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 and 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 details about the crowd-sourcing implementation for our large-scale experiment and we present extra annotation and clustering examples.

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. 2(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.

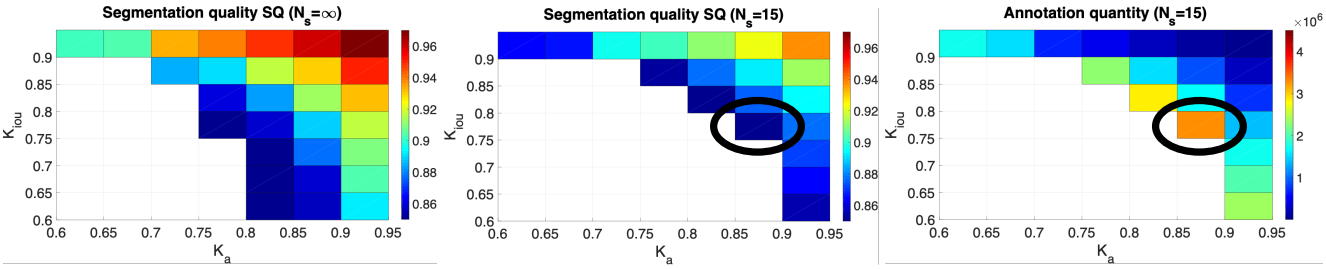
To ensure good quality, the annotators first read a simple set of instructions with several examples (Fig. 2(b)). Then, they go through a simple qualification test, at the end of which we provide detailed feedback on how well they performed. Annotators who successfully pass this test can proceed to the 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 submit if they fail to achieve a high accuracy.

Batching HITs and estimating quality. The tree search and annotation procedure described in Sec. 4 of the main paper is performed in a completely sequential way meaning that only one cluster is selected and annotated at a time. In practice, to run our pipeline on millions of masks on AMT, we efficiently batch AMT HITs (Human Intelligence Tasks) and obtain many cluster quality estimates at the same time. For every object category, we perform our search and choose up to 30 candidate clusters to annotate according to our selection criteria and ordering. We ask $N_s = 15$ questions per cluster, and after we obtain the annotators responses, we automatically update the cluster qualities. On subsequent annotation rounds, we repeat this process, sampling up to 30 clusters, but we reuse responses where pos-

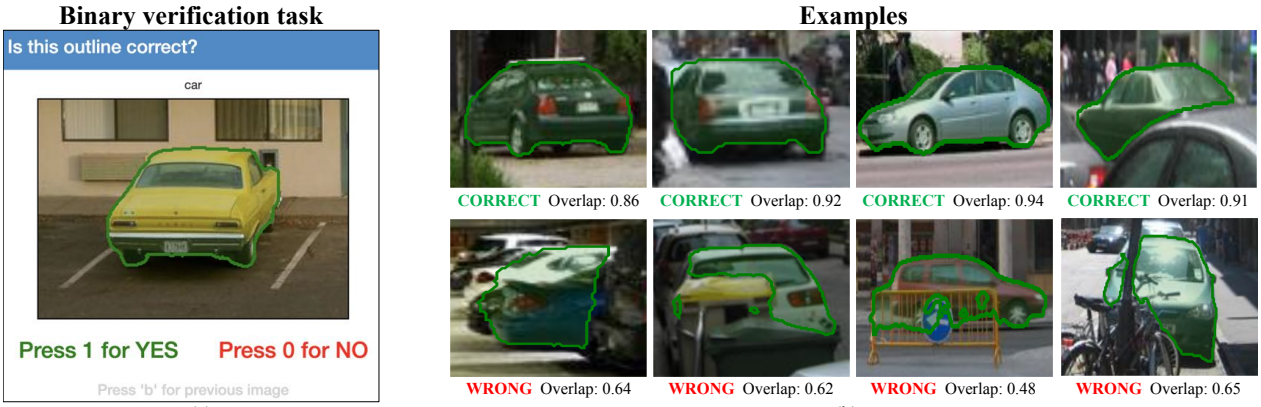
*Denotes equal contribution

[1] https://mturk.com/



(a) (b) (c)

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.



(a) (b)

Figure 2. **Crowd-sourcing 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 AMT annotators with many instruction examples and an interactive qualification test.

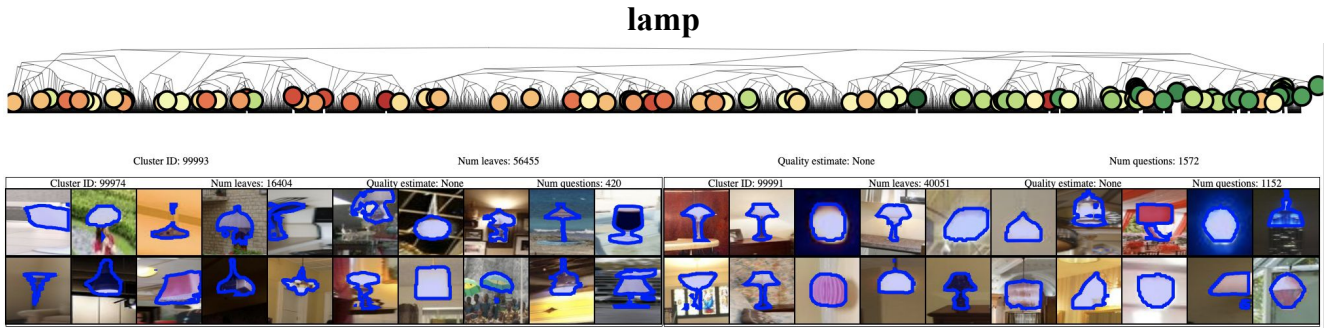


Figure 3. **Cluster tree viewer.** (Top) We show a subtree for the lamp category. The annotated clusters of the tree are color-coded according to the quality estimate (green for high quality and red for low). Notice that clusters without colored dots have either been split early based on the cluster score S or 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.

sible to ask the minimum number of questions to satisfy the N_s per cluster. By asking questions in batches and automatically updating the cluster qualities based on human responses, the Places results in the main paper were obtained in only a few days. In the end, we ran 6 rounds of batched HITs for each object category and the AMT annotation process for each batch took up to 6 hours. With a higher annotation budget, we could continue this process for more rounds to reach high-quality clusters appearing deeper in the tree and obtain more high quality masks.

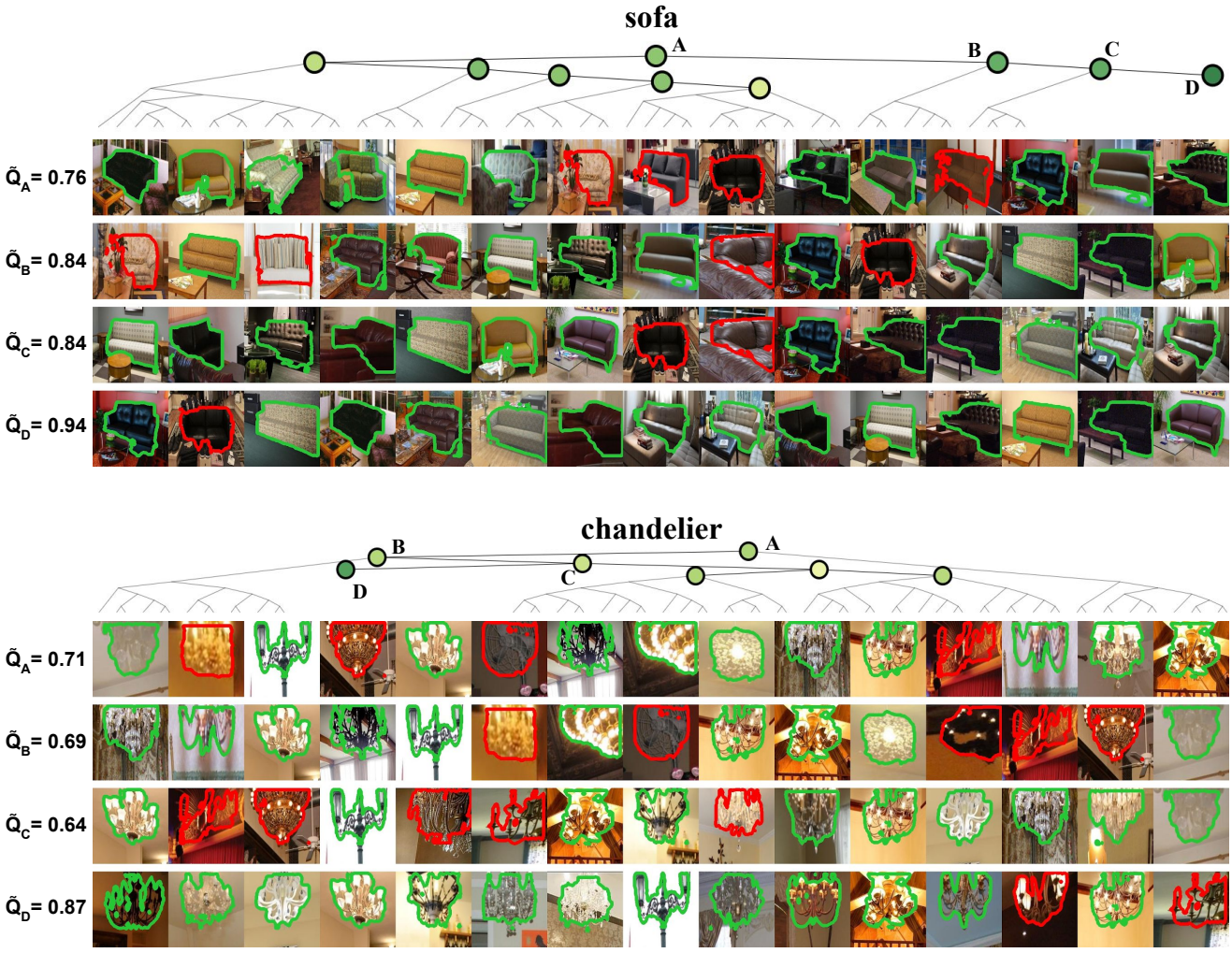


Figure 4. **Cluster examples (Part 1/2).** For each category (sofa and chandelier), we show the subtree for cluster A. Below, we show the quality estimate based on AMT responses. Green outlines mean positively verified, and red outlines mean negatively verified. Notice that clusters D are high quality ($\tilde{Q} \geq 0.85$), pruned, and added to our obtained dataset.

Clustering qualitative results. In Fig. 3(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. 3(bottom)). The selected displayed cluster is near the actual root node of T, and it contains 56,455 masks so the appearance and the quality of the masks vary a lot.

In Fig. 4, we show examples of clusters with their corresponding quality estimates for the object categories sofa and chandelier. Specifically, the green outlined masks are positively verified by human annotators on AMT and the red outlined masks are negatively verified. In each case, we show the subtree for clusters A and we show how the tree search procedure leads to the clusters D which are of high-quality ($\tilde{Q} \geq 0.85$). The part of the trees below the clusters D are pruned and the masks they contain are added to our obtained annotations.

In Fig. 5, we show an example subtree for the object category bicycle. We show the quality estimate $\tilde{Q}$ for three clusters based on the AMT human responses. The cluster A is a high-quality cluster, the part of the tree below it is pruned and the masks it contains are added to our obtained annotations. The clusters B and C are lower quality clusters. Note that the parent clusters of A, B, and C do not have quality estimates, meaning that our algorithm decided to split them without any human intervention based on the cluster scores S.

Mask annotation examples. In Fig. 6 and 7, we show class-specific, cropped annotation masks that are obtained

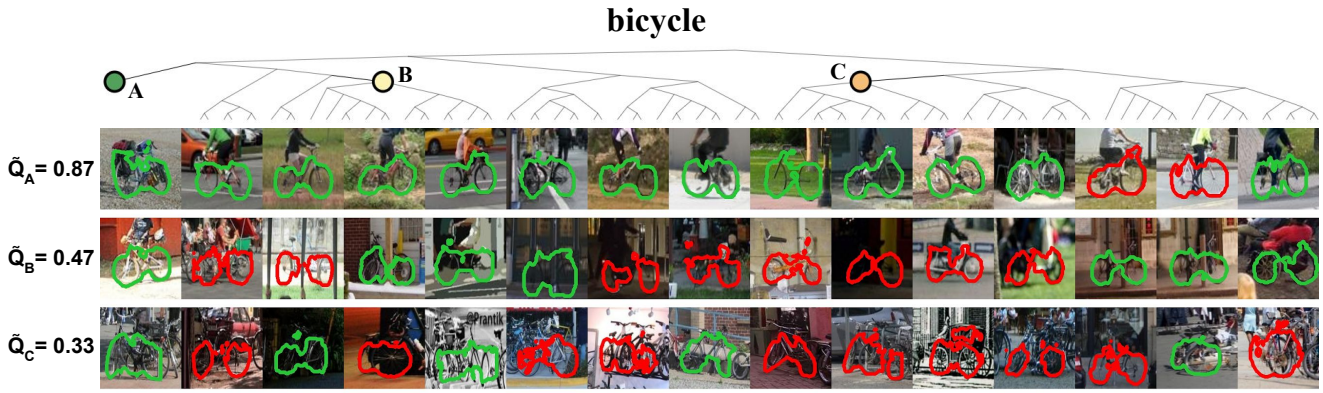


Figure 5. **Cluster examples (Part 2/2).** For the bicycle category, we show the quality estimate for 3 clusters based on AMT human responses. Cluster A is high quality and added to our database. Clusters B and C are low quality but not rejected. Notice that the parent clusters of A, B, and C do not have quality estimates, meaning our algorithm split early based on the cluster scores S.

for 36 different object categories in the 1M unlabeled Places images. In Fig. 8, 9 and 10, we show our obtained mask annotations in full images of the Places dataset using our proposed pipeline. Notice that some images are more sparsely annotated than others, as shown in Fig. 11(a) of the main paper. We expect density coverage to improve by increasing the annotation budget and further annotating the tree deeper in the Places experiment.

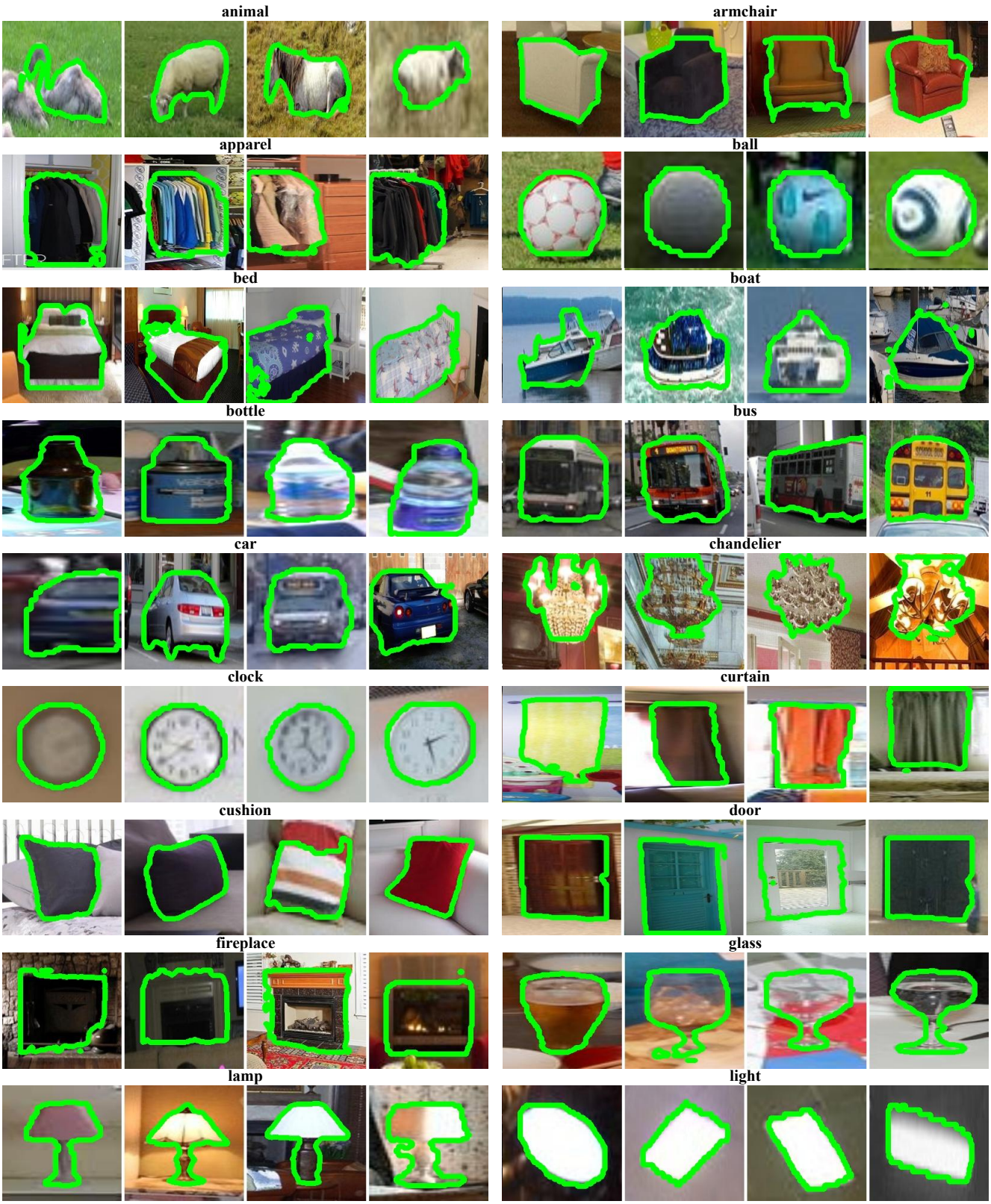


Figure 6. **Category annotations (Part 1/2).** We show four category-specific mask annotations for 18 different object categories that are obtained in the 1M unlabeled Places images.

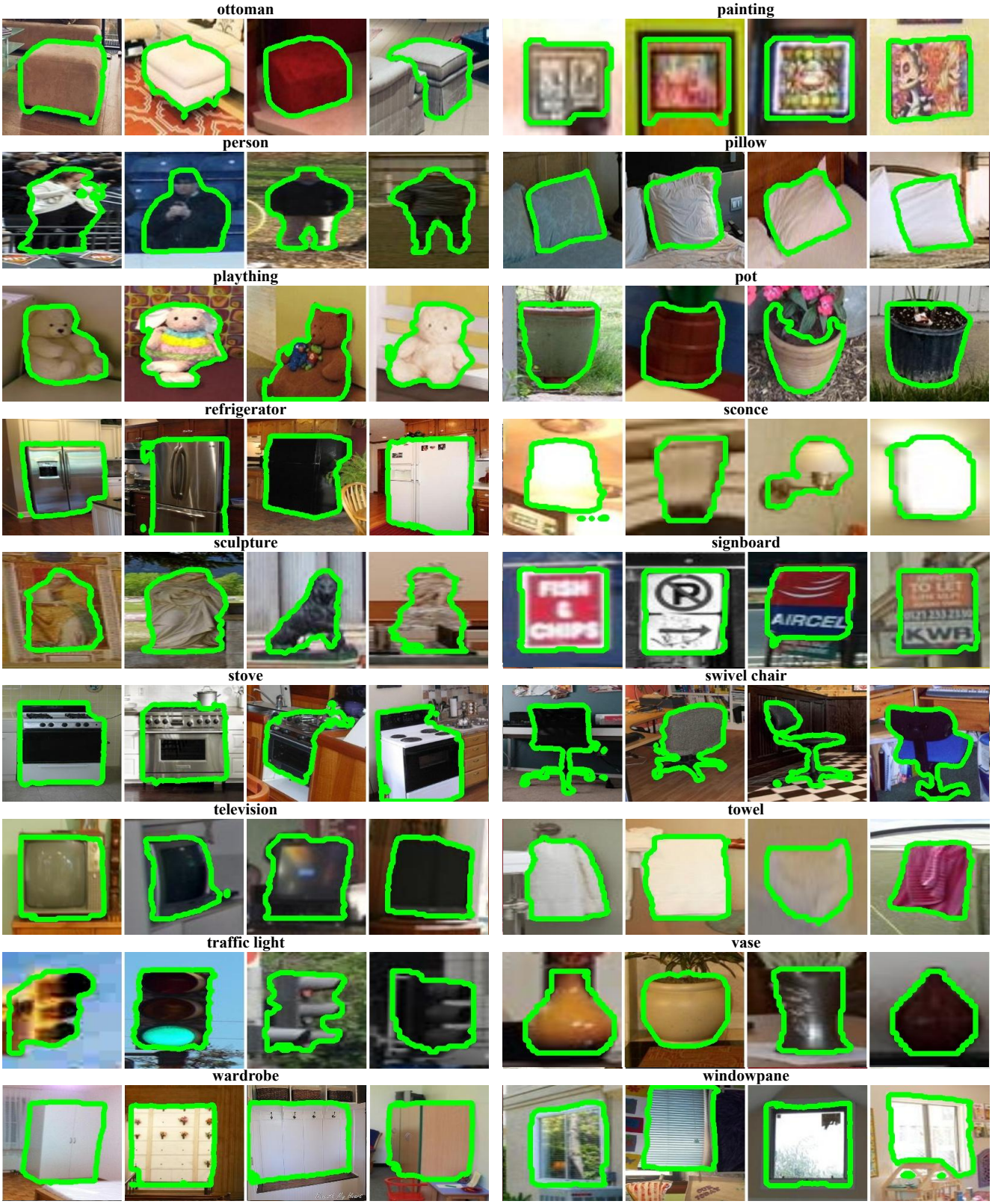


Figure 7. **Category annotations (Part 2/2).** We show four category-specific mask annotations for 18 different object categories that are obtained in the 1M unlabeled Places images.

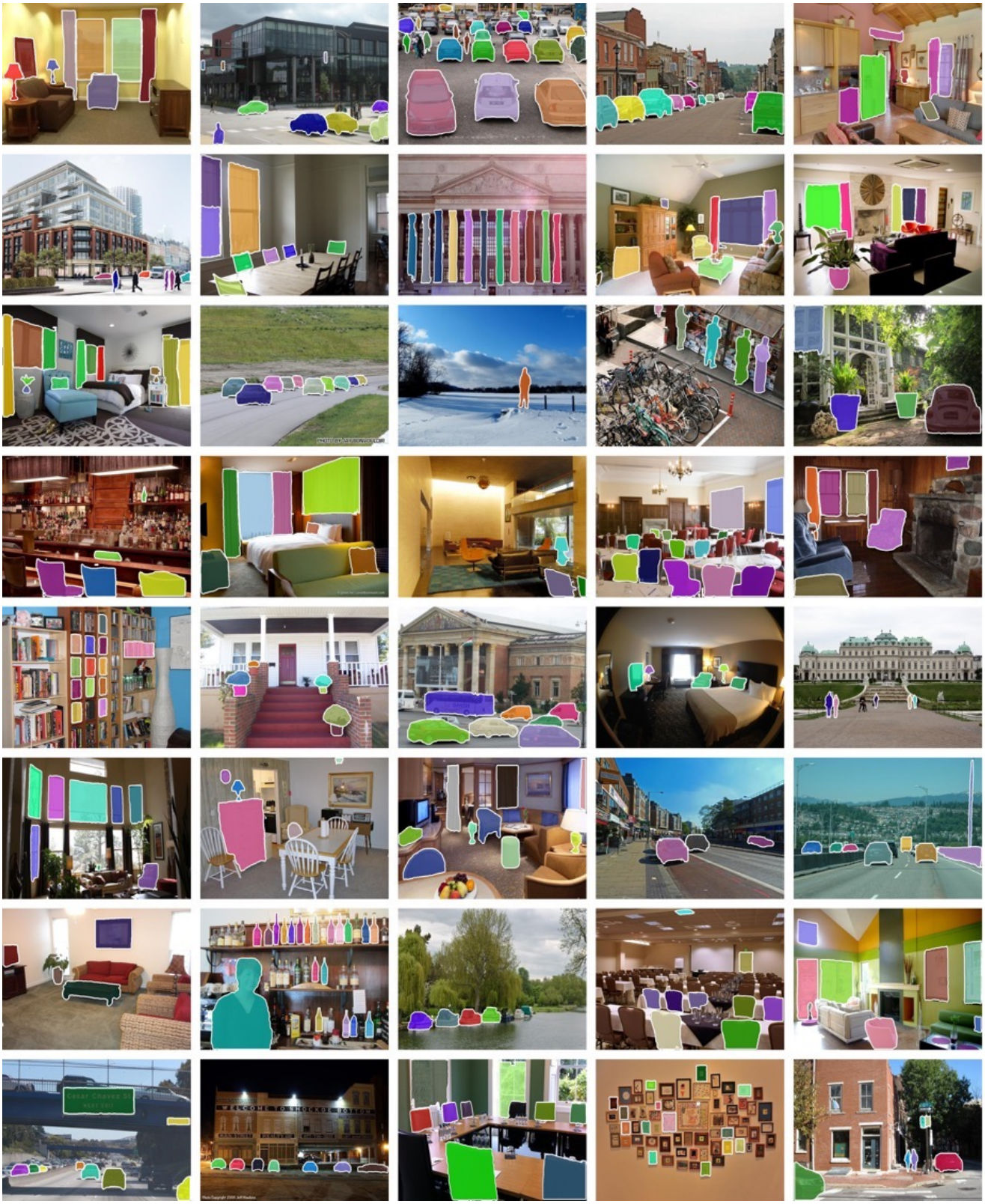


Figure 8. **Obtained mask annotations in Places (Part 1/3).** We show here example images from the 1M unlabeled Places images with our obtained mask annotations using our pipeline under a small fixed annotation budget.

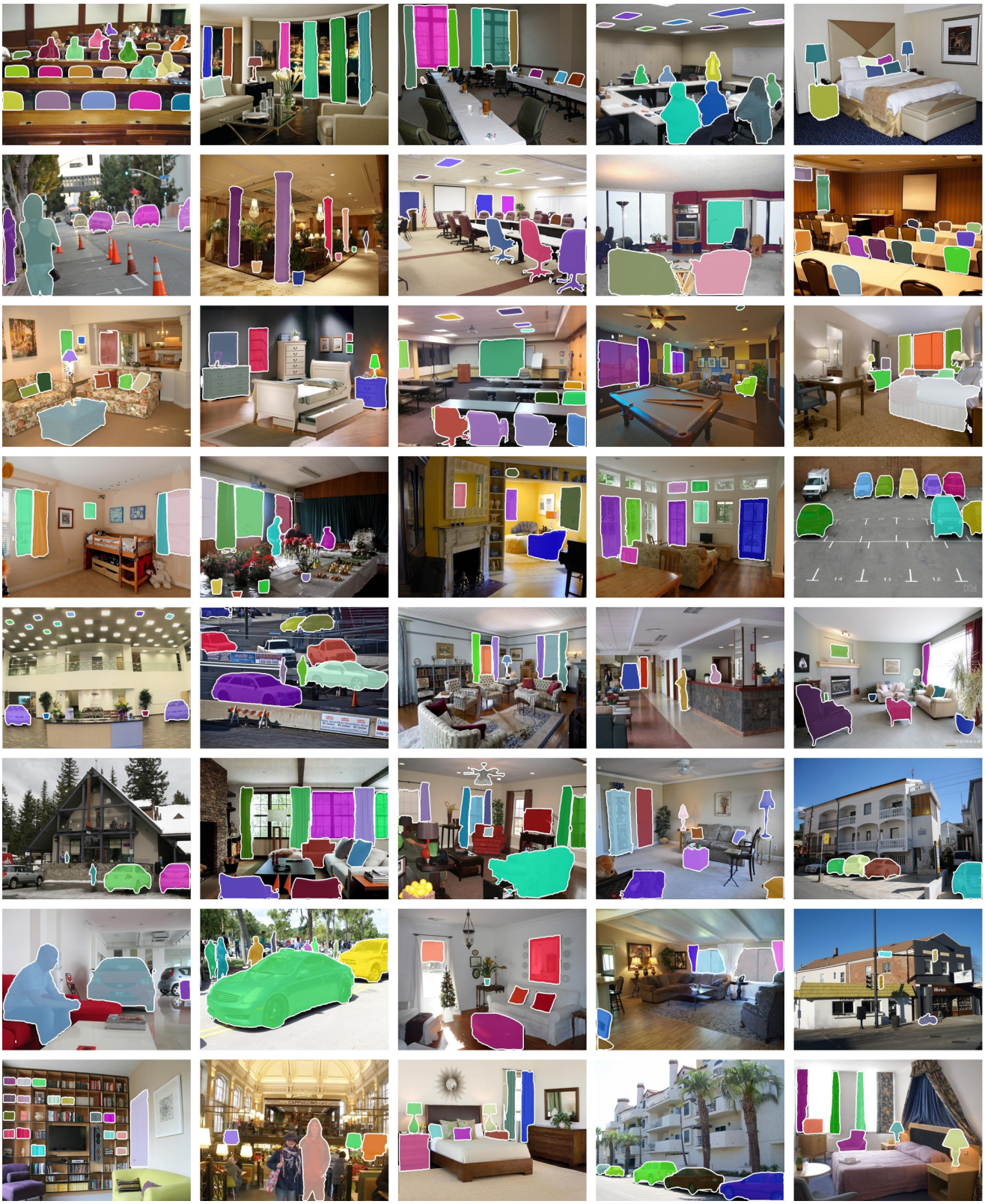


Figure 9. **Obtained mask annotations in Places (Part 2/3).** We show here example images from the 1M unlabeled Places images with our obtained mask annotations using our pipeline under a small fixed annotation budget.

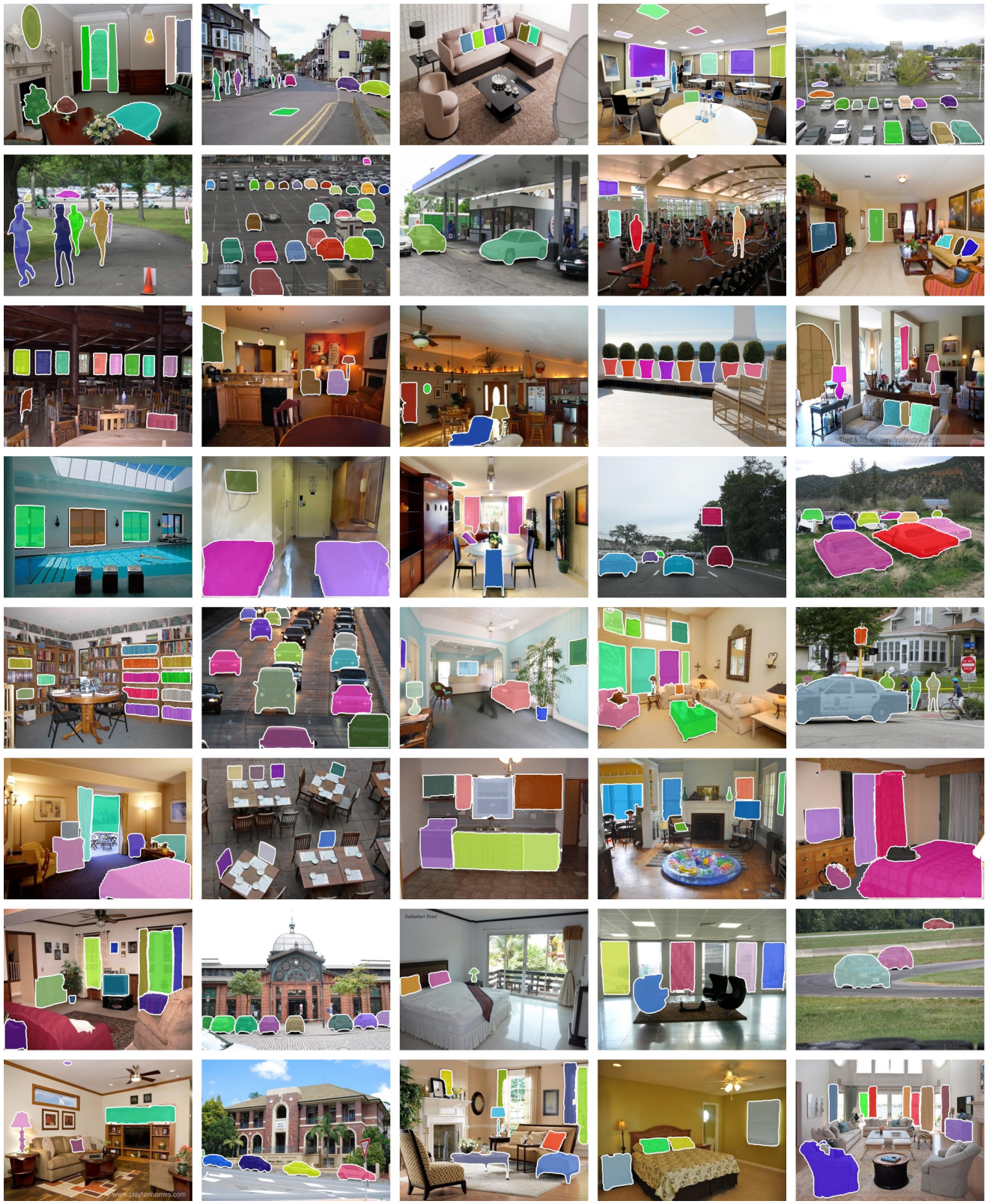


Figure 10. **Obtained mask annotations in Places (Part 3/3).** We show here example images from the 1M unlabeled Places images with our obtained mask annotations using our pipeline under a small fixed annotation budget.