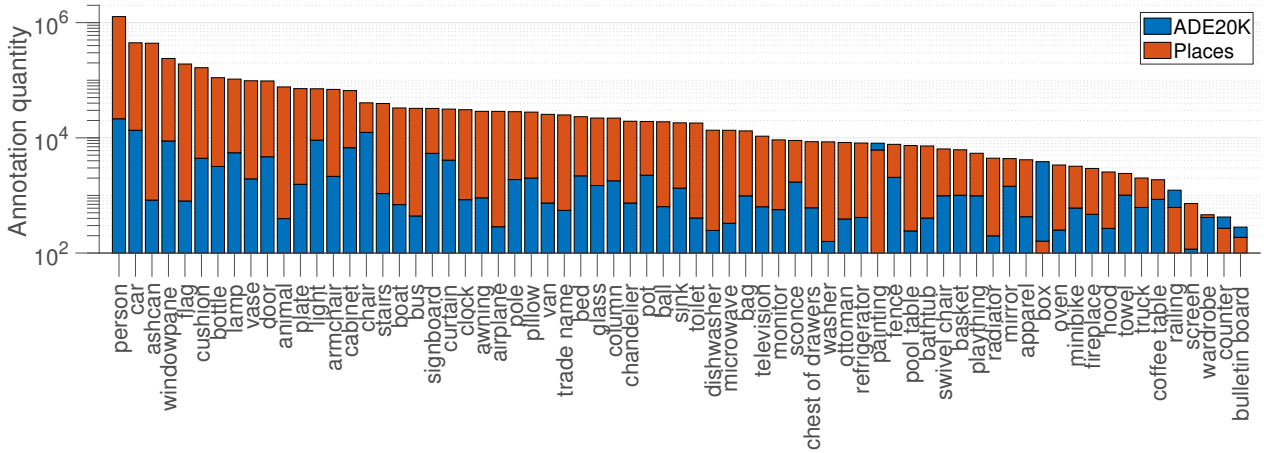

Figure 2. **Object category distribution.** We compare the number of annotations for the initial ADE training set (blue) and the obtained annotations in Places (red). Note that the graph is in log scale and the number of the obtained annotations is one to two orders of magnitude larger than the annotations in ADE20K for most categories.

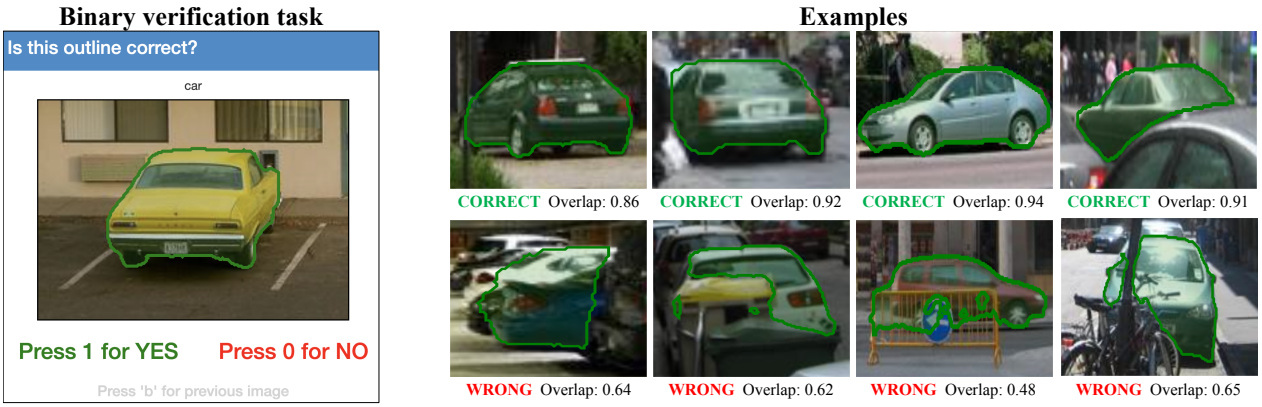

Figure 3. **Binary verification task.** (a) The interface used for fast binary verification. Annotators press "1" (CORRECT), "0" (WRONG), or "b" (GO BACK) until all questions in the batch are complete. (b) Correct and wrong instruction examples for the car category. We train MTurk annotators with many instruction examples and a qualification test.

submit if they fail to achieve a high accuracy.

Clustering qualitative results. In Fig. 4 and 5, we show examples of clusters with their corresponding quality estimates. Specifically, the green outlined masks are positively verified by human annotators on AMT and the red outlined masks are negatively verified. Note that some quality estimates may be computed with more than 15 samples due to

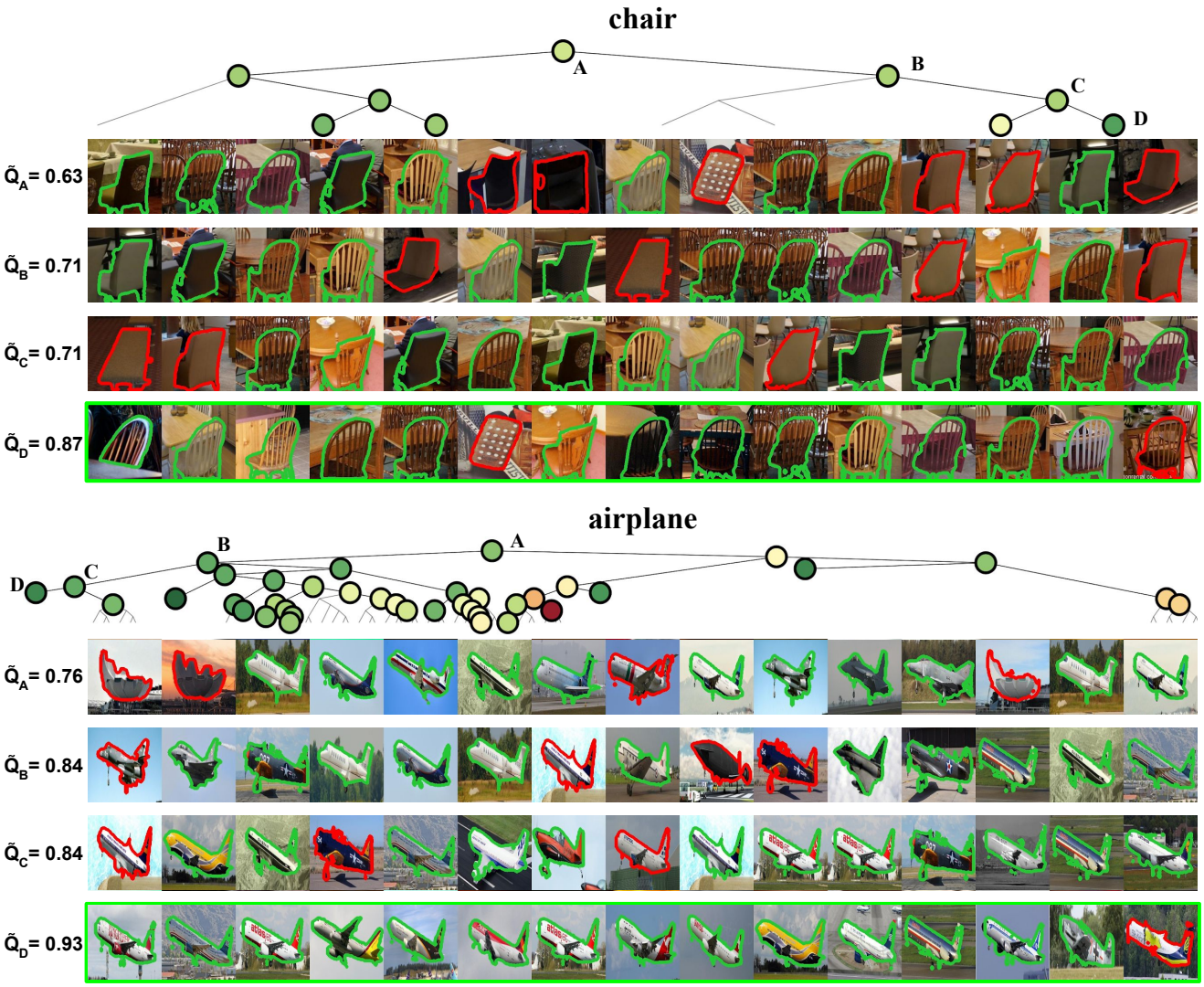

Figure 4. **Cluster examples (Part 1/2).** For each category (chair and airplane), we show the subtree for cluster A. Below, we show the quality estimate based on AMT responses. Green outlines mean positively verified, and red outlines mean negatively verified. Notice that clusters D are high quality, pruned, and added to our obtained dataset.

crowd-sourcing implementation details when near the root node. To speed up the search procedure on AMT and enable running iterative binary verification tasks in batches, we initiate the annotation procedure by asking 150 questions at the root cluster.

In Fig. 6(top), we show a class-specific subtree for the lamp category. The high quality clusters are colored in green, and the low quality ones in red. For the root cluster of this subtree, we show randomly selected masks from the left and right children (Fig. 6(bottom)). The selected cluster is near the actual root node of T containing more than 50K masks and that is why the appearance and the quality of the masks vary a lot

Mask annotation examples. In Fig. 7,8, we show class-specific, cropped annotation masks that are obtained for 36 different object categories in Places. In Fig. 9, 10, 11, 12, 13, we show a random selection of our obtained mask annotations in full images of the Places dataset using our proposed pipeline. We will release the obtained 4.3M mask annotations upon acceptance of the paper.

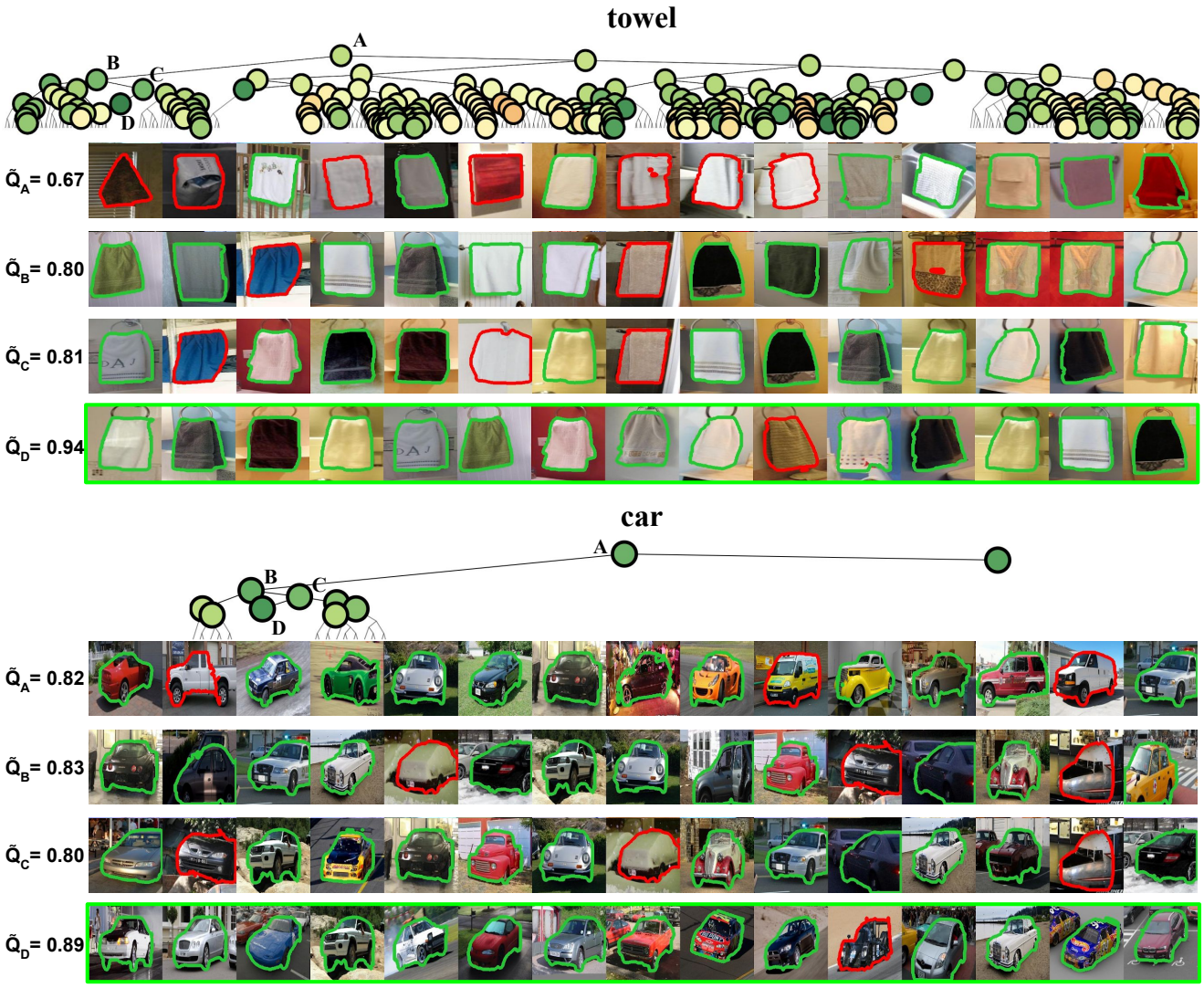

Figure 5. **Cluster examples (Part 2/2).** For each category (towel and car), we show the subtree for cluster A. Below, we show the quality estimate based on MTurk responses. Green outlines mean positively verified, and red outlines mean negatively verified. Notice that clusters D are high quality, pruned, and added to our obtained dataset.

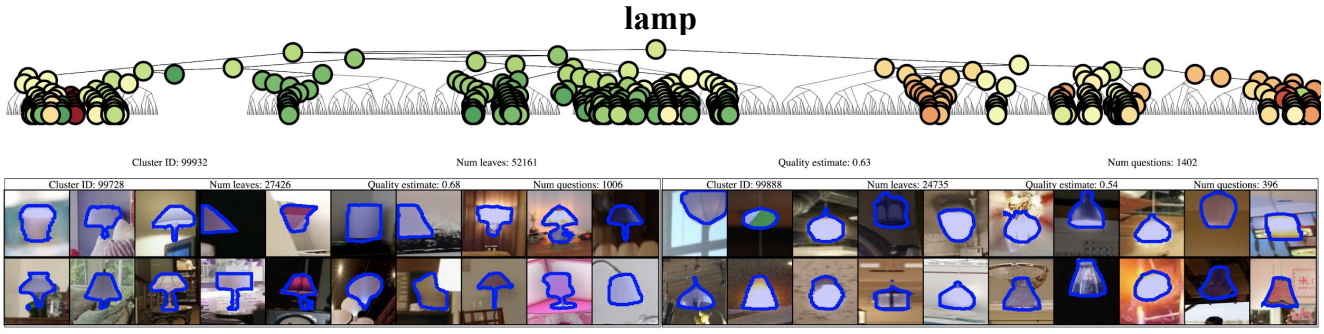

Figure 6. **Cluster tree viewer.** (Top) We show a subtree for the lamp category. The annotated clusters of the tree are color-coded according to the quality estimate (green for high quality and red for low). Notice that clusters without colored dots have yet to be explored by the search procedure. Accepted or rejected clusters, however, are pruned and set as a leaf. (Bottom) For the current root node of the subtree, we show randomly selected masks from the left and right children.

Figure 7. **Category annotations (Part 1/2).** We show four category-specific annotations for 18 different object categories that are obtained in the Places dataset.

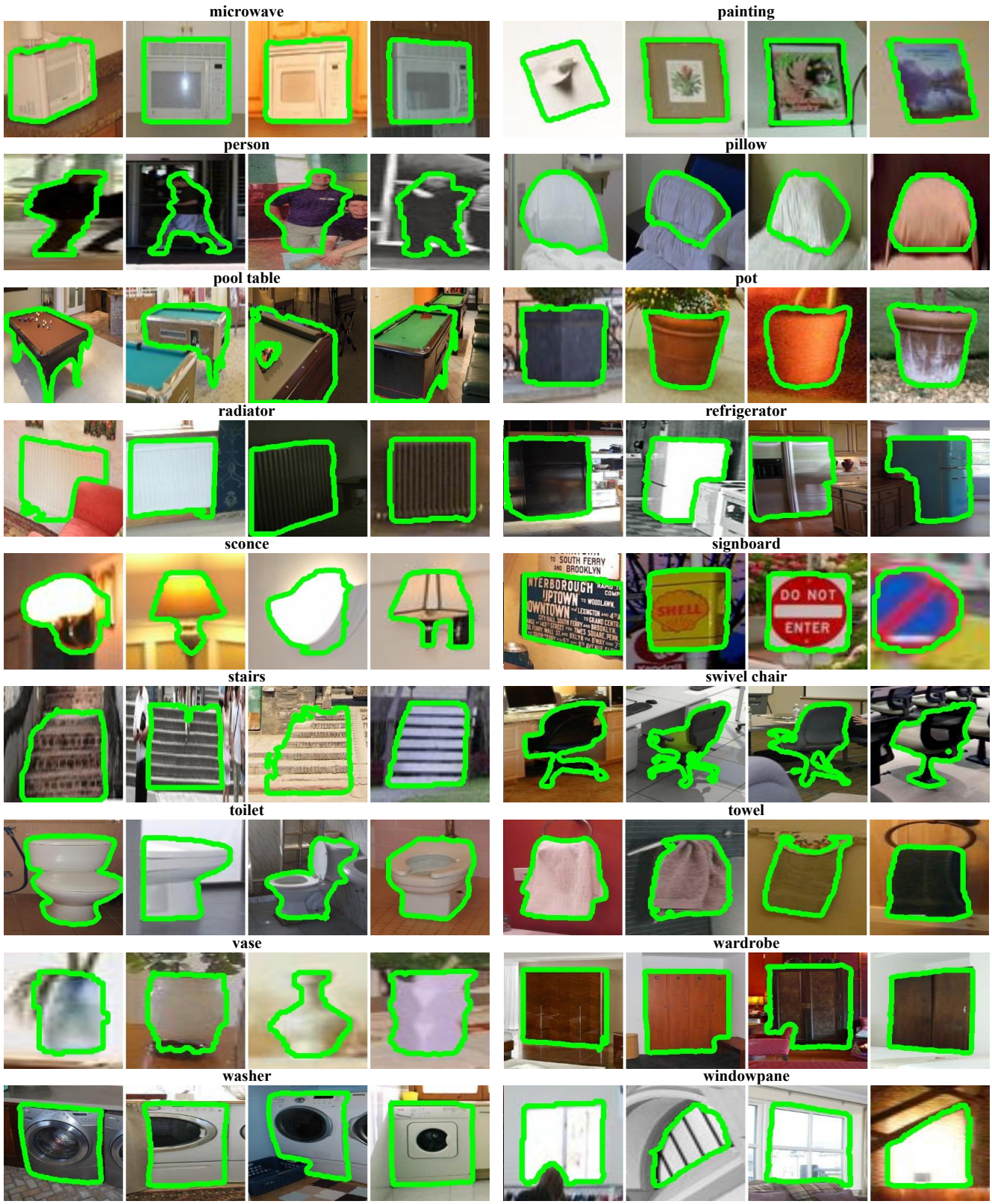

Figure 8. **Category annotations (Part 2/2).** We show four category-specific annotations for 18 different object categories that are obtained in the Places dataset.

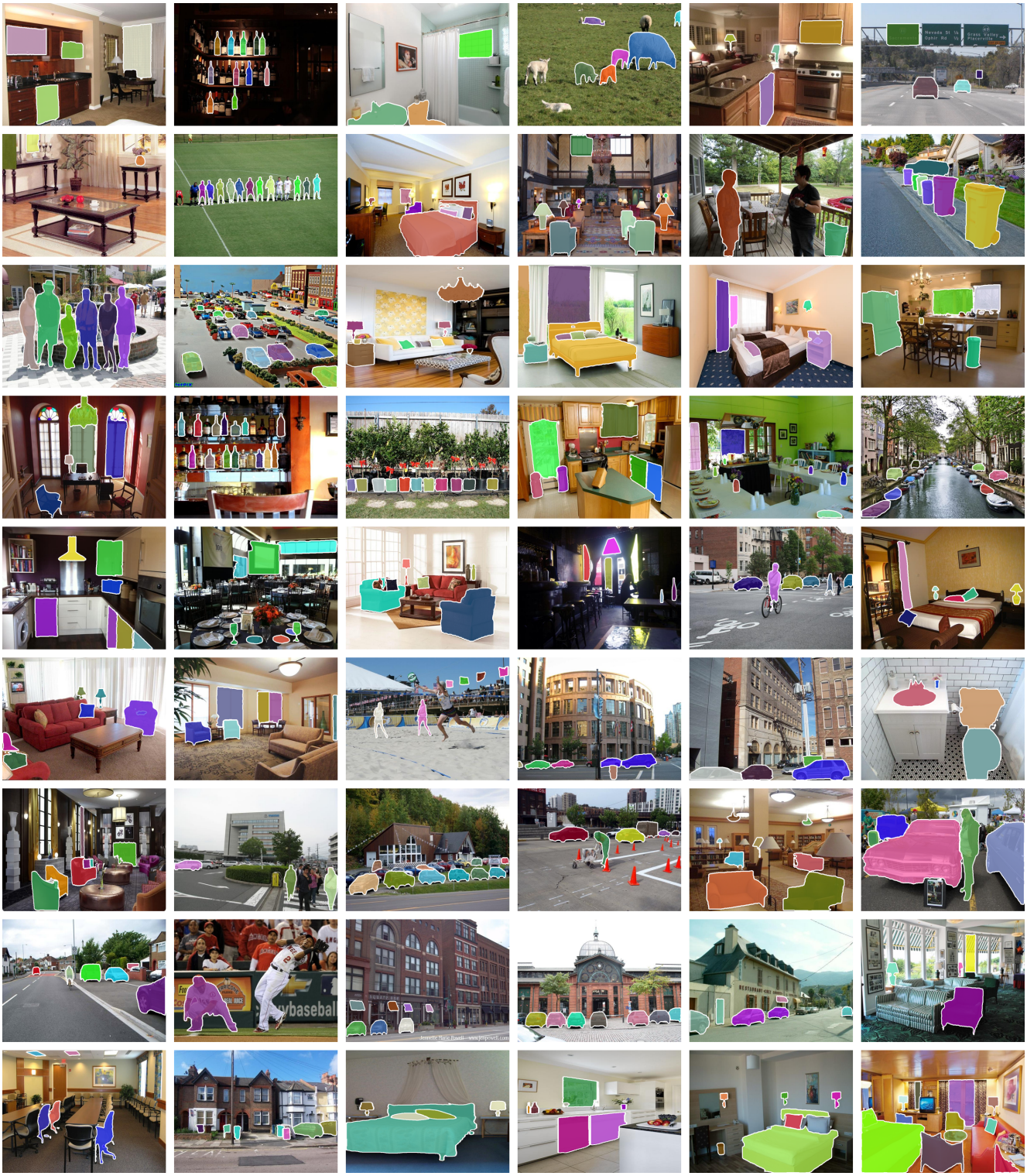

Figure 9. **Obtained mask annotations in Places (Part 1/5).** We show here a random selection of example images from the Places dataset with our obtained mask annotations using our pipeline under a small fixed annotation budget.

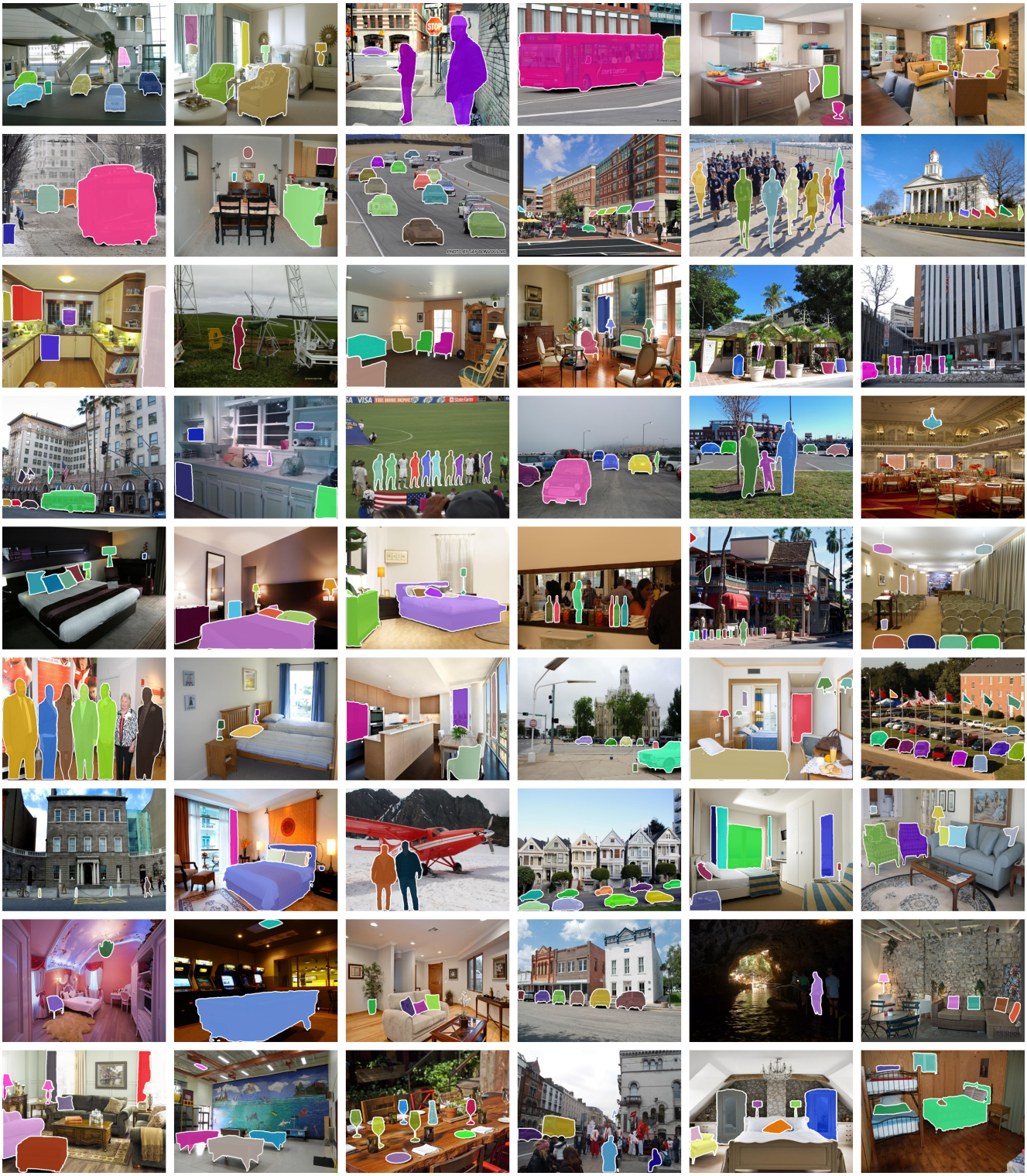

Figure 10. **Obtained mask annotations in Places (Part 2/5).** We show here a random selection of example images from the Places dataset with our obtained mask annotations using our pipeline under a small fixed annotation budget.

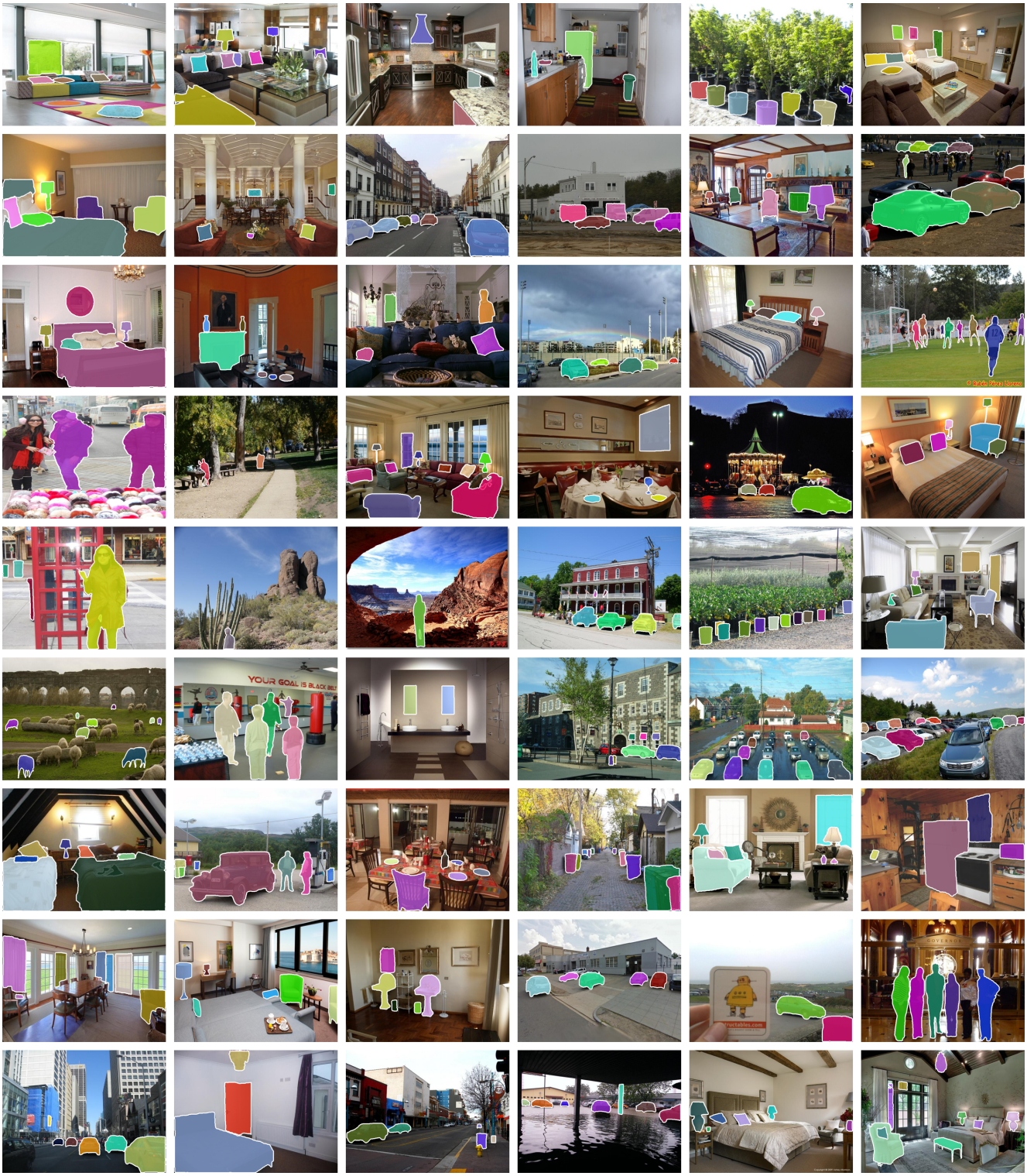

Figure 11. **Obtained mask annotations in Places (Part 3/5).** We show here a random selection of example images from the Places dataset with our obtained mask annotations using our pipeline under a small fixed annotation budget.

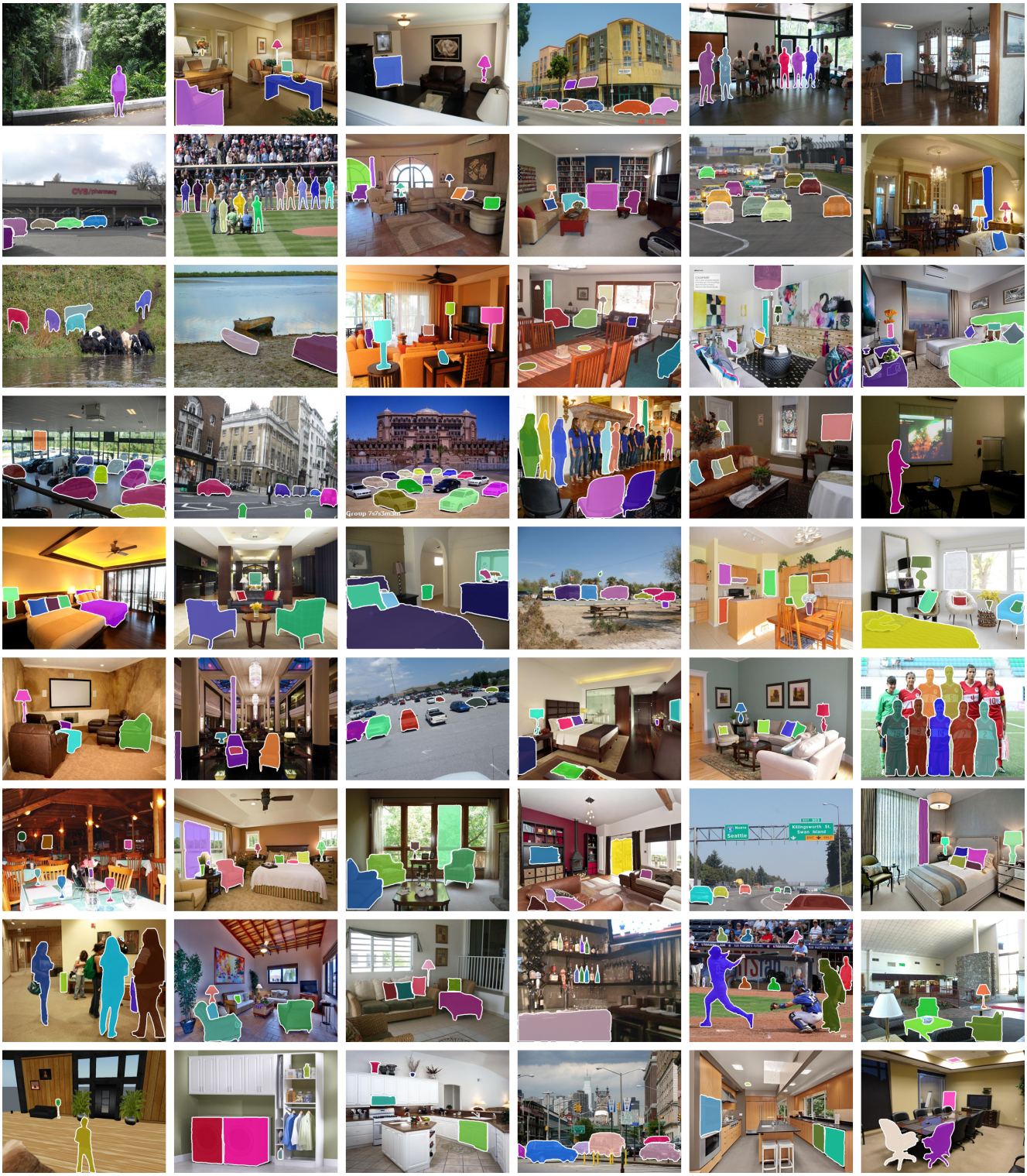

Figure 12. **Obtained mask annotations in Places (Part 4/5).** We show here a random selection of example images from the Places dataset with our obtained mask annotations using our pipeline under a small fixed annotation budget.

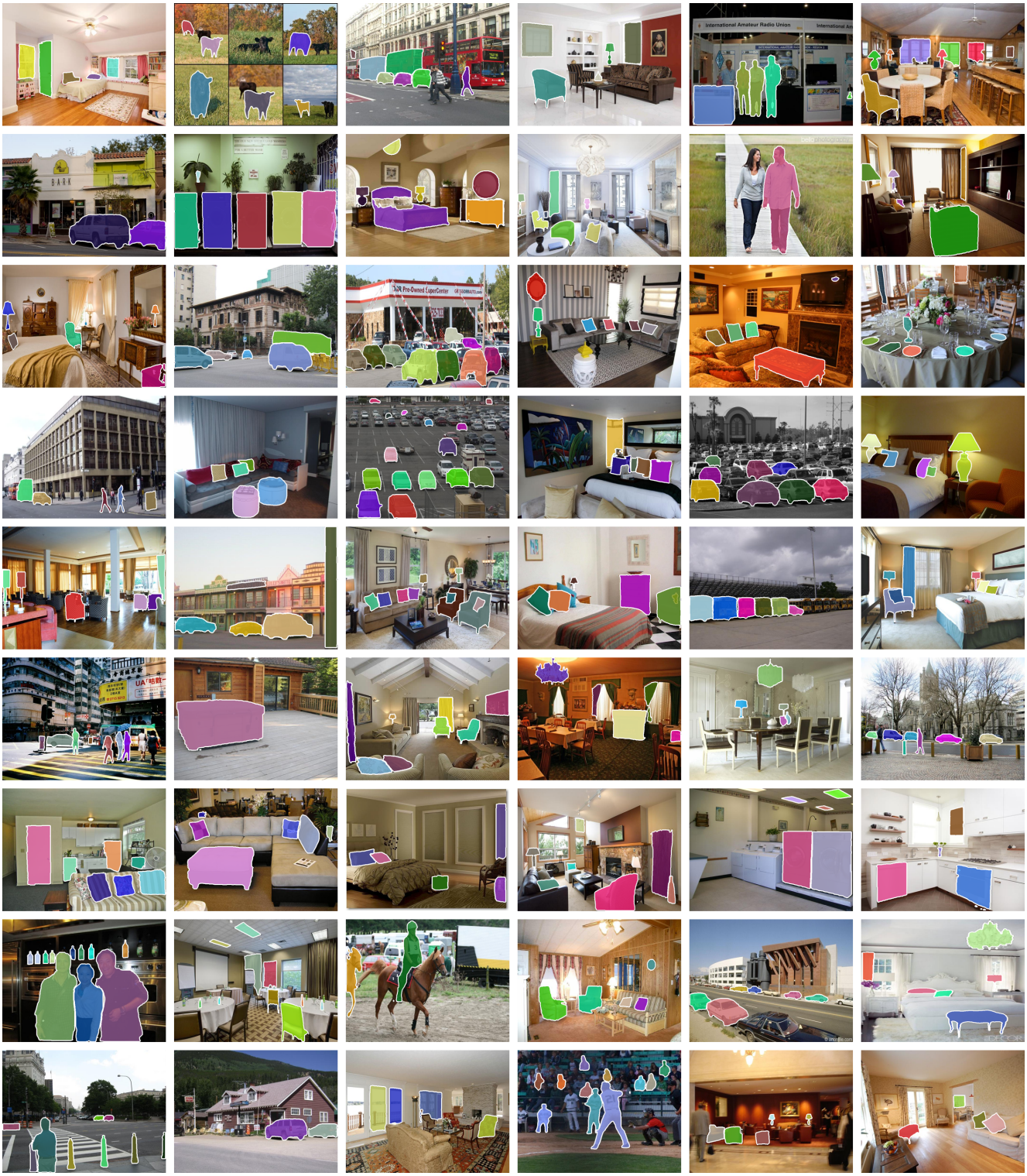

Figure 13. **Obtained mask annotations in Places (Part 5/5).** We show here a random selection of example images from the Places dataset with our obtained mask annotations using our pipeline under a small fixed annotation budget.