Annotation Propagation in Large Image Databases via Dense Image Correspondence

Supplemental Material

1 Additional Results

Here we show more results of our system on the datasets we experimented with. Description of the datasets and the experiment setups are given in Section 5 in the paper. **We recommend viewing the results electronically** and zoom in for more details. Notice that some figures span more than one page. We indexed the figures in a grid to allow referencing particular results. We note that all the results are for images that are originally untagged and unlabeled in the database (the subset of images $I \setminus I_t$).

In Fig. 1 we show more results on LabelMe Outdoors (LMO) dataset [1]. In addition to the final result, we also show the MAP labels based on local evidence alone (the appearance model in Section 3, Eqn. 2), similar to Fig. 3(c) in the paper.

Fig. 2 shows the estimated spatial prior of each word (Eqn. 7) in LMO's vocabulary. It can be seen that the prior agrees with the true spatial prior, computed from the ground truth labels, for more frequent words. The estimated prior is somewhat blurrier than the ground truth, indicating some errors in classification, however the general layout is captured correctly. For example, *sky* is mostly at the top of the image, *building* is in the middle, and *road* and *sea* are at the bottom.

Fig. 3 shows more results on SUN dataset [2], as well as comparison with the results by [3], similar to Fig. 8 in the paper. The results of [3] were produced using the authors' original implementation (available online), modified by us to account for tagged images as described in Section 3 in their paper (termed "weak supervision"). Taken together with Fig. 1, these results show that the algorithm can handle large variety of both indoor and outdoor scenes. Notice that while SUN has a relatively large vocabulary (500+ words), the tags inferred by the algorithm tend to correspond to words with higher frequency in the dataset. That is because words that occur frequently, and co-occur frequently with other words, are considered more probable by the algorithm (Eqn. 3).

Fig. 4 and 5 show more results on the ESP game dataset [4] and IAPR benchmark [5], where we used the same images and vocabulary as in [6] (available online). These two datasets are much noisier in terms of both image content and vocabulary, and so are more challenging for the algorithm. In particular, both datasets include more abstract words (*e.g. smile*, *night*) that are harder to model, as well as words that might not correspond to a particular image region (*e.g. photo*).

Finally, Fig. 6 shows more failure cases on all datasets. Limitations of the system include incorrect classification under similar visual appearance or insufficient exemplars of particular words (*e.g.* row 1 columns 2-3, row 3 columns 1,3 in (a), row 1 column 1 in (b), row 2 columns 2-3 in (c)), and errors due to incorrect inter-image correspondence (*e.g.* row 1 column 1, row 5 columns 1,3 in (a), row 2 column 3 in (b), row 1 column 1 in (c)).

2

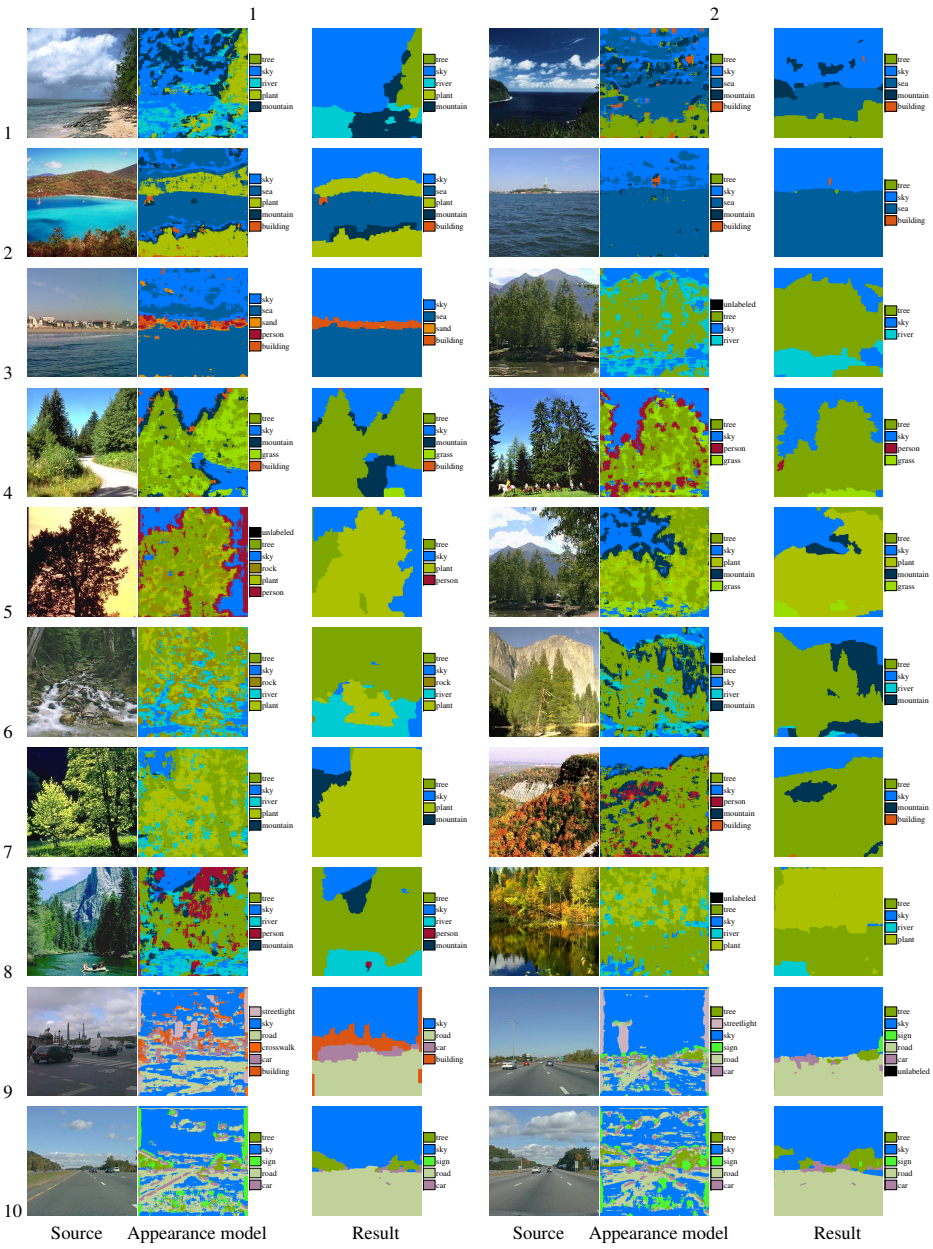

1
2
1
2
3
4
5
6
7
8
9
10
Source
Appearance model
Result
Source
Appearance model
Result

Fig. 1. More results on LMO. For each example, we show the source image on the left, the MAP labeling using the appearance model only in the middle (computed independently at each pixel; see Fig. 3 in the paper), and the final result of the annotation propagation algorithm (appearance model + spatial regularization + regularization via dense image correspondences) on the right. Note that the final result might not contain all tags from the appearance model result.

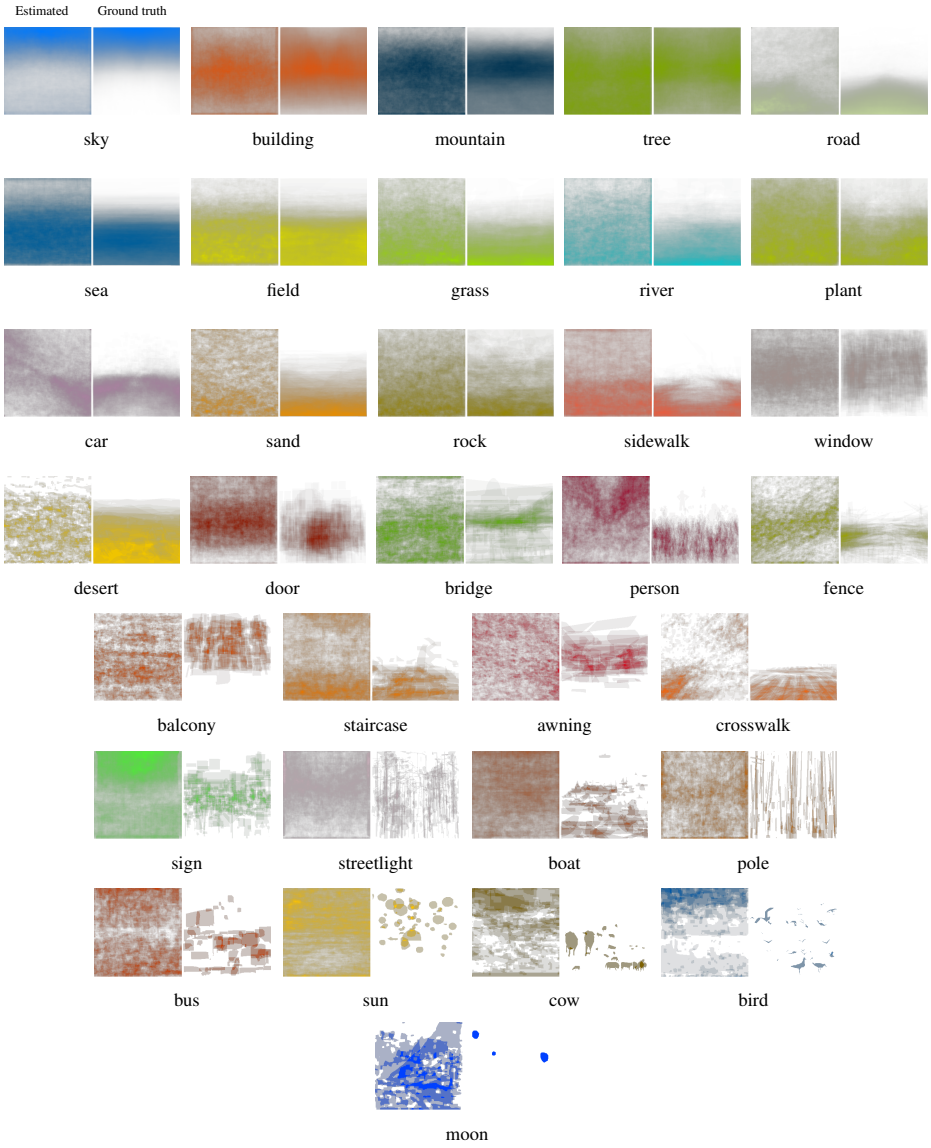
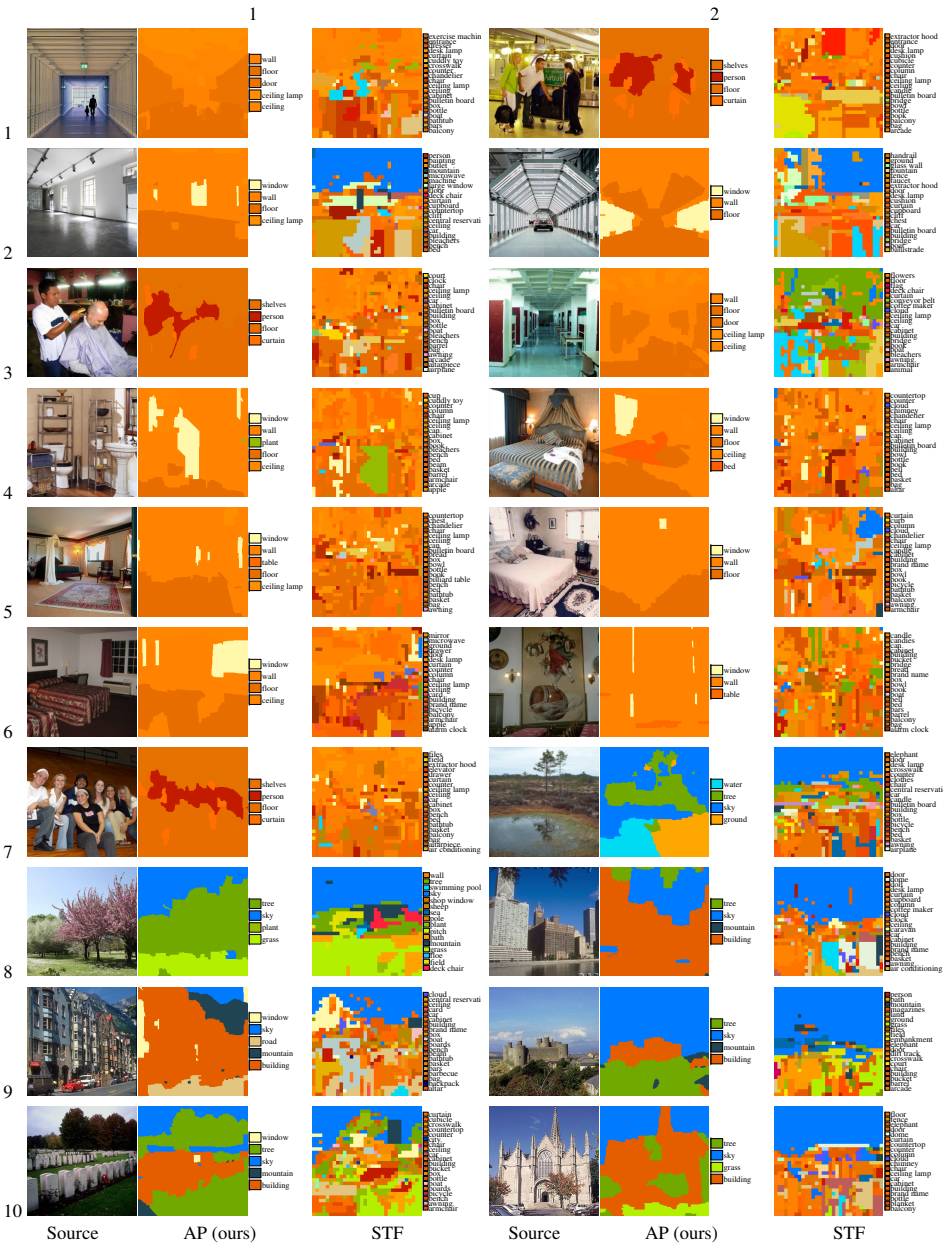
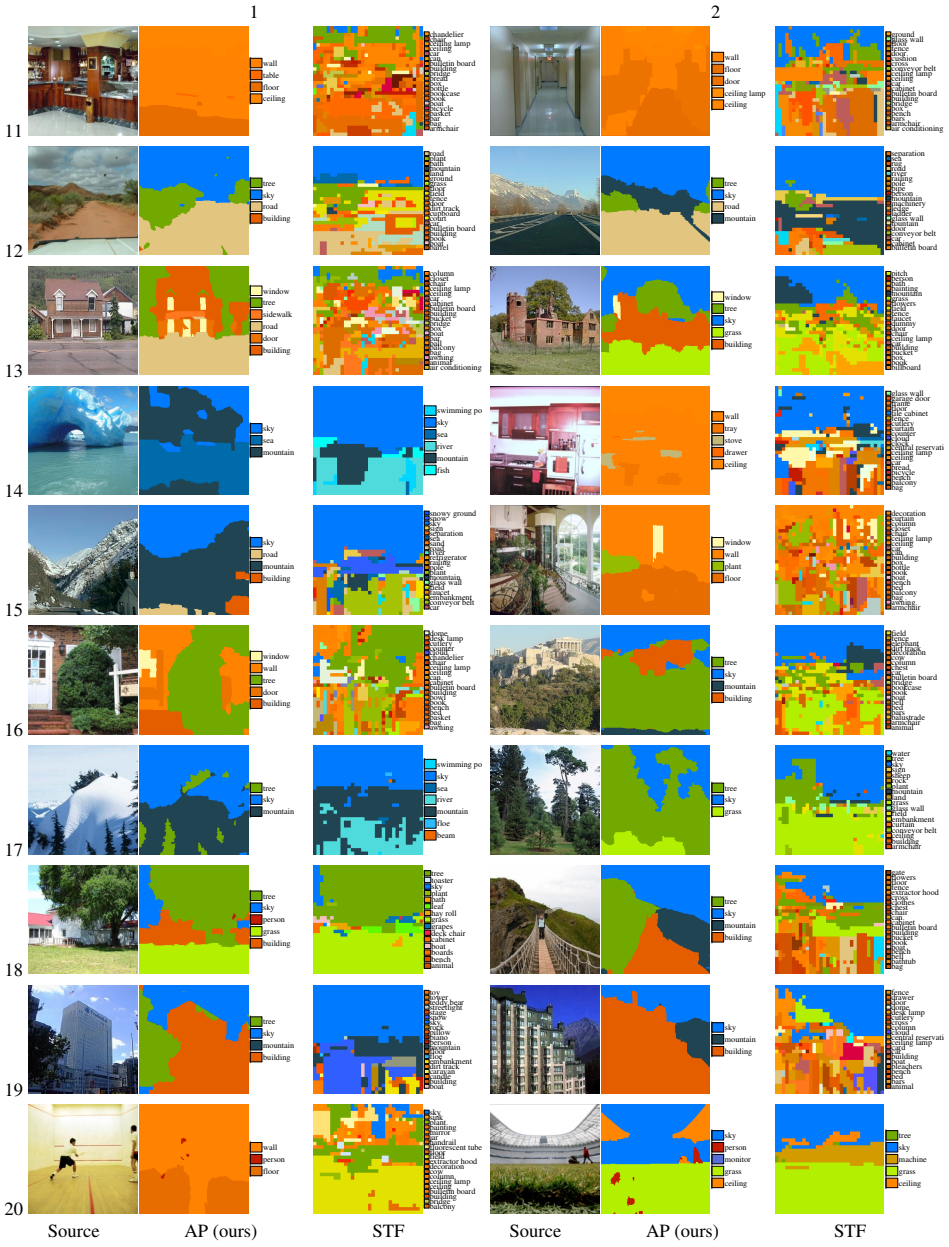


Fig. 2. The estimated spatial prior h_l^s (Eqn. 7) for the LMO vocabulary. Words are ordered from top left to bottom right according to their frequency in the dataset. For each word, the left image is the estimated prior and the right image is the true prior according to human labels. The colormap is the same as Fig. 1 above, with saturation corresponding to probability, from white (zero probability) to saturated (high probability).


1
2
1
2
3
4
5
6
7
8
9
10
Source
AP (ours)
STF
Source
AP (ours)
STF


1
2
11
12
13
14
15
16
17
18
19
20
Source
AP (ours)
STF
Source
AP (ours)
STF

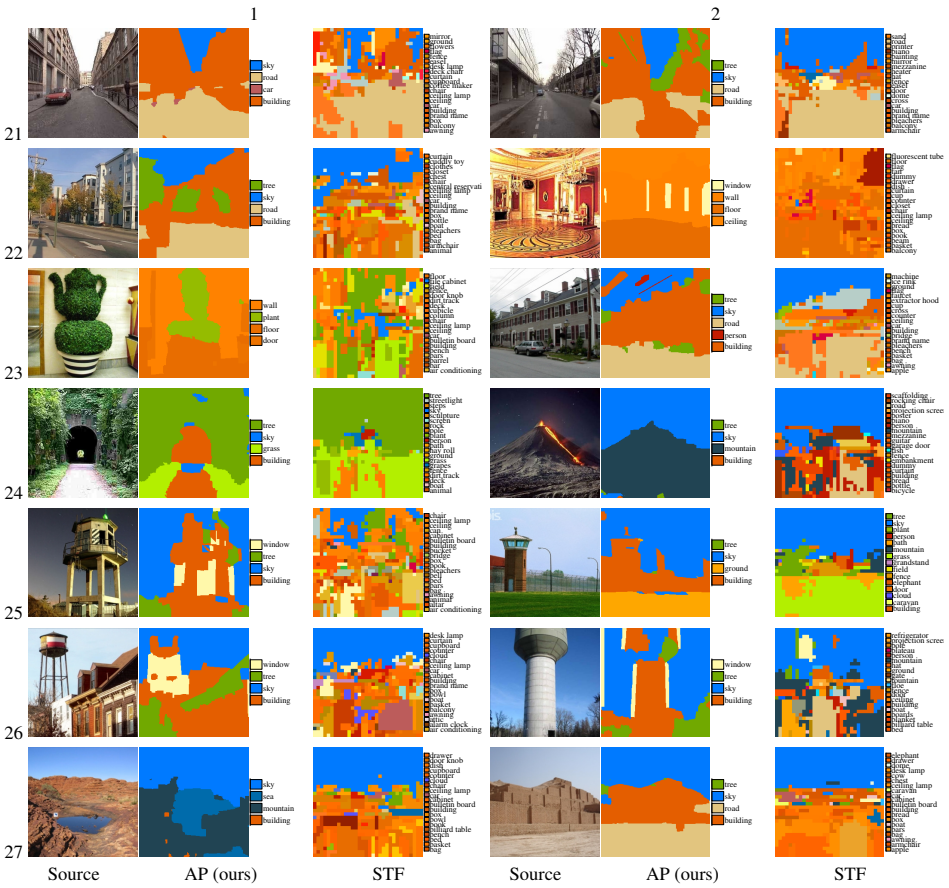


Fig. 3. More results on SUN, and comparison with semantic texton forests (STF) [3]. For some images, STF assigned too many tags for a clear visualization, and so we limited the legends to 20 words (omitting any excess words; for visualization purposes only).

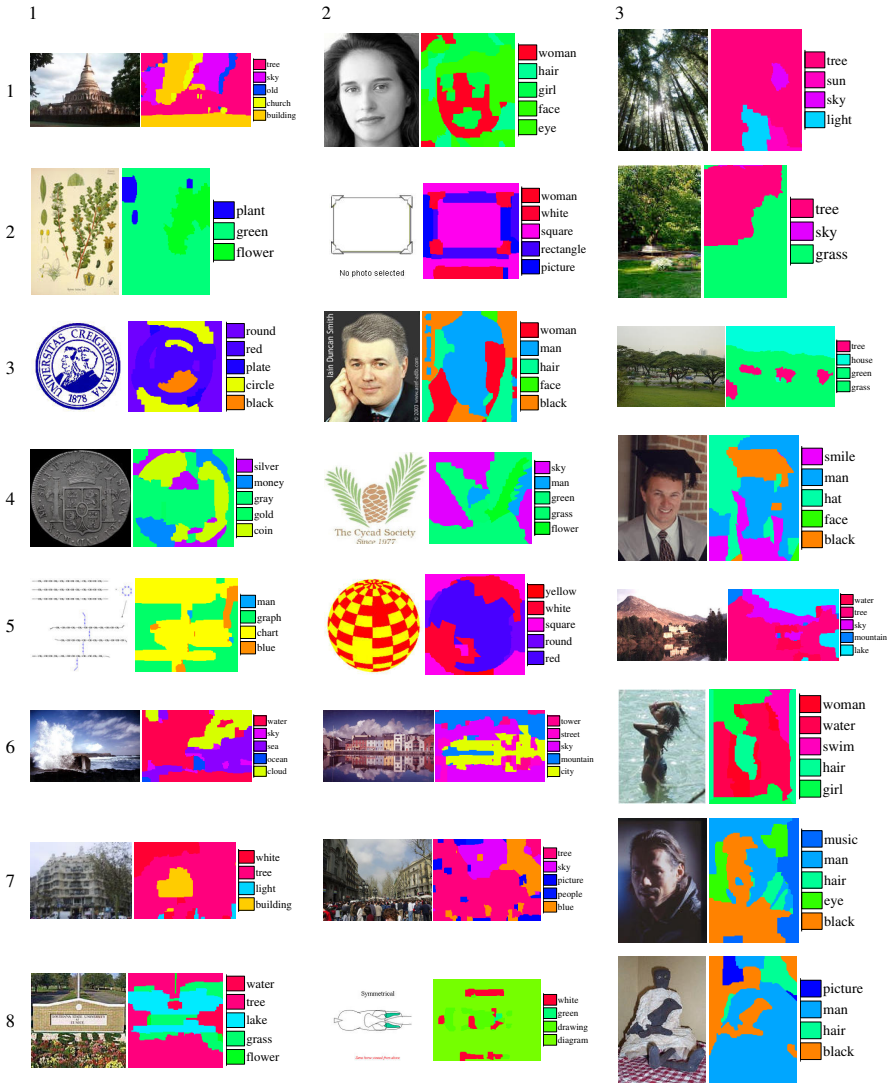
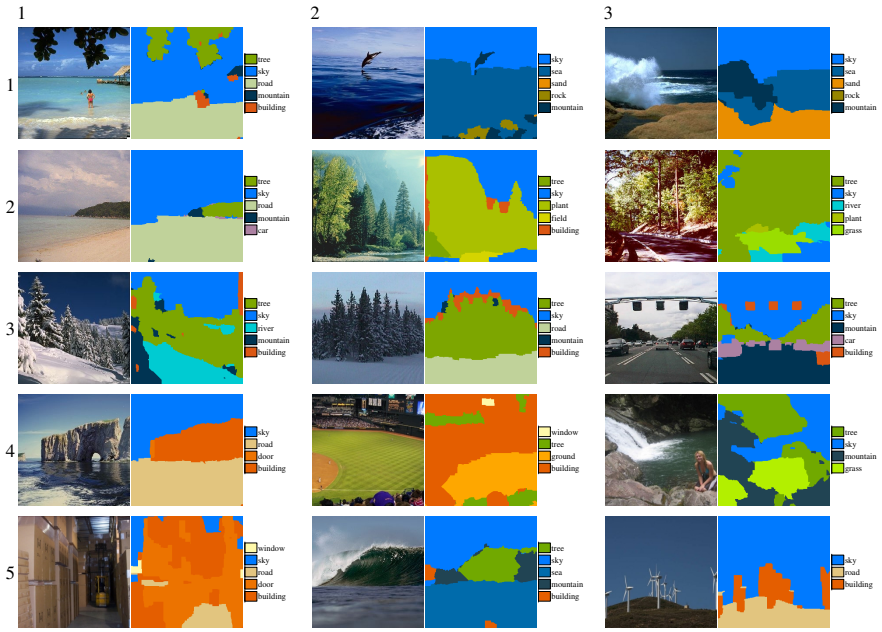


Fig. 4. More results on ESP.

Fig. 5. More results on IAPR.



(a) SUN (and LMO)

(b) ESP

(c) IAPR

Fig. 6. More failure cases on SUN, ESP and IAPR.

References

1. Russell, B., Torralba, A., Murphy, K., Freeman, W.: Labelme: a database and web-based tool for image annotation. IJCV **77** (2008) 157–173
2. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR. (2010) 3485–3492
3. Shotton, J., Johnson, M., Cipolla, R.: Semantic texton forests for image categorization and segmentation. In: CVPR. (2008)
4. Von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: SIGCHI. (2004)
5. Grubinger, M., Clough, P., Müller, H., Deselaers, T.: The iapr benchmark: A new evaluation resource for visual information systems. In: LREC. (2006) 13–23
6. Makadia, A., Pavlovic, V., Kumar, S.: Baselines for image annotation. IJCV **90** (2010)