Sketch/Image-Based 3D Scene Retrieval: Benchmark, Algorithm, Evaluation

Juefei Yuan[1], Hameed Abdul-Rashid[1], Bo Li[1,*], Yijuan Lu[2]
[1] School of Computing Sciences and Computer Engineering, University of Southern Mississippi, USA
[2] Department of Computer Science, Texas State University, San Marcos, USA

Abstract

Sketch/Image-based 3D scene retrieval is to retrieve man-made 3D scene models given a user's hand-drawn 2D scene sketch or a 2D scene image usually captured by a camera. It is a brand new but also very challenging research topic in the field of 3D object retrieval due to the semantic gap in their representations: 3D scene models or views differ from either non-realistic 2D scene sketches or realistic 2D scene images. Due to the intuitiveness in sketching and ubiquitous availability in image capturing, this research topic has vast applications such as 3D scene reconstruction, autonomous driving cars, 3D geometry video retrieval, and 3D AR/VR entertainment. To boost this interesting and important research, we build the currently largest and most comprehensive 2D scene sketch/image-based 3D scene retrieval benchmark[1], develop a convolutional neural network (CNN)-based 3D scene retrieval algorithm and finally conduct an evaluation on the benchmark.

1. Introduction

The huge and fast growing amount of 2D/3D scene files (sketches, images, and models) produced in our daily life makes researchers very excited since they shed bright light on many promising applications including autonomous driving cars, 3D scene reconstruction/retrieval, 3D geometry video retrieval, virtual reality (VR) and augmented reality (AR) in 3D entertainment, 3D game, animation and movie, 3D printing, and mobile applications. For instance, 2D sketch-based 3D scene retrieval and reconstruction facilitate the ability to create 3D scene contents for many 4D immersive programs, such as one of the most popular Disney World's programs - Avatar Flight of Passage Ride [14, 4, 12], and help us to find domain-specific indoor/outdoor 3D scenes, such as battlefield scenes for soldiers in training, sand table models for real estate marketing, or desert and mountain scenes for cartoon or game production. Therefore, the benefits of this 3D scene retrieval research direction will be in its applications as well as across many related fields.

Sketch-based 3D scene retrieval is searching for relevant 3D scenes based on a 2D scene sketch input, which provides an intuitive and convenient scheme for users to learn and retrieve 3D scenes. However, existing sketch-based 3D model retrieval algorithms [8] have mainly focused on single object retrieval since they assume that there is only one object in a query sketch. Typically, there are several objects within a scene that may overlap one another, also occlusions are very common, and location configuration between objects provides helpful contextual information but is not trivial to interpret during retrieval. Nevertheless, considering its vast application scenarios, we believe that this research topic deserves further exploration and hopes to raise more interest and attention from people both inside and outside of the 3D object retrieval research community.

Image-based 3D scene retrieval finds relevant 3D scenes based on the content in a 2D scene query image, which also has vast related applications, including highly capable autonomous vehicles like the Renault SYMBIOZ [9] [13], multi-view 3D scene reconstruction, VR/AR scene content generation, and consumer electronics apps. However, similarly, there is a lack [19] of substantial research in this field due the challenges involved and lack of related retrieval benchmarks. Seeing the benefits of advances in retrieving 3D scene models based on a scene image query makes this research direction useful, promising, and interesting as well.

Although sketch/image-based 3D scene retrieval is an important and interesting research topic, it is extremely challenging due to the semantic gap in their representations: 3D scene models differ from either non-realistic 2D scene sketches or realistic 2D scene images. Due to the wider representation gap between roughly outlined 2D scene sketches and precise 3D scene models, 2D scene sketch-based 3D

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

[1]Our **Scene_SBR_IBR** benchmark: `http://orca.st.usm.edu/~bli/Scene_SBR_IBR/`.

scene retrieval (SceneSBR) is one of the most challenging research topics in 3D object retrieval.

To promote this interesting yet challenging research, we organized two 2018 Eurographics Shape Retrieval Contest (SHREC) tracks [21, 2, 22, 3] titled "2D Scene Sketch-Based 3D Scene Retrieval" and "2D Scene Image-Based 3D Scene Retrieval", by building the first 2D scene sketch/image-based 3D scene retrieval benchmark **SceneSBR** and **SceneIBR**. They share the same 1,000 3D scene models as the target dataset. **SceneSBR** contains 250 scene sketches, while **SceneIBR** has 10,000 2D scene images. All the sketches, images and models are equally classified into 10 indoor as well as outdoor classes.

However, as can be seen, both benchmarks contain only 10 scene classes, and this is one of the reasons that all the three deep learning-based participating methods have achieved excellent performance. Considering this, after the track we have tripled the size of **SceneSBR** and **SceneIBR** to make each has 30 classes, and built an extended version for each, resulting a unified benchmark **Scene_SBR_IBR**, which has 750 2D scene sketches, 30,000 scene images, and 3,000 3D scene models. Similarly, all the 2D sketches, 2D images and 3D scenes are equally classified into 30 classes.

In addition, to evaluate the new benchmark, we also propose a classification-based 3D scene retrieval algorithm based on the VGG [10] deep learning model, considering the existing semantic gap in their representations. We develop our own 3D scene view sampling method; then simultaneously classify a query sketch/image and all the target 3D scenes via their scene views; and finally rank the 3D scenes accordingly to generate the retrieval algorithm's performance, which is provided as a baseline accuracy for the new benchmark. The main contributions introduced in this work are highlighted as follows:

- We propose a brand new research direction, that is sketch/image-based 3D scene retrieval, in the 3D Object Retrieval research area.
- We build a new and the currently largest sketch/image-based 3D scene retrieval benchmark to promote this research direction.
- We evaluate the benchmark by using our newly proposed deep learning classification-based 3D scene retrieval algorithm and provide the baseline performance for the community.

2. Related work

3D scene retrieval. Compared to sketch-based retrieval in the context of a single sketch, sketch-based retrieval using a 2D scene sketch query is much less studied. Fisher and Hanrahan [7] proposed a novel 3D model retrieval scheme named context-based 3D model retrieval, which is to retrieve models according to their spatial context in a 3D scene. They adopted a new pipeline of model retrieval by first locating the position of the model by drawing a 3D box and then searching relevant 3D models based on the dimensionality and context information. Models within scenes are extracted beforehand. Both geometry and tags are utilized to find related models based on the assumption that those models appear with similar contexts. However, our research topic differs from theirs in several facets. Firstly, their input is already a 3D scene consisting of several models. Secondly, the retrieval scheme is mainly for scene completion rather than scene reconstruction from scratch. Finally, they do not draw sketches to represent a 3D model or use an image as a template to build the scene. Recently, Xu et. al [18] proposed Sketch2Scene for automatic 2D sketch-based 3D scene composition by representing 3D scene objects' functional and spatial relationships based on structural groups.

2D/3D scene datasets. Xiao et. al [15] built the Scene UNderstanding (SUN) image database for the purpose of fostering improvements in large scale scene recognition. SUN was initially comprised of 899 scene categories and 130,519 images. Later, SUN was extended to include 908 distinct classes [17]. Xiao et. al [16] further created SUN3D, a RGB-D video database with camera pose and object labels, to capture full extend of 3D places. They used the videos for partial 3D reconstruction and propagated labels from one frame to another, then used the labels to refine the partial reconstruction. Song et. al [11] constructed SUNCG, a database of synthetic 3D scenes with manually labeled voxel occupancy and semantic labels. SUNCG has 45,622 different scenes and 2,644 objects across 84 categories. Zhou et. al [23] compiled Places, a database of 10,624,928 scene images across 434 scene categories. While Places is not annotated at the object level, it provides the most diverse scene composition as well as insights into solutions to scene understanding problems.

3. Our Scene_SBR_IBR benchmark

Overview. We have built a unified 3D scene benchmark supporting both sketch and model queries by substantially extending **SceneSBR** and **SceneIBR** by means of identifying and consolidating the same number of sketches/images/models for another additional 20 classes from the most popular 2D/3D data resources. Our work is the first to form a new and larger benchmark corpus for both sketch-based and image-based 3D scene retrieval. This benchmark provides an important resource for the community of 3D scene retrieval and will likely foster the development of practical sketch-based and image-based 3D scene retrieval applications.

Motivation. As mentioned in Section 1, to foster the research direction of sketch-based and image-based 3D scene retrieval, we built the first benchmark **SceneSBR** and **Scene_IBR** respectively and organized two related Shape Retrieval Contest (SHREC) tracks [21, 2]. During the competitions, we found that both of these two benchmarks are not challenging and comprehensive enough since they cover only 10 categories, each of which is clearly distinct from one another. Considering this, we decided to further increase the comprehensiveness of the benchmarks by building a significantly larger and unified benchmark which supports both types of retrieval.

Building process. The first thing for the benchmark design is category selection, for which we have referred to several of the most popular 2D/3D scene datasets, such as Places [23] and SUN [15]. Finally, we selected the most popular 30 scene classes (including the initial 10 classes in **SceneSBR** and **Scene_IBR**) from the 88 available category labels in the Places88 dataset [23], via a voting mechanism among three people (two graduate students as voters and a faculty member as the moderator) based on their judgments. We want to mention that the 88 common scenes are already shared by ImageNet [5], SUN [15], and Places [23]. Then, to collect data (sketches, images, and models) for the additional 20 classes, we gathered from Flicker and Google Image for sketches and images, and downloaded SketchUp 3D scene models (originally in .SKP format, but we provide .OBJ format as well after transformation) from 3D Warehouse [1].

Benchmark details. Our 3D scene retrieval benchmark **Scene_SBR_IBR** is publicly available[2]. All of its 30 classes has the same number of 2D scene sketches (25), 2D scene images (1000), and 3D scene models (100). It supports both sketch-based (via **Scene_SBR**) and image-based (via **Scene_IBR**) 3D scene retrieval by providing two different query datasets and a common 3D scene model target dataset:

- **2D Scene Sketch Query Dataset (Subset 1).** The 2D scene sketch query set comprises 750 2D scene sketches (30 classes, each with 25 sketches). One example per class is demonstrated in **Fig. 1**.
- **2D Scene Image Query Dataset (Subset 2).** The 2D scene image query set is composed of 30,000 scene images (30 classes, each with 1,000 images). One example per class is demonstrated in **Fig. 2**.
- **3D Scene Model Target Dataset (Subset 3).** The 3D scene dataset is built on the selected 3,000 3D scene models downloaded from Google 3D Warehouse. Each class has 100 3D scene models. One example per class is shown in **Fig. 3**.

[2]Available on the URL: `http://orca.st.usm.edu/~bli/Scene_SBR_IBR/`

Therefore, **Scene_SBR_IBR** is divided into two sub-level and task-specific benchmarks: **Scene_SBR** comprising **Subsets** 1 and 3 for sketch-based 3D scene retrieval and **Scene_IBR** comprising **Subsets** 2 and 3 for image-based 3D scene retrieval.

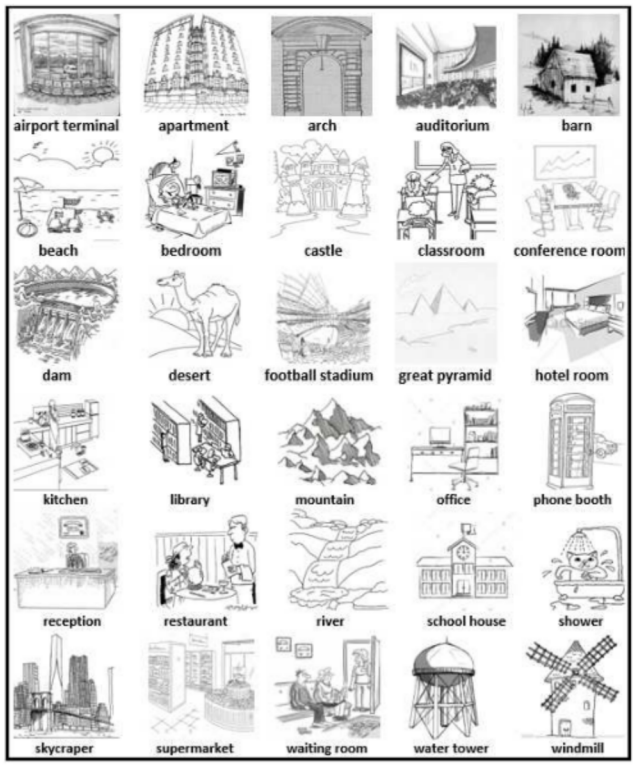

Figure 1. Example 2D scene query sketches in our Scene_SBR_IBR benchmark. One example per class is shown.

To help evaluate learning-based 3D scene retrieval algorithms, we randomly select 18 sketches, 700 images, and 70 models from each class for training and use the remaining 7 sketches, 300 images, and 30 models for testing, as indicated in **Table** 1. The users are required to generate results based on the testing dataset if they use learning in their approach(es). Otherwise, the retrieval results can be generated based on the complete dataset.

Evaluation metrics. To have a comprehensive evaluation of a retrieval algorithm on our benchmark, we employ seven commonly adopted performance metrics in the 3D model retrieval community [8]: Precision-Recall (PR) diagram,

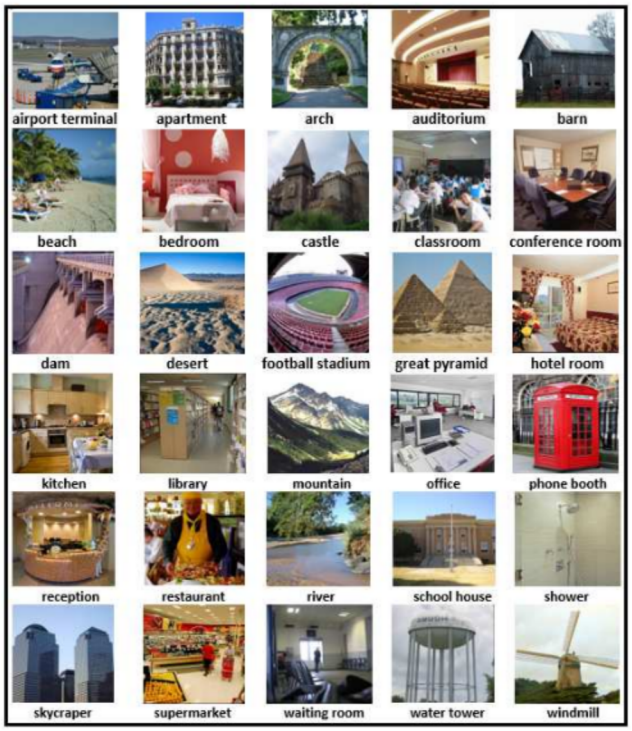

Figure 2. Example 2D scene query images in our Scene_SBR_IBR benchmark. One example per class is shown.

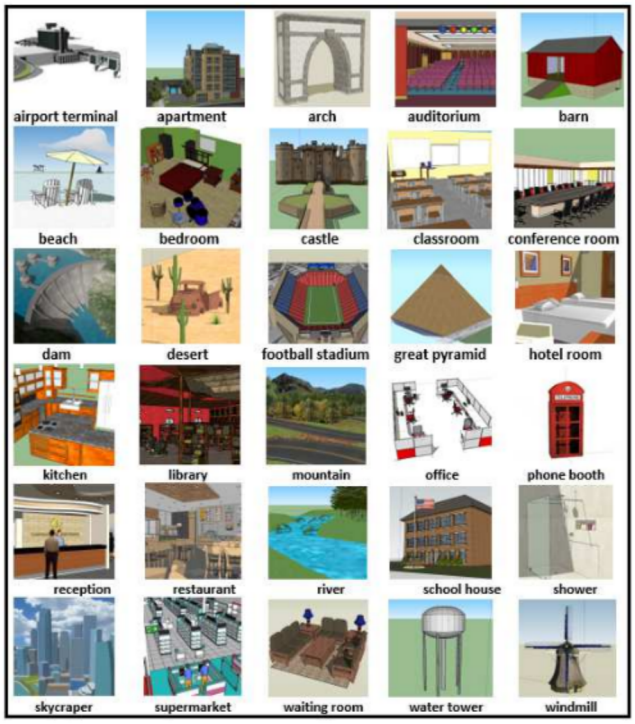

Figure 3. Example 3D target scenes in our Scene_SBR_IBR benchmark. One example per class is shown.

Nearest Neighbor (NN), First Tier (FT), Second Tier (ST), E-Measures (E), Discounted Cumulated Gain (DCG) and Average Precision (AP). We also have developed the related code to compute them for our benchmark.

4. Our retrieval algorithm VMV-VGG

Based on the VGG-16 deep learning model [10] and our prior work [20], we propose a View and Majority Vote based 3D scene retrieval algorithm (**VMV-VGG**), as illustrated in **Fig.** 4. It employs two different VGG-16 based classification models (VGG1 and VGG2): one for 2D scene sketches/images, and the other for 2D scene views. For the model, we have the following parameter settings: the input image size: 256x256; the number of output categories: 30 for our proposed benchmark; the batch size: 64; and the learning rate: 0.00001.

Table 1. Training and testing dataset information of our Scene_SBR_IBR benchmark.

Datasets	Sketches	Images	Models
Training (per class)	18	700	70
Testing (per class)	7	300	30
Total (per class)	25	1,000	100
Total (all 30 classes)	750	30,000	3,000

The retrieval algorithm comprises the following six steps: (1) Scene view sampling: we center each 3D scene model in a 3D sphere, and then automatically sample 13 scene view images based on a QMacro script program developed by us which automates the operations of the SketchUp software to perform the view sampling. For the 13 viewpoints, we uniformly sample 12 views along the equator of the sphere starting from the front view, and then sample one view from the north pole of the sphere. One example is demonstrated in **Fig.** 5. (2) **Data augmentation:** to counter overfitting, before each pre-training or training, we first employ the same data augmentation technique we developed in [20] to increase the related dataset's size by 500 times based on a set of random rotations, shifts and flips. (3) **Pre-training of VGG1 and VGG2:** for sketch-based 3D scene retrieval we pre-train the VGG1 model on the TU-Berlin sketch dataset [6] at epoch 500, and pre-train

VGG2 on the Places scene image dataset [23] at epoch 100; while for the image-based retrieval mode, we use Places to pre-train both VGG1 and VGG2 and stop at epoch 100. (4) **Fine-tuning:** fine-tune the pre-trained VGG1/VGG2 on the corresponding 2D scene sketches/images or 2D scene views training dataset at epoch 100/50, respectively. (5) **Sketch/image/view classification:** we respectively feed the well-trained model VGG1 or VGG2 with its corresponding testing query sketch/image or target scene view to obtain two classification vectors. (6) **Majority vote-based label matching:** finally we generate the rank list for each query by using majority vote-based label matching method based on the query's classification vector and a target 3D scene's 13 classification vectors.

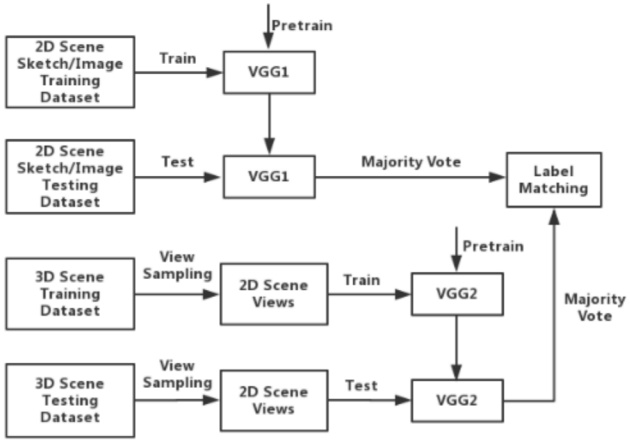

Figure 4. VBV-VGG architecture.

Figure 5. Scene view sampling example: a set of 13 sample views of an apartment scene model.

5. Evaluation

The purpose of this evaluation is to provide the baseline performance for either sketch-based or image-based 3D scene retrieval on our benchmark **Scene_SBR_IBR** and also to examine the benchmark's comprehensiveness and difficulty level. To reach this purpose, we run our **VMV-VGG** algorithm presented in Section 4 on the two sub-level benchmarks **Scene_SBR** and **Scene_IBR** of **Scene_SBR_IBR** respectively, by following the same parameter settings mentioned in Section 4.

Their performance is shown in **Fig.** 6 and **Table** 2 of **Scene_SBR_IBR** based on the seven performance evaluation metrics mentioned in Section 3. We have found that compared with the performance that has been achieved in the two SHREC'18 tracks [21, 2] which used a much smaller benchmark containing only 10 classes, in contrast to the current 30 classes available in our new benchmark, the overall performance dropped significantly for either type of retrieval. For example, Li's MMD-VGG method, which also utilizes VGG, has achieved an excellent overall performance in terms of DCG (0.856) or AP (0.685) on the **SceneSBR** benchmark, while they drop to DCG (0.533) and AP (0.244) respectively based on our **VMV-VGG**. This should be a direct and natural result after a substantial increase in the comprehensiveness and challenge level that exist in **Scene_SBR_IBR** after we incorporate much more scene categories. The addition of more classes will cause more ambiguities during the retrieval process and the retrieval algorithm may fail to properly distinguish between classes that share certain similarities.

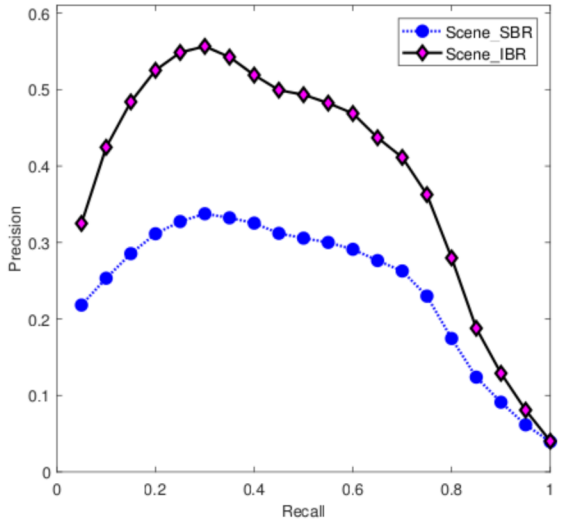

Figure 6. Precision-Recall diagram performance of our VMV-VGG on our Scene_SBR_IBR benchmark.

Table 2. Performance metrics generated by running our VMV-VGG on our Scene_SBR_IBR benchmark.

Benchmark	NN	FT	ST	E	DCG	AP
Scene_SBR	**0.081**	**0.281**	**0.369**	**0.280**	**0.533**	**0.244**
Scene_IBR	**0.122**	**0.458**	**0.573**	**0.452**	**0.644**	**0.392**

6. Conclusions and future work

Sketch/Image-based 3D scene retrieval are brand new, interesting, and challenging research topics with a lot of application potentials. There is extremely limited preliminary work in this field, which allows us to explore many promising ideas and interesting results. In this paper, the currently largest 3D scene retrieval benchmark **Scene_SBR_IBR** is proposed with the hope to advance this research direction. To assist other interested researchers, the baseline performance on the benchmark has been provided by conducting evaluation based on a proposed CNN classifier-based 3D scene retrieval algorithm **VMV-VGG**. Our future goals include: (1) building a large-scale and/or multimodal 2D scene sketch/image-based 3D scene retrieval benchmark; (2) semantics-driven 2D scene sketch/image-based 3D scene retrieval.

Acknowledgments

This project is supported by the University of Southern Mississippi Faculty Startup Funds Award to Dr. Bo Li and the Texas State Research Enhancement Program and NSF CRI-1305302 Awards to Dr. Yijuan Lu. We gratefully acknowledge the support from NVIDIA Corporation for the donation of the Titan X/Xp GPUs used in this research.

References

[1] 3D Warehouse. http://3dwarehouse.sketchup.com/?hl=en, 2018.

[2] H. Abdul-Rashid and et al. SHREC'18 track: 2D scene image-based 3D scene retrieval. In *3DOR*, pages 1–8, 2018.

[3] H. Abdul-Rashid, J. Yuan, B. Li, and Y. Lu. SHREC'18 2D Scene Image-Based 3D Scene Retrieval Track Website. http://orca.st.usm.edu/~bli/SceneIBR2018/, 2018.

[4] W. Attractions. New ride!!!! Disney world animal kingdom: Avatar flight of passage ride video 4k hd video (pov). http://www.youtube.com/watch?v=f-cw7iCUY3c, 2018.

[5] J. Deng and et al. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[6] M. Eitz and et al. How do humans sketch objects? *ACM Trans. Graph.*, 31(4):44:1–44:10, 2012.

[7] M. Fisher and P. Hanrahan. Context-based search for 3D models. *ACM Trans. Graph.*, 29:182:1–182:10, 2011.

[8] B. Li and et al. A comparison of 3D shape retrieval methods based on a large-scale benchmark supporting multimodal queries. *Computer Vision and Image Understanding*, 131:1–27, 2015.

[9] Renault. Renault SYMBOIZ Concept. http://www.renault.co.uk/vehicles/concept-cars/symbioz-concept.html.

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[11] S. Song and et al. Semantic scene completion from a single depth image. In *CVPR*, pages 190–198. IEEE Computer Society, 2017.

[12] I. the Magic. New flight of passage ride queue, preshow in pandora - the world of avatar at walt Disney world. http://www.youtube.com/watch?v=eM8f47Igtu8, 2018.

[13] L. T. Tips. Driving a multi-million dollar autonomous car. http://www.youtube.com/watch?v=vlIJfV1u2hM\&feature=youtu.be.

[14] Wikipedia. Avatar flight of passage. http://en.wikipedia.org/wiki/Avatar_Flight_of_Passage, 2018.

[15] J. Xiao and et al. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492. IEEE Computer Society, 2010.

[16] J. Xiao and et al. SUN3D: A database of big spaces reconstructed using SfM and object labels. In *ICCV*, pages 1625–1632, 2013.

[17] J. Xiao and et al. SUN database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119(1):3–22, 2016.

[18] K. Xu and et al. Sketch2Scene: sketch-based co-retrieval and co-placement of 3D models. *ACM Trans. Graph.*, 32(4):123, 2013.

[19] K. Xu, V. G. Kim, Q. Huang, N. Mitra, and E. Kalogerakis. Data-driven shape analysis and processing. In *SIGGRAPH ASIA 2016 Courses*, page 4. ACM, 2016.

[20] Y. Ye and et al. 3D sketch-based 3D model retrieval with convolutional neural network. In *23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016*, pages 2936–2941. IEEE, 2016.

[21] J. Yuan and et al. SHREC'18 track: 2D scene sketch-based 3D scene retrieval. In *3DOR*, pages 1–8, 2018.

[22] J. Yuan, B. Li, and Y. Lu. SHREC'18 2D Scene Sketch-Based 3D Scene Retrieval Track Website. http://orca.st.usm.edu/~bli/SceneSBR2018/, 2018.

[23] B. Zhou and et al. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, 2018.