Supplementary Materials:
Open Vocabulary Scene Parsing

1. Data association protocol

To learn the joint embeddings of images and word concepts, we need to augment ADE20K dataset by adding information about how the label classes (> 3000) are semantically related. We associate each class in ADE20K dataset with a synset in WordNet, representing a unique concept. The data association process requires semantic understanding, so we resort to Amazon Mechanical Turk (AMT). The annotation protocol is detailed as follows, and screen shots of our AMT interface are shown in Figure 1.

We search for each class in the dataset, for all the synsets having the same name. We find 3 different cases: (1) a single synset is found for the given class; (2) multiple synsets are found due to polysemy; (3) no sysnets are found, either because the correct synset has a different name or because that concept is not in WordNet.

In the first case, we automatically match classes in the dataset with the obtained synsets, and then ask workers on AMT to verify by looking at the image labels and the definitions of synsets in the WordNet.

In the second case where multiple synsets were found, we show an image displaying such concept and ask workers to select the synset whose definition matches the given class.

In the last case where no synset candidate was found, we show an image with the concept and ask workers to find the best matching synset by looking over WordNet online API. They also have the option to indicate when no synset can match.

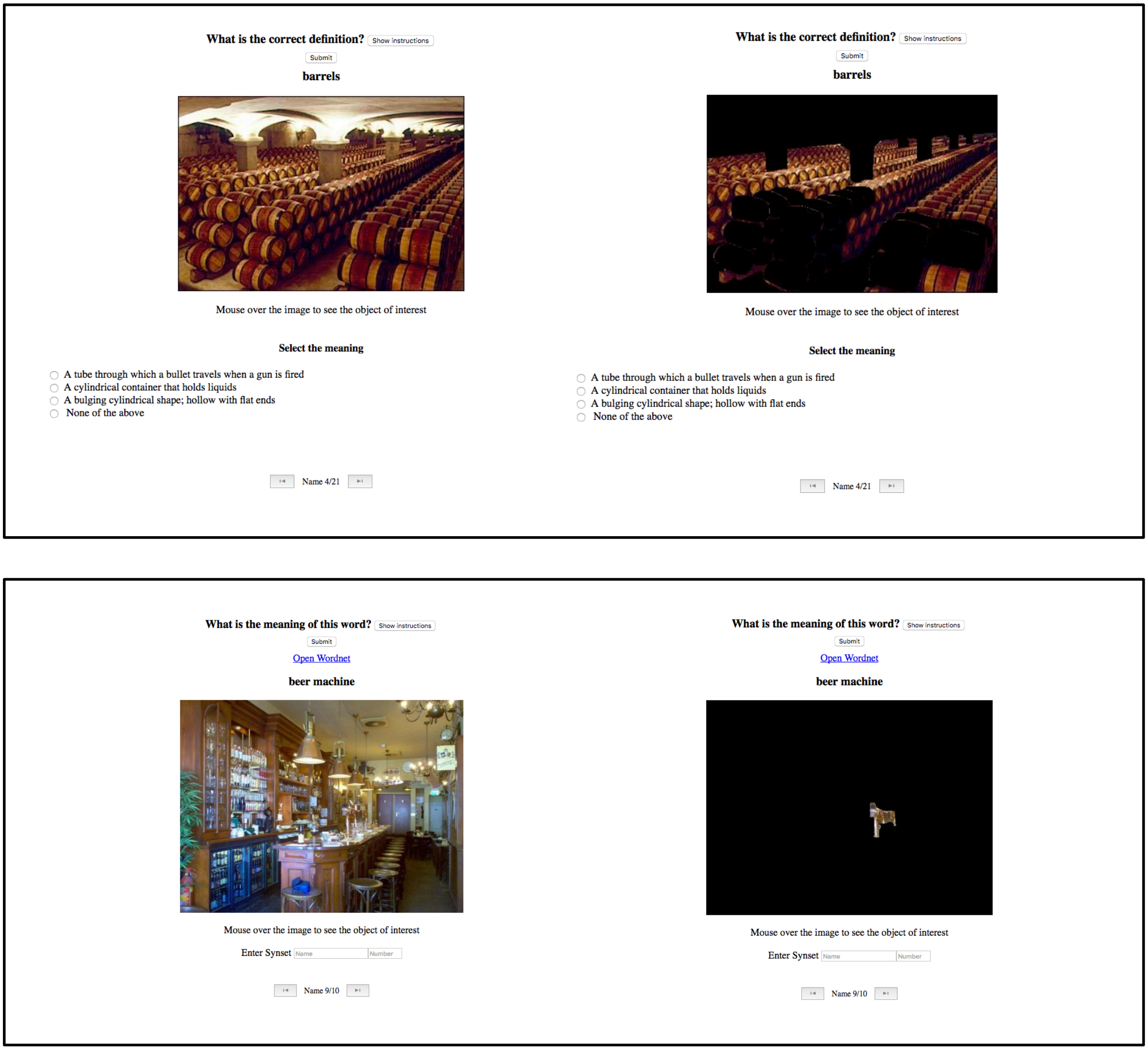
2. Concept graph

After data association, we end up with 3019 classes in the dataset having synset matches. Out of these there are 2019 unique synsets forming a DAG. All the matched synsets have *entity.n.01* as the top hypernym and there are in average 8.2 synsets in between. The depths of the ADE20K dataset annotations range from 4 to 19.

A detailed visualization of the concept graph built is shown in Figure 2. The node radii indicate the class frequencies in the ADE20K dataset. The figure only shows part of the full graph, nodes with 5 descendents or less have been hidden.

3. Full zero-shot predictions

Our model gives each sample a list of predictions in hierarchical order. Due to the space limitation, full prediction lists are not shown in the main paper. In Figure 3, we give details of zero-shot predictions, both ground truth and prediction lists are shown in the texts beneath the images. Correct predictions are marked in green, inconsistent items are marked in orange. It can be seen that for hard examples, *e.g.* “dome” (row1, column3), a general and conservative prediction is made; when the test sample is easy and similar to training samples, *e.g.* “wagon” (row1, column1), our model gives specific and aggressive predictions.



What is the correct definition?

barrels

Mouse over the image to see the object of interest

Select the meaning

○ A tube through which a bullet travels when a gun is fired
○ A cylindrical container that holds liquids
○ A bulging cylindrical shape; hollow with flat ends
○ None of the above

Name 4/21

What is the correct definition?

barrels

Mouse over the image to see the object of interest

Select the meaning

○ A tube through which a bullet travels when a gun is fired
○ A cylindrical container that holds liquids
○ A bulging cylindrical shape; hollow with flat ends
○ None of the above

Name 4/21

What is the meaning of this word?

Open Wordnet

beer machine

Mouse over the image to see the object of interest

Enter Synset Name Number

Name 9/10

What is the meaning of this word?

Open Wordnet

beer machine

Mouse over the image to see the object of interest

Enter Synset Name Number

Name 9/10

Figure 1. Screen shots of AMT interface for data association.

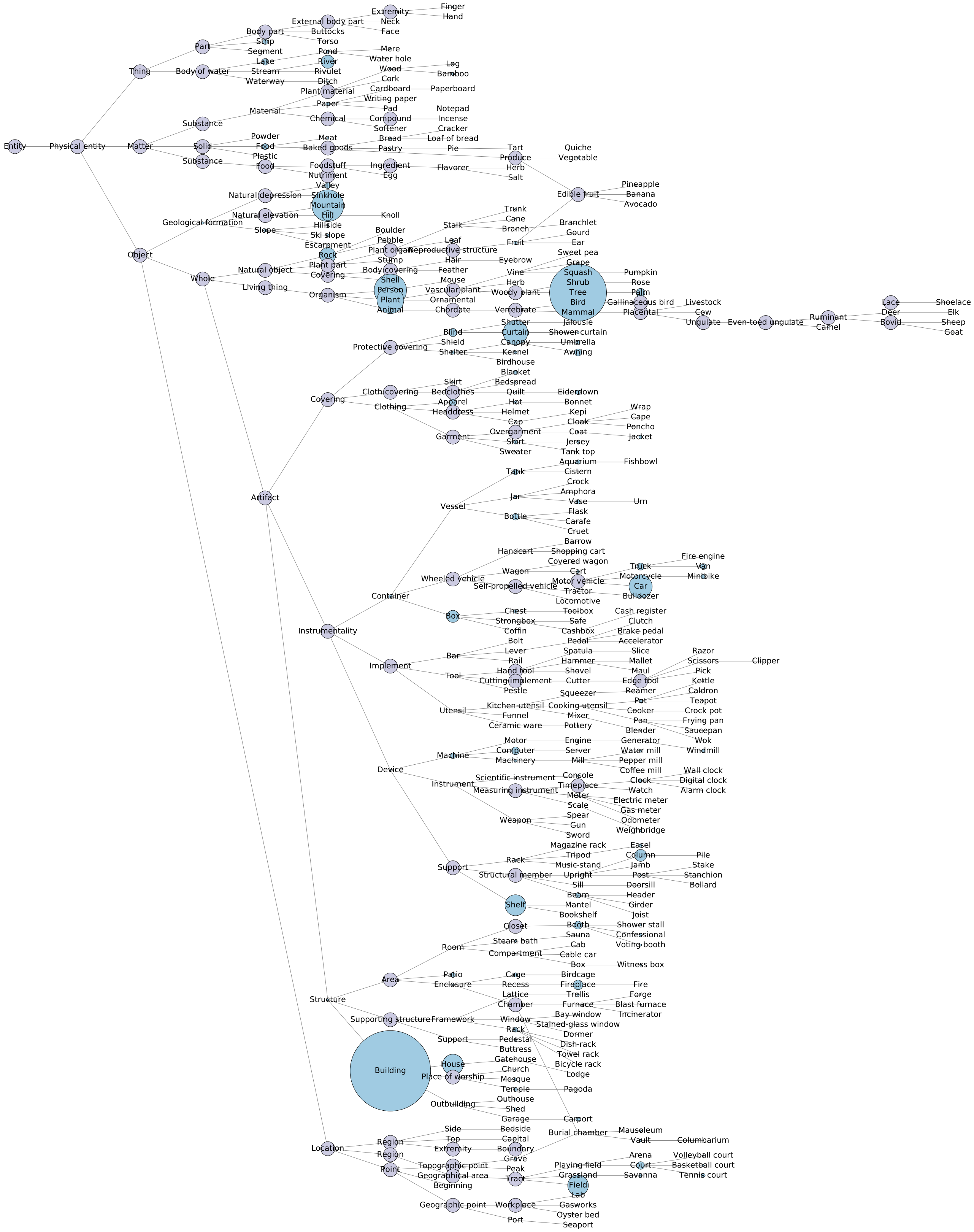


Figure 2. Part of the concept graph built based on WordNet and ADE20K label frequencies.

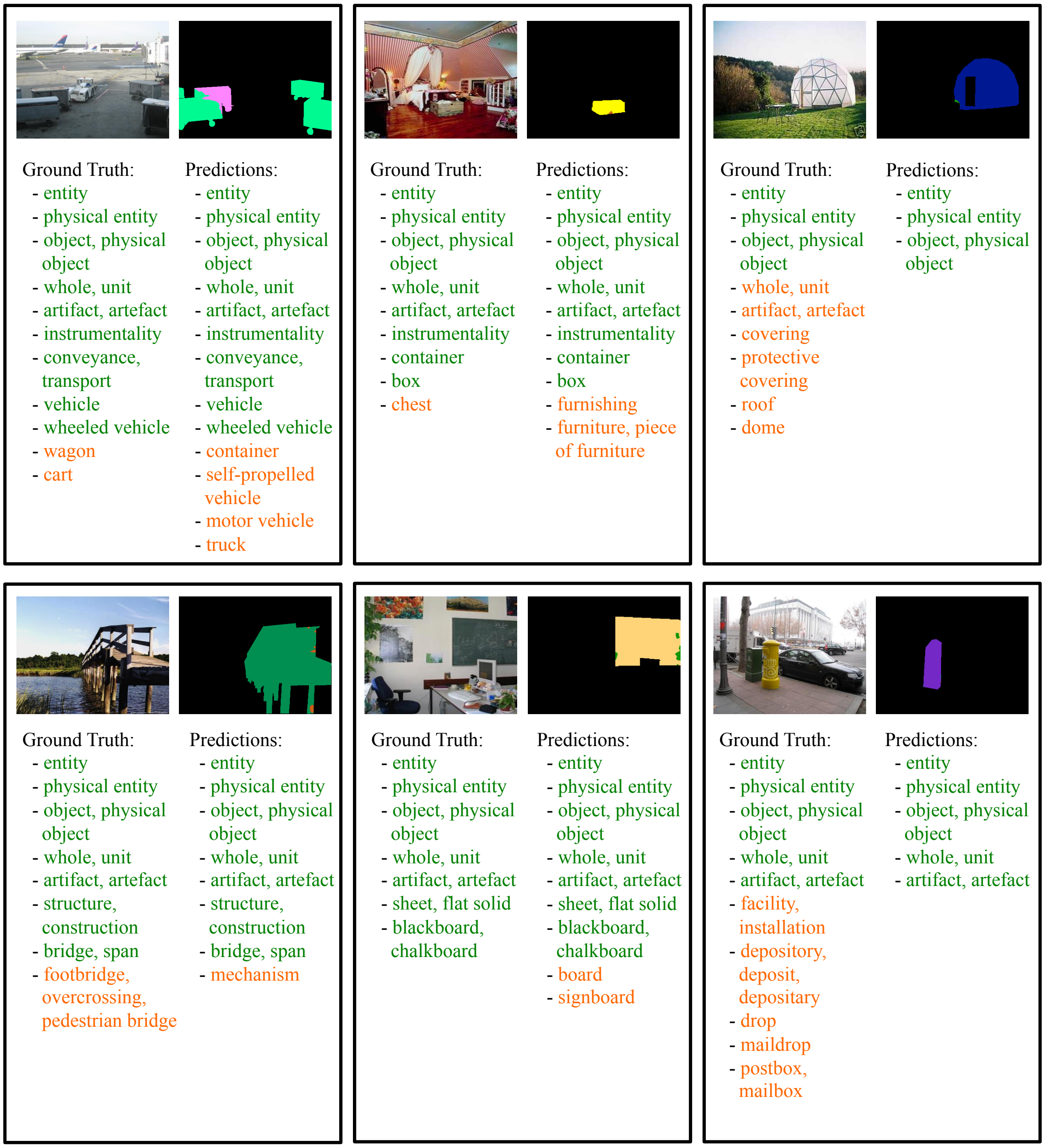


Figure 3. Full prediction results of zero-shot scene parsing.