**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 promote this important research and engage fellow researchers, we built the current largest and most comprehensive 2D scene sketch/image-based 3D scene retrieval benchmark, developed a convolutional neural network (CNN)-based 3D scene retrieval algorithm and finally conducted an evaluation on the benchmark.

**Introduction**

The exponential growth of 2D/3D scene files (sketches, images, and models) produced in our daily lives instills excitement among researchers for many promising applications such as 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 assist us 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]. They also help us 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 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 by assuming only one object exists 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 hope to raise more 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. In a similar manner, there is a lack [19] of substantial research in this field due to 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 among representations: 3D scene models differ from either non-realistic 2D scene sketches or realistic 2D scene images. Because roughly outlined 2D scene sketches and precise 3D scene models exists on opposite ends of the representational spectrum, 2D scene sketch-based 3D scene retrieval 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 respectively 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 the current 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. While, they mainly focus on indoor scenes, such as offices, kitchens and bedrooms. 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 poses 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 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 were not challenging and comprehensive enough since they cover only 10 distinctive categories. 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 state-of-the-art 2D/3D scene datasets, such as Places [23] and SUN [15]. The criteria for the category selection is popularity. Finally, we selected the most popular 30 scene classes (including the initial 10 classes in SceneSBR and SceneIBR) 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 . All 30 classes possess the same number of 2D scene sketches (25), 2D scene images (1000), and 3D scene models (100). It supports both sketch-based and image-based 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 3D Warehouse. Each class has 100 3D scene models. One example per class is shown in Fig. 3.

Therefore, Scene SBR IBR is divided into two sublevel 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. 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, 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.

**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. The code is also publicly available3 . 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.

**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 goal, 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 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 difficulty level that exist in Scene SBR IBR after we incorporate much more scene categories. The addition of more classes will led to further ambiguity during the retrieval process and the retrieval algorithm may fail to properly distinguish between classes that share certain similarities.

**Conclusion and Future Work**

Sketch/Image-based 3D scene retrieval are brand new, interesting, and challenging research topics with a lot of application potentials. Due to extremely limited preliminary work in this field, we are able to explore many ideas and formulate new results. In this paper, the current largest 3D scene retrieval benchmark Scene SBR IBR, which supports both sketch and image queries, 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) conducting research in semantics-driven 2D scene sketch/image-based 3D scene retrieval.