# Semantic Tree-Based 3D Scene Model Recognition

Juefei Yuan[1], Tianyang Wang[2], Shandian Zhe[3], Yijuan Lu[4], Bo Li[1*]
[1] School of Computing Sciences and Computer Engineering, University of Southern Mississippi, USA
[2] Department of Computer Science & Information Technology, Austin Peay State University, USA
[3] School of Computing, University of Utah, USA
[4] Department of Computer Science, Texas State University, USA

## Abstract

*3D scene recognition is important for many applications including robotics, autonomous driving cars, augmented reality (AR), virtual reality (VR), 3D movie and game production. A lot of semantic information (i.e. objects, object parts and object groups) is existing in 3D scene models. To significantly improve 3D scene recognition accuracy, we incorporate such semantic information into the recognition process by building a semantic scene tree and propose a deep random field (DRF) model-based semantic 3D scene recognition approach. Experiments demonstrate that the semantic approach can effectively capture semantic information of 3D scene models, accurately measure their similarities, and therefore greatly enhance the recognition performance. Code, data and experimental results can be found on the project homepage*[1].

## 1 Introduction

Scene understanding is one of the key questions in the community of computer vision. It may involve several components such as object detection, semantic segmentation, and scene recognition/classification. 3D scene recognition is to recognize the category of a given 3D scene which often involves multiple objects in a scene. It is important for a lot of related applications such as robotics [8, 4], autonomous driving cars [7], augmented reality (AR), virtual reality (VR), 3D movie and game production. Existing scene recognition/classification algorithms usually consider a 3D scene model as a common 3D object, and classify 3D scenes in the same way as 3D object classification. However, as a common sense, in 3D scene models there exists a lot of semantic information at different levels, such as object level, part level, and object groups level. Neglecting such important and helpful information during 3D scene classification or recognition will have significant negative impact on the performance.

Motivated by the above facts, to significantly improve the recognition accuracy, we propose a semantic-tree based scene recognition approach by first building a semantic scene tree based on the semantic ontology of WordNet [20] to host related semantic information, and then utilizing this tree to incorporate such semantic information during the scene recognition process. Experimental results demonstrate a significant improvement in recognition accuracy after utilizing such semantic information. They also prove that our semantic approach can effectively capture semantic information of 3D scene models, accurately measure their similarities, and therefore greatly enhance the recognition performance.

## 2 Related work

### 2.1 Deep learning based 3D scene understanding

According to Goodfellow et al. [13], the human visual system does much more than just recognizing objects. It is able to understand entire scenes including many objects and relationships between objects, and process rich 3D geometric information needed for our bodies to interface with the world.

YOLO (v1 [22], v2 [23], v3 [24]) is a state-of-the-art, end-to-end, one-stage, real-time object detection system. It can be used to detect objects either from videos or images. Compared with other object detection methods, YOLOv3 [24] has faster image processing speed and we adopt it in our object occurrence prediction in Section 4.2.2. Based on YOLOv1 [22] and YOLO9000 (v2) [23], in order to improve the performance, YOLOv3: (1) improves its backbone net structure (from v2's darknet-19 to v3's darknet-53, which has deeper layers); (2) changes the loss function (from softmax to logistic loss) to solve the problem

*Corresponding author. For any questions, please contact Bo Li. E-mail: bo.li@usm.edu or li.bo.ntu0@gmail.com.

[1]Project homepage URL: https://github.com/juefeiyuan/3D-scene-recognition/.

978-1-7281-4272-2/20/$31.00 ©2020 IEEE
DOI 10.1109/MIPR49039.2020.00024

of overlapping labels (e.g., woman and person); (3) adopts multi-scale predictions to solve the problem that small objects cannot be well detected.

Zhao et al. [29] proposed a framework to parse scene images at both pixel feature and word concept levels by jointly embedding the two levels of information into a high-dimensional vector space. At the word concept level, they incorporated the semantic word-word relations (hypernym/hyponym) based on WordNet [20] and compared their jointly embedding framework with other models, such as Word2Vec [19] and demonstrated better performance.

Choi et al. [12] proposed a hierarchical visual scene understanding model named 3D Geometric Phrase Model, which captures both semantic and geometric relationships of the objects in a scene, as well as their grouping information. Su et al. [26] devised a multi-view convolutional neural network (MVCNN) framework for 3D shape recognition by first learning features from multiple rendered views of a 3D model via a CNN model, and then fusing all the extracted features via a max-pooling like view pooling approach, and finally using another CNN as a classifier for the 3D shape recognition. We also utilized the MVCNN framework in our approach. PointNet [21] is a deep neural network designed on top of point clouds, and it directly consumes point clouds. Such an interaction better preserves the permutation invariance of points in the input, and thus mitigates the issues caused by transforming point cloud to regular 3D voxel grids or collections of images. The proposed unified architecture is applicable for a wide range of applications including object classification, part segmentation and scene parsing, and has demonstrated promising results, as well.

### 2.2 WordNet and its semantics-driven multimedia applications

WordNet [20] is a lexical database of concepts/synsets, represented by a set of synonyms. Each node in the tree represents one word, which has one or more senses (meanings). Each sense has its synset and a set of words are related through the following three relationships: hypernyms/hyponyms (IS_A relation), holonyms (MEMBER_OF relation) and meronyms (PART_OF relation). As a lexical dictionary of semantic concepts, WordNet has been widely applied in semantics-driven multimedia applications.

Marszalek and Schmid [18] proposed to utilize WordNet to build a semantic and hierarchical graph for the visual objects to be recognized. Based on labeled training data, they learned a binary classifier for each node in the graph. Wang et al. [27] proposed to build an ontology based on WordNet for a 3D model benchmark, infer 3D semantic properties by a rule engine based on Semantic Web Rule Language (SWRL), and perform semantic retrieval using the ontology. WordNet has been extensively used in visual understanding [1], at image (object)-level (i.e., ImageNet [25]), 3D model level (i.e., ShapeNet [10]), scene-level (i.e., Places [30]), and video (activity)-level (i.e., [14]). In addition, it is also used as a knowledge graph (like Freebase [5] for generic human knowledge, and GeneOntology [3] for biology). It is also useful for natural language understanding [6], and building its connection to visual understanding, such as Visual Genome [15].

## 3 Semantic tree-based 3D scene model recognition — a deep random field (DRF) model

### 3.1 Overview

In this paper, we propose a semantic tree-based 3D scene model recognition approach. We follow the recognition framework of multi-view convolutional neural network (MVCNN) [26] for 3D scene recognition, while we also incorporate semantics loss during the learning process and propose a deep random field (DRF) model. As illustrated in **Fig. 1**, our approach is composed of the following five steps.

(1) **2D Scene Semantic Tree construction**: build a Scene Semantic Tree (SST) for 3D scene models selected from the currently largest large-scale online 3D model repository 3D Warehouse [2], as described in Section 3.2. This semantic tree forms a network of semantic classes, attributes (i.e., semantic objects), and 3D scene model files.

(2) **3D scene view sampling**: since most (i.e., >90%) of the collected 3D scene models are in the upright position, starting from the front view, we sample 13 views for each 3D scene by first uniformly setting 12 cameras along the equator of the bounding sphere of the model and then raising their elevation angle by 20 degrees, together with a bird's-eye view generated by setting the camera on the north pole of the sphere.

(3) **Semantic object instance segmentation**: segment each scene view image into a set of consistent semantic objects. For example, as shown in **Fig. 1**, a view image of a 3D kitchen is segmented into the following semantic instances: several bottles, bowels, chairs, forks, tables, and wine glasses, together with a TV. These object categorical names, together with their number of appearances, form the *semantics* of the scene view.

(4) **Semantic loss computation**: compute the semantic similarity between the semantics of the unknown 3D scene model and that of each target scene category, based on the appearing scene objects' categorical names and their numbers of occurrence in the corresponding scene view images and the semantic information of the target category prelearned in Step (1).

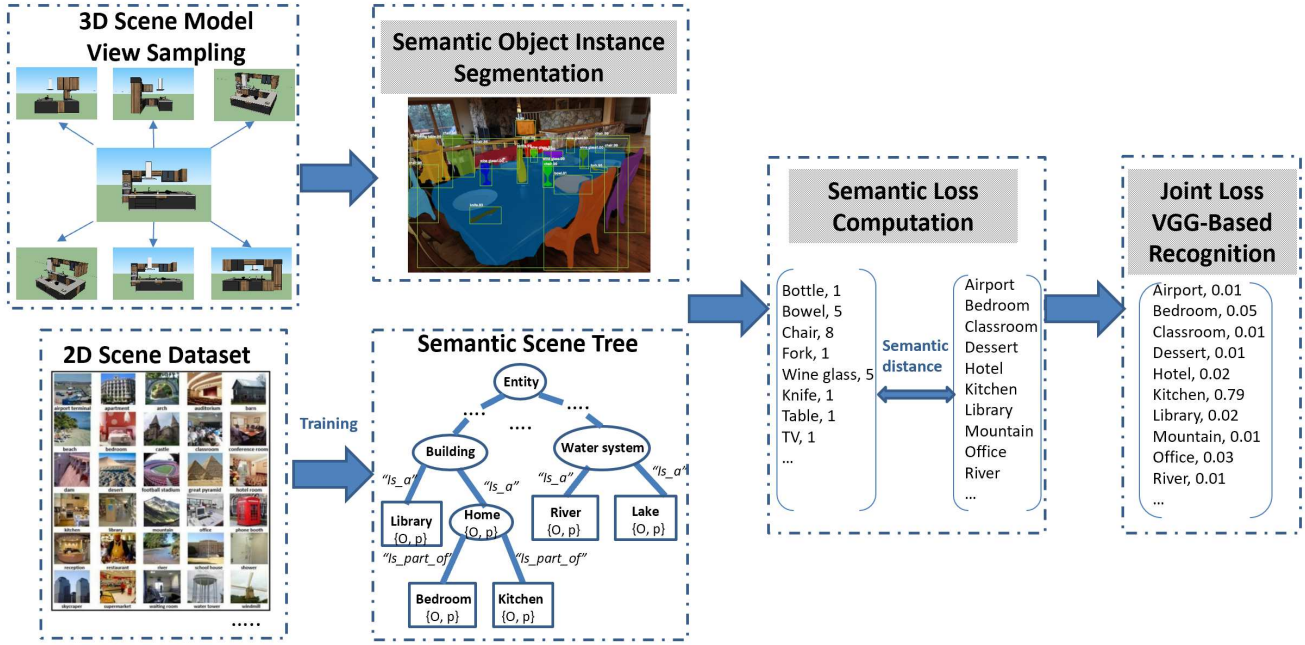



Figure 1: Semantic tree-based 3D scene model recognition approach: a deep random field (DRF) model. $\{O, p\}$ is used to represent the object occurrence distribution of each scene category: $O$ means an object class, while $p$ is its occurrence probability in that scene.

(5) **Joint loss VGG-based recognition (DRF)**: similar to MVCNN, we use the VGG16 framework to train our DRF model, but replace its loss function by a joint loss which combines VGG's cross-entropy loss and semantic-tree based loss, which will be detailed in Section 3.3. Finally, we utilize the trained DRF model to recognize each testing 3D scene model.

### 3.2 Scene Semantic Tree construction

**Semantic information extraction.** The YOLOv3 [24] model is adapted to help us detect all the possibly appearing objects in each 3D scene model. However, the original pre-trained YOLOv3 model can detect only 183 classes [17] existing in COCO stuff [9], a dataset for large-scale object detection, segmentation, and captioning, we often need to enlarge the training dataset to make it also contains manually-annotated object instances for other additional object classes. For example, in our experiments (Section 4.2.1, we add 27 additional classes whose names can be found on the project homepage. It is important and necessary since these object classes may have a high chance to appear in certain scenes. For example, desert is one of the 30 scene classes of the benchmark, while cactus objects are often present in desert scenes.

Assume the set of all possible object classes is $O$=$\{O_1, \ldots, O_n\}$. Based on YOLOv3, we detect the number of occurrence $c_i$ of each object class $O_i$ in a scene view generated from the training dataset, thus forming the object occurrence statistics $C$=$\{c_1, \ldots, c_n\}$. We then train a simple 9-layer DNN model based on the statistics to learn an object occurrence probability distribution to estimate the chance that each object will appear in a 3D scene, and this distribution is regarded as the scene semantics information of that scene: **Object occurrence probability** $\{P(O_i|S)\}$ indicating the conditional probability that an object class $O_i$ appears in a scene $S$. The number of nodes in each layer of the DNN model is: 500, 625, 500, 400, 600, 300, 200, 120, and 210, respectively.

**Scene Semantic Tree definition.** WordNet [20] provides a broad and deep taxonomy with over 80K distinct synsets representing distinct noun concepts arranged as a directed acyclic graph (DAG) network of hyponym relationships (e.g., "table" is a hyponym of "furniture"). As shown in **Fig. 1**, a Scene Semantic Tree (SST) is a hierarchy of classes with corresponding 3D scene models organized based on the semantic hierarchy in WordNet synsets. Each class (synset) of the Scene Semantic Tree has several attributes (i.e., via is-a, has-part, or is-made-of relations) according to its gloss defined in WordNet. Each leaf node of the Scene Semantic Tree has a number of 2D images belonging to the leaf node class. It also contains the scene semantics information (Object occurrence probability) learned in Section 3.2. Therefore, the Scene Semantic Tree forms a network of classes, attributes (i.e. scene object categori-

cal names and their estimated distribution), and related 3D scene model files.

### 3.3 Joint loss definition and DRF model training

The standard way to classify the objects in a scene or an image is to treat each object independently and train a deep neural network (DNN) to classify each object. To improve the recognition accuracy, we plan to incorporate the semantically relatedness relationships between the detected scene objects' labels and the candidate scene category labels into the training and prediction, as well, by utilizing the Scene Semantic Tree. For example, an object *table* detected from an unknown scene is more likely to help us classify the scene to be *kitchen* or *restaurant* rather than *phone booth* or *shower*, because *table* is a "PART_OF" *kitchen* or *restaurant* — they are more semantically related. We name our model as deep random field (DRF), because the way to encode the relationships resembles Markov random fields [11]. The loss function of our DRF model is,

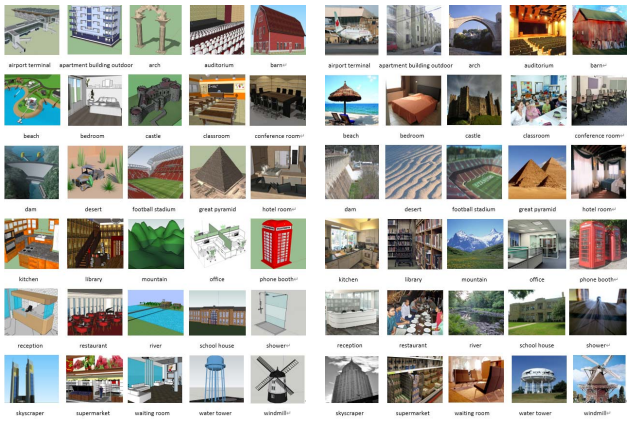
$$\mathcal{L} = \lambda \mathcal{L}_{\text{DNN}} + (1-\lambda)\mathcal{L}_{\text{SST}}(\{R_i * c_i\}, \{P(O_i|S)\}),$$

where, $\mathcal{L}_{\text{DNN}}$ and $\mathcal{L}_{\text{SST}}$ are the standard loss of a DNN classifier and the semantic loss based on the Scene Semantics Tree (SST), respectively, while both are defined based on the cross-entropy loss function (binary cross-entropy (BCE) for $\mathcal{L}_{\text{SST}}$); $\lambda$ is a hyper-parameter that represents the strength of the standard DNN part; $R_i$ is the WordNet-based semantic relatedness between two semantically related concepts: the object class name $O_i$ and a candidate scene category $S$ to classify the scene view. In our experiments, we adopt the Lesk [16] algorithm as the relatedness measurement; $c_i$ is the detected number of occurrences of $O_i$ in the scene view image; $\{P(O_i|S)\}$ is the scene semantics information of $S$ learned in Section 3.2. The learning will be optimizing the loss function to jointly estimate the weights of DNN. Before loss combination, we scale both DNN and semantic losses to be in the range of [0, 1].

## 4 Experiments and discussions

### 4.1 Dataset

We conduct a comprehensive evaluation of our semantic scene recognition algorithm based on the latest sketch/image-based 3D scene retrieval benchmark built by us, named **Scene_SBR_IBR** [28]. **Scene_SBR_IBR** was also used by us in organizing two 2019 Eurographics Shape Retrieval Contest (SHREC'19) tracks on 3D scene shape retrieval. It contains three subsets: 750 2D scene sketches, 30,000 2D scene images, and 3,000 3D scene models. All the 2D sketches/images and 3D scene models are equally classified into 30 classes. For each class, 18 sketches, 700 images and 70 models were randomly chosen for training while the rest 7 sketches, 300 images and 30 models were kept for testing. We utilize its 3D scene subset (testing dataset portion) to evaluate our 3D scene recognition algorithm, while using its image subset (training dataset portion) for scene semantic information extraction (See Section 3.2 for details and results in Section 4.2), considering its much larger size than that of the 3D scene dataset, much higher overall accuracy in scene details, and much more realistic scene features. A 3D scene example and a 2D scene image instance for each class are demonstrated in **Fig. 2** (a) and (b), respectively.




(a) models (b) images

Figure 2: 3D target scene model and 2D scene image examples in our **Scene_SBR_IBR** benchmark. One example per class is shown.

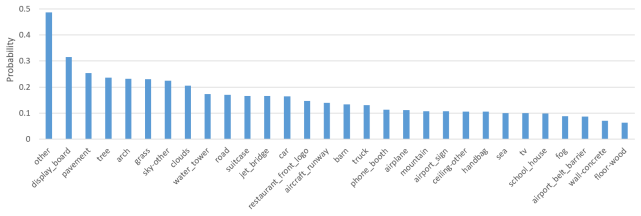
### 4.2 Semantics learning results

#### 4.2.1 Scene object categories

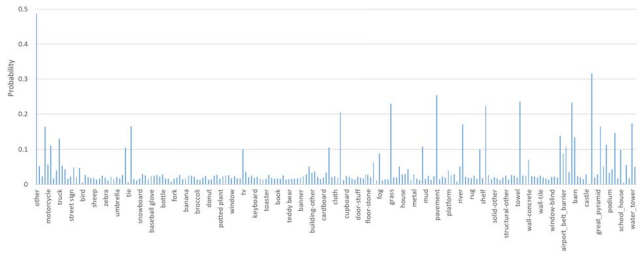
To learn the scene semantics information for the target 3D scenes in the **Scene_SBR_IBR** benchmark, we choose the maximum number of possible different object categories that may appear in the 3D scenes to be 210 by adding 27 additional classes, together with their manually annotated object instances to meet the needs of the **Scene_SBR_IBR** benchmark. The list of the 27 additional classes can be found on our project homepage.

#### 4.2.2 Object occurrence probabilities

By following the approach presented in Section 3.2, for each scene category, we first adopt YOLOv3 [24] framework to detect the objects in each scene image within the category to form the image's scene object statistics, and then individually employ a 9-layer deep neural network to



(a) top 30 classes



(b) all the 210 classes

Figure 3: Object occurrence probabilities for the airport terminal scene category.

train on all the obtained object statistics of the scene images to build the object occurrence probability for that scene category. **Fig. 3** shows an example result on the airport terminal scene class. Similarly, all the 30 scene categories' object occurrence probability distributions are available on the project homepage.

### 4.3 3D scene recognition results

We evaluate our DRF approach based on the testing dataset of the 3D scene subset of the **Scene_SBR_IBR** benchmark, and compare with the adapted MVCNN [26] approach (i.e., using the Places365 [30] pretrained model for the VGG part) for 3D scene recognition, which was named scene-based MVCNN (sMVCNN) by us. As described in the Section 3, DRF shares with MVCNN in terms of the recognition framework but incorporates the additional semantic-tree based loss into the loss function definition. Since we are dealing with scene models, rather than single object models like MVCNN, we adopt the scene image recognition model Places365 which is also based on VGG.

Firstly, we respectively train sMVCNN and DRF based on the training dataset (sampled scene images) of the target 3D scene dataset **Scene_SBR_IBR**, by starting with the Places365 pretrained model [30] or a randomly initialized Places365 model. We search the best $\lambda$ values based on a coarse-to-fine grid search with a search step of 0.1 and 0.01, respectively. The best $\lambda$ values are 0.67 and 0.57 for DRF started with a pre-trained and randomly initialized Places365 model, respectively. Secondly, we test the trained sMVCNN and our DRF model with the corresponding testing dataset based on their scene images as well. **Table 1** compares their recognition accuracies.

Table 1: Scene recognition accuracy comparison on the testing dataset of **Scene_SBR_IBR**.

| Accuracy | Pre-trained | Randomly initialized |
|---|---|---|
| sMVCNN [26] | 0.529 | 0.537 |
| DRF (Our) | 0.594 | 0.585 |

We can find that on the **Scene_SBR_IBR** dataset, for either way of model initialization, after incorporating the scene semantic information, our DRF has achieved an improvement of 12.3% and 8.94% in the accuracy, respectively, if compared to sMVCNN which does not consider the available scene semantic information. This demonstrates that our semantic-tree based approach has successfully captured the scene semantic information existing in the 3D scenes, and also accurately measured their similarities, thus significantly improved the 3D scene recognition performance.

## 5 Conclusions and future work

Our work aims to address the challenges in 3D scene recognition. We develop a probabilistic deep learning algorithm which incorporates the semantic relationships of the objects into the scene semantics learning process. The semantic information contains objects' occurrence information. Experiments demonstrate that the semantic approach can effectively capture semantic information of 3D scene models, accurately measure their similarities, and therefore greatly enhance the recognition performance.

We plan to expand the definition of semantic information to also include the following two additional pieces of semantic information: objects co-occurrence and spatial relations. (1) **Object co-occurrence probability** $\{P(O_i, O_j)|S)\}$: the conditional probability that both of two object classes $O_i$ and $O_j$ appear simultaneously in a 3D scene $S$; (2) **Spatial relation probability** $\{P(SR(O_i, O_j)|S)\}$: the conditional probability that two object classes $O_i$ and $O_j$ have a certain spatial relation (SR, a spatial preposition) in $S$ , e.g., $SR(O_i, O_j)$ = support / surround / near, that is, $O_i$ supports / surrounds / is near to $O_j$.

## References


[1] Deep learning for visual understanding: A review. *Neurocomputing*, 187:27 – 48, 2016. 2

[2] 3D Warehouse. http://3dwarehouse.sketchup.com/?hl=en, 2018. 2

[3] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000. 2

[4] J. Bellingham, P. E. Dupont, P. Fischer, L. Floridi, R. J. Full, N. Jacobstein, V. Kumar, M. McNutt, R. D. Merrifield, B. J. Nelson, B. Scassellati, M. Taddeo, R. Taylor, M. M. Veloso, Z. L. Wang, and R. J. Wood. The grand challenges of *Science Robotics*. *Science Robotics*, 3(14), 2018. 1

[5] K. D. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, pages 1247–1250, 2008. 2

[6] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642, 2015. 2

[7] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser. Simultaneous localization and mapping: A survey of current trends in autonomous driving. *IEEE Trans. Intelligent Vehicles*, 2(3):194–220, 2017. 1

[8] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. D. Reid, and J. J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Trans. Robotics*, 32(6):1309–1332, 2016. 1

[9] H. Caesar, J. R. R. Uijlings, and V. Ferrari. COCO-Stuff: Thing and stuff classes in context. *CoRR*, abs/1612.03716, 2016. 3

[10] A. X. Chang, T. A. Funkhouser, L. J. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An information-rich 3D model repository. *CoRR*, abs/1512.03012, 2015. 2

[11] R. Chellappa and A. Jain. Markov random fields: Theory and application. *Boston: Academic Press, 1993, edited by Chellappa, Rama; Jain, Anil*, 1993. 4

[12] W. Choi, Y. C. andc Caroline Pantofaru, and S. Savarese. Understanding indoor scenes using 3D geometric phrases. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 33–40, 2013. 2

[13] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. 1

[14] A. G. Huth, S. Nishimoto, A. T. Vu, and J. L. Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210 – 1224, 2012. 2

[15] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 2

[16] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC 1986, Toronto, Ontario, Canada, 1986*, pages 24–26, 1986. 4

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014. 3

[18] M. Marszalek and C. Schmid. Semantic hierarchies for visual object recognition. In *CVPR*, 2007. 2

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. 2

[20] G. A. Miller. WordNet: A lexical database for English. *Commun. ACM*, 38(11):39–41, 1995. 1, 2, 3

[21] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 77–85, 2017. 2

[22] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015. 1

[23] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6517–6525. IEEE Computer Society, 2017. 1

[24] J. Redmon and A. Farhadi. YOLOv3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. 1, 3, 4

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 2

[26] H. Su, S. Maji, E. Kalogerakis, and E. G. Learned-Miller. Multi-view convolutional neural networks for 3D shape recognition. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 945–953, 2015. 2, 5

[27] X. Wang, T. Lv, S. Wang, and Z. Wang. An Ontology and SWRL based 3D model retrieval system. In *AIRS*, pages 335–344, 2008. 2

[28] J. Yuan, H. Abdul-Rashid, B. Li, and Y. Lu. Sketch/image-based 3D scene retrieval: Benchmark, algorithm, evaluation. In *2nd IEEE Conference on Multimedia Information Processing and Retrieval, MIPR 2019, San Jose, CA, USA, March 28-30, 2019*, pages 264–269, 2019. 4

[29] H. Zhao, X. Puig, B. Zhou, S. Fidler, and A. Torralba. Open vocabulary scene parsing. In *ICCV*, pages 2021–2029. IEEE Computer Society, 2017. 2

[30] B. Zhou, À. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, 2018. 2, 5