Scene Category Context for 3D Object Detection with RGBD cameras

Carl J. Olsson
Department of Computer Science
Brown University
Providence, RI 02912
colsson@cs.brown.edu

Abstract

In the problem of 3D object detection for indoor scene understanding, contextual relationships such as scene category can provide informative clues that improve object detection performance. In this paper, we extend our previous 3D object detection framework [7] to leverage positive context clues, by integrating scene appearance information with the appearance and geometry features from our previous work. As in the previous work, we test the effectiveness of our approach on the same SUN-RGBD dataset [8]. The experimental results demonstrate that through effective integration of positive context clues such as scene appearance, our approach achieves a significant improvement over the previous work's performance [7].

1 Introduction

With the advent of inexpensive RGB-D cameras and advances in algorithms for 2D object detection, interest in the subject of indoor scene understanding and 3D object detection has surged. This problem is important for a variety of reasons including but not limited to robotics. For example, in domestic robotics, a robot must be able to navigate indoor environments as well as interact with objects in these environments. In addition, accurate 3D object detection and localization is the basis for successful object grasping.

In our previous work, we demonstrated that 3D object detection and localization is possible with inexpensive RGB-D sensors. Our approach used a structured prediction framework with geometry and appearance features to learn an algorithm that aligns 3D cuboid hypotheses to RGB-D data. One major contribution of this previous work was a novel appearance feature called cloud of oriented gradients (COG) [7], an extension of the popular 2D histogram of oriented gradients (HOG) descriptor to 3D RGB-D images [2]. Simple contextual relationships between other object detections and categories were then utilized in a second cascade classifier to reduce false positives, further improving object detection accuracy to beyond that of a state-of-the-art CAD-model detector [9].

In this paper, we extend our previous 3D object detection framework to leverage positive context clues. Specifically, we are interested in exploiting the contextual relationships between scene appearance and objects. As in the previous work, we represent the objects in an indoor scene in terms of 3D cuboids and model various appearance and geometry features of objects. In this work, we integrate new scene appearance features to model the contextual relationships between objects and scene appearance. This approach encourages object labels to agree with the scene appearance (i.e. beds are often found in bedrooms).

We evaluate our approach on the challenging and large SUN-RGBD dataset [8] and show significant improvements over our previous work's performance, demonstrating the usefulness of the contextual relationship between scene category and objects.

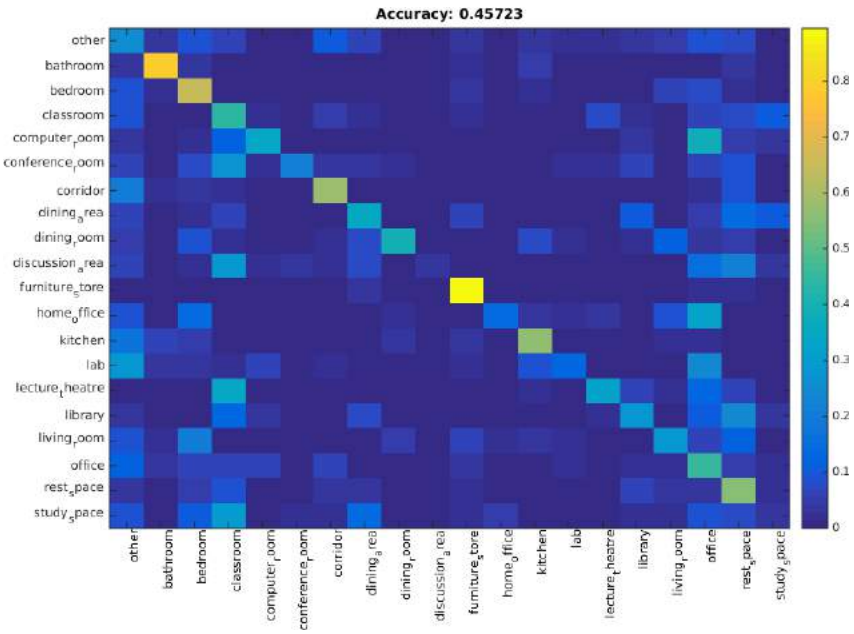

Figure 1: Confusion matrix for scene categorization. The figure title displays the average accuracy for classification.

1.1 Related Work

There have been many interesting papers on RGB-D scene understanding and 3D object detection [8]. In addition, scene categorization is a very popular task for scene understanding. However, few of them have incorporated scene appearance contextual clues into their object detection frameworks [6].

1.2 Dataset

Each of the 10,335 RGB-D images in the SUN-RGBD dataset are annotated with a scene category, a total of 44 categories ranging from living room to bathroom. For these images, we have 64,595 3D bounding boxes (with accurate orientations for objects) annotated. In addition, each RGB-D image has pre-computed deep features using the scene classification approach by Xiao et al. [11]. This feature is learned using a Deep Convolutional Neural Net [5] with 2.5 million scene images [12].

2 Scene categorization

For this task, we used the 19 scene categories with more than 80 images and an 'other' category, as in the SUN-RGBD paper to learn a model that predicts scene category for RGB-D images. We used the pre-computed deep features with a linear SVM [1] to achieve state-of-the-art scene categorization performance on the SUN-RGBD dataset [8], with an accuracy of 45.72% versus 39.00% of Song et al [8] over the same 19 scene categories as well as an additional 'other' category (See Fig. 1).

3 Feature-based modeling of geometry, layout, and appearance

Our object detectors are learned from 3D oriented cuboid annotations in the SUN-RGBD dataset [7]. We discretize each cuboid into a $6 \times 6 \times 6$ grid of (large) voxels, and extract features for these $6^3 = 216$ cells. Voxel dimensions are scaled to match the size of each instance.

As in our previous work, we use the standard descriptors for the 3D geometry of the observed depth image, the novel *cloud of oriented gradient* (COG) descriptor of RGB appearance, and the novel *Manhattan voxel* scene layout descriptor [7]. In addition, we introduce a new scene appearance feature of RGB-D images.

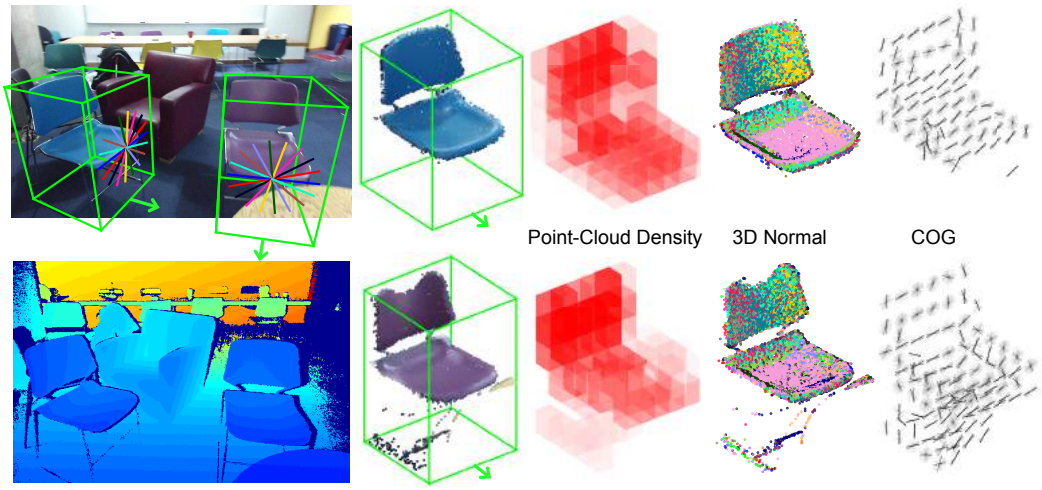

Figure 2: Given input RGB and Depth images (left), we align oriented cuboids and transform observed data into a canonical coordinate frame. For each voxel in a $6 \times 6 \times 6$ grid, we then extract (from left to right) point cloud density features, 3D normal orientation histograms, and our COG model of back-projected image gradient orientations. On the left, COG bins are colored to show alignment between instances. The value of the point cloud density feature is proportional to the voxel intensity, each 3D orientation histogram bin is assigned a distinct color, and COG feature intensities are proportional to the normalized energy in each orientation bin, similarly to HOG descriptors [2].

3.1 3D geometry and appearance features

As in our previous work [7], we use the point cloud density feature $\phi^a_{i\ell}$, 25 surface normal histogram features $\phi^b_{i\ell}$, and 9 COG appearance features $\phi^c_{i\ell}$ (See Fig. 2).

3.2 Scene layout geometry features

As in our previous work [7], we use the novel *Manhattan voxel* discretization for 3D layout prediction. As such, the overall space is discretized in $12 \times 6 = 72$ bins [7].

3.3 2D scene appearance feature

In order to integrate new scene appearance features that model contextual relationships between objects and scene appearance, we introduce a new descriptor over the 20 scene categories s in Sec. 2 as

$$\phi_s(s = u) = \sigma(t_u) \tag{1}$$

Where t_u denotes the classifier score for scene category u and σ is the multinomial logistic function. We utilize the pre-computed deep features from the SUN-RGBD dataset [8] with the linear SVM [1] from Sec. 2. We do not apply any scaling or normalization.

4 Learning to detect cuboids & layouts

For each voxel ℓ in some cuboid B_i annotated in training image I_i, we have one point cloud density feature $\phi^a_{i\ell}$, 25 surface normal histogram features $\phi^b_{i\ell}$, and 9 COG appearance features $\phi^c_{i\ell}$. Our overall feature-based representation of cuboid i is then $\phi(I_i, B_i) = \{\phi^a_{i\ell}, \phi^b_{i\ell}, \phi^c_{i\ell}\}_{\ell=1}^{216}$. Cuboids are aligned via annotated orientations as illustrated in Fig. 2, using the gravity direction provided in the SUN-RGBD dataset [8]. Similarly, for each of the Manhattan voxels ℓ in layout hypothesis M_i we compute point cloud density and surface normal features, and $\phi(I_i, M_i) = \{\phi^a_{i\ell}, \phi^b_{i\ell}\}_{\ell=1}^{72}$.

4.1 Structured prediction of cuboids

As in the previous work [7], we learn a prediction function $h_c : I \rightarrow B$ that maps an RGB-D image I to a 3D bounding box $B = (L, \theta, S)$ for each object category c independently, using those images which contain visible instances of that category,. Here, L is the center of the cuboid in 3D, θ is the cuboid orientation, and S is the physical size of the cuboid along the three axes determined by its orientation. We assume objects have a base upon which they are usually supported, and thus θ is a scalar rotation with respect to the ground plane.

Given n training examples of category c, we use an n-slack formulation of the structural support vector machine (SVM) objective [4] with margin rescaling constraints:

$$\min_{w_c, \xi \geq 0} \quad \frac{1}{2} w_c^T w_c + \frac{C}{n} \sum_{i=1}^{n} \xi_i \quad \text{subject to}$$
$$w_c^T [\phi(I_i, B_i) - \phi(I_i, \bar{B}_i)] \geq \Delta(B_i, \bar{B}_i) - \xi_i,$$
$$\text{for all } \bar{B}_i \in \mathcal{B}_i, i = 1, \ldots, n. \quad (2)$$

Here, $\phi(I_i, B_i)$ are the features for oriented cuboid hypothesis B_i given RGB-D image I_i, B_i is the ground-truth annotated bounding box, and $\mathcal{B}_i$ is the set of possible alternative bounding boxes.

For training images with multiple instances, we add images multiple times to the training set, each time removing the subset of 3D points contained in other instances, as in previous work on 2D detection [10].

Cuboid loss function Given some ground truth cuboid B and estimated cuboid $\bar{B}$, we use the following loss function from the previous work [7]:

$$\Delta(B, \bar{B}) = 1 - \text{IOU}(B, \bar{B}) \cdot \left(\frac{1 + \cos(\bar{\theta} - \theta)}{2} \right). \quad (3)$$

Here, $\text{IOU}(B, \bar{B})$ is the volume of the 3D intersection of the cuboids, divided by the volume of their 3D union. The loss is bounded between 0 and 1, and is smallest when the $\text{IOU}(B, \bar{B})$ is near 1 *and* the orientation error $\theta - \bar{\theta} \approx 0$. Loss approaches 1 if either position or orientation is wrong. We solve the loss-sensitive objective of Eq. (2) using a cutting-plane method [4].

Cuboid hypotheses We precompute features for candidate cuboids in a sliding-window fashion using discretized 3D world coordinates, with 16 candidate orientations.

We discretize cuboid size using empirical statistics of the training bounding boxes: $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ width quantiles, $\{0.25, 0.5, 0.75\}$ depth quantiles, and $\{0.3, 0.5, 0.8\}$ height quantiles. Every combination of voxel size, and 3D location and orientation, is then evaluated.

4.2 Structured prediction of manhattan layouts

We again use the S-SVM formulation of Eq. (2) to predict Manhattan layout cuboids $M = (L, \theta, S)$. The loss function $\Delta(M, \bar{M})$ is as in Eq. (3), except we use the "free-space" definition of IOU from [8], and account for the fact that orientation is only identifiable modulo 90° rotations. Because layout annotations do not necessarily have Manhattan structure, the ground truth layout is taken to be the cuboid hypotheses with largest free-space IOU [7].

Layout hypotheses We predict floors and ceilings as the 0.001 and 0.999 quantiles of the 3D points along the gravity direction, and discretize orientation into 18 evenly spaced angles between 0 and 180°. We then propose layout candidates that capture at least 80% of all 3D points, and are bounded by the farthest and closest 3D points. For typical scenes, there are 5,000-20,000 layout hypotheses [7].

5 Cascaded learning of spatial context

As in the previous work [7], we adapt *cascaded classification* [3] to the modeling of contextual relationships in 3D scenes. In this approach, "first-stage" detections as in Sec. 4 become input features

										
Sliding-Shape [9]	42.95	19.66	20.60	28.21	60.89	-	-	-	-	-
Geom	8.29	15.06	26.20	24.53	1.15	-	-	-	-	-
Geom+COG	52.98	28.64	42.16	45.14	43.00	28.17	7.93	14.25	12.83	47.69
Geom+COG+Context-5	58.72	44.04	42.50	54.81	63.19	-	-	-	-	-
Geom+COG+Context-10	61.29	48.68	49.80	59.03	66.31	44.58	12.97	25.14	30.05	56.78
Geom+COG+Context-10+Layout	63.67	51.29	51.02	62.17	70.07	45.19	15.47	27.36	31.80	58.26
With softmax scene category	67.94	52.04	51.41	63.33	74.98	46.15	18.09	30.30	**33.85**	65.30
With ground truth scene category	**70.35**	**52.75**	**52.25**	**63.46**	**80.53**	**47.18**	**21.03**	**30.71**	33.75	**69.21**

Table 1: Average precision scores for all object categories, from left to right: *bed, table, sofa, chair, toilet, desk, dresser, night-stand, bookshelf, bathtub*. Notice that using the new softmax scene appearance descriptor ϕ_s boosts performance and that using a one-hot ground truth scene category feature outperforms the softmax scene descriptor ϕ_s in all but one category.

to "second-stage" classifiers that estimate confidence in the correctness of cuboid hypotheses. We also integrate our new scene appearance features ϕ_s as additional contextual features.

Contextual features For an overlapping pair of detected bounding boxes B_i and B_j, we denote their volumes as $V(B_i)$ and $V(B_j)$, their volume of their overlap as $O(B_i, B_j)$, and the volume of their union as $U(B_i, B_j)$. We characterize their geometric relationship via three features: $S_1(i,j) = \frac{O(B_i,B_j)}{V(B_i)}$, $S_2(i,j) = \frac{O(B_i,B_j)}{V(B_j)}$, and the IOU $S_3(i,j) = \frac{O(B_i,B_j)}{U(B_i,B_j)}$. To model object-layout context [6], we compute the distance $D(B_i, M)$ and angle $A(B_i, M)$ of cuboid B_i to the closest wall in layout M.

The first-stage detectors provide a most-probable layout hypothesis, as well as a set of detections (following non-maximum suppression) for each category. For a bounding box B_i with confidence score z_i, there may be several overlapping bounding boxes of categories $c \in \{1, \ldots, C\}$. Letting i_c be the instance of category c with maximum confidence z_{i_c}, features ψ_i for bounding box B_i are created via a quadratic function of z_i, $S_{1:3}(i, i_c)$, $A(B_i, M)$, and a radial basis expansion of $D(B_i, M)$. In addition, the new scene appearance feature ϕ_s is appended to ψ_i.

Contextual learning During training, each detected bounding box for each class is marked as "true" if its intersection-over-union score to a ground truth instance is greater than 0.25, and is the largest among those detections. We train a standard binary SVM with a radial basis function (RBF) kernel

$$K(B_i, B_j) = \exp\left(-\gamma ||\psi_i - \psi_j||^2\right). \quad (4)$$

The bandwidth parameter γ is chosen to be $\gamma = 0.1$ using validation data and the second-stage layout predictor is trained as in the previous work [7].

Contextual Prediction During testing, given the set of cuboids found in the first-stage sliding-window search, we apply the second-stage cascaded classifier to each cuboid B_i to get a new contextual confidence score z_i'. The overall confidence score used for precision-recall evaluation is then $z_i + z_i'$, to account for both the original belief from the geometric and COG features and the correcting power of contextual cues. The second-stage layout prediction is directly provided by the second-stage S-SVM classifier [7].

6 Experiments

We test our cascaded model on the SUN RGB-D dataset [8] and compare with the state-of-the-art *sliding shape* [9] cuboid detector and the baselines from the previous work [7]. As before, we learn and evaluate RGB-D appearance models of 10 object categories, five more than [9]. Object cuboid and 3D layout hypotheses are generated and evaluated as described in previous sections.

To evaluate the detection performance, we calculate the intersection-over-union score on 3D bounding boxes and consider the predicted bounding box to be correct when the score is above 0.25. We provide several comparisons to demonstrate the effectiveness of our new scene appearance feature ϕ_s, and the importance of both appearance and context features.

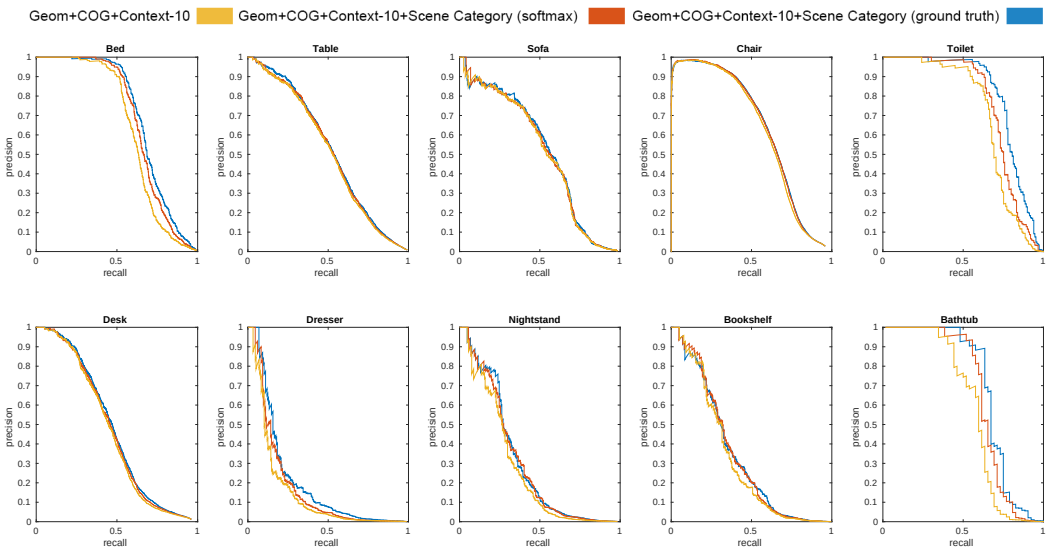

Figure 3: Precision-recall curves for 3D cuboid detection of the 5 object categories considered by [9] (top), and 5 additional categories (bottom). We also test the importance of modeling scene category context with our new softmax scene category descriptor ϕ_s and a one-hot ground truth scene category feature. See legend at top.

6.1 The importance of scene appearance

To test the importance of modeling scene category we trained our detector with our new softmax scene appearance descriptor ϕ_s. There is a very clear improvement in detection accuracy for the 10 object categories tested (see precision-recall curves in Fig. 3).

In addition, we trained our detector with a one-hot scene category feature that used the ground truth scene category. This was to test the potential of improving our scene classification in Sec. 2. Unsurprisingly, the ground truth scene category feature outperformed the performance of the new softmax scene category descriptor ϕ_s in most categories and had lower performance on one object category.

Despite this, the performance gain of using ground truth scene category is marginal for most categories besides toilets and beds, which have a strong contextual relationship with scene category (i.e. beds are often in bedrooms and toilets in bathrooms). On the same token, modeling scene category doesn't seem to hurt object detection performance for object categories that don't have strong contextual relationships with scene category (i.e. bookshelfs and desks appear in many scene categories).

6.2 Learned scene category weights

To elucidate what our detector learned about scene category in regards to object detection, we visualized the learned weights of the new softmax scene category descriptor ϕ_s for 3D cuboid detection of the toilet and bookshelf object categories, the categories had the most and least improvement of using the ground truth scene category ($+5.55\%$ and -0.1% respectively).

The learned weights of the new softmax scene appearance descriptor ϕ_s of the toilet detector (shown in Fig. 4) support our intuition that modeling scene category can provide informative clues that improve object detection performance (i.e. toilets are often found in bathrooms and not in other types of scenes). Likewise, the learned weights of the new softmax scene appearance descriptor ϕ_s of the bookshelf detector implies that bookshelfs don't have strong contextual relationships with scene category and shows that modeling scene category doesn't seem to hurt object detection performance.

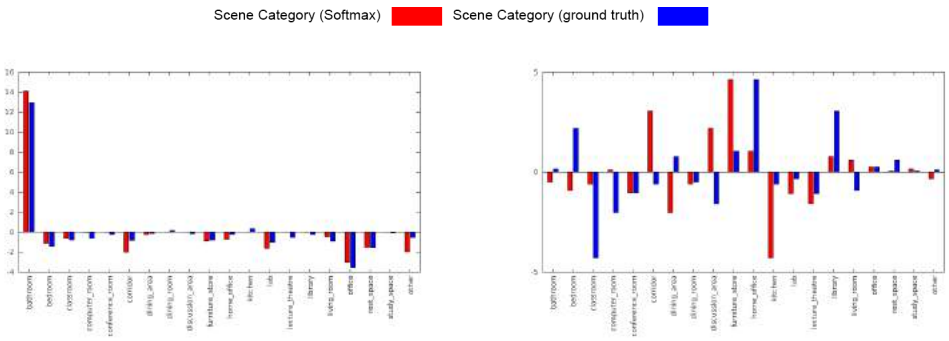

Figure 4: Learned weights of the new softmax scene category descriptor ϕ_s and the one-hot ground truth scene category descriptor for 3D cuboid detection of the toilet and bookshelf categories, the categories had the most and least improvement of using the ground truth scene category ($+5.55\%$ and -0.10% respectively).

7 Conclusion

We extend our previous work to leverage positive context clues, by integrating scene category information with the appearance and geometry features. We have also provided a evaluation of our approach on the same SUN-RGBD dataset. The experimental results demonstrate that through effective integration of positive context clues such as scene category, our approach achieves a significant improvement over the previous work's performance.

References

[1] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[2] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In*Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[3] Geremy Heitz, Stephen Gould, Ashutosh Saxena, and Daphne Koller. Cascaded classification models: Combining models for holistic scene understanding. In *NIPS*, pages 641–648, 2009.

[4] Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59, 2009.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[6] Dahua Lin, Sanja Fidler, and Raquel Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1417–1424, 2013.

[7] Zhile Ren and Erik Sudderth. Three-dimensional object detection and layout prediction using clouds of oriented gradients. In *CVPR 2016*. IEEE, 2016.

[8] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 567–576, 2015.

[9] Shuran Song and Jianxiong Xiao. Sliding shapes for 3d object detection in depth images. In *Computer Vision–ECCV 2014*, pages 634–651. Springer, 2014.

[10] A. Vedaldi and A. Zisserman. Structured output regression for detection with partial occulsion. In *NIPS*, 2009.

[11] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010.

[12] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.