Eurographics Workshop on 3D Object Retrieval (2018)
A. Telea and T. Theoharis, R. Veltkamp (Editors)

SHREC'18 Track: 2D Scene Sketch-Based 3D Scene Retrieval

Juefei Yuan[†1], Bo Li*[†1], Yijuan Lu[†2], Song Bai[‡3], Xiang Bai[‡3], Ngoc-Minh Bui[‡4], Minh N. Do[‡5], Trong-Le Do[‡4], Anh-Duc Duong[‡6], Xinwei He[‡3], Tu-Khiem Le[‡4], Wenhui Li[‡7], Anan Liu[‡7], Xiaolong Liu[‡3], Khac-Tuan Nguyen[‡4], Vinh-Tiep Nguyen[‡6], Weizhi Nie[‡7], Van-Tu Ninh[‡4], Yuting Su[‡7], Vinh Ton-That[‡4], Minh-Triet Tran[‡4], Shu Xiang[‡7], Heyu Zhou[‡7], Yang Zhou[‡3], Zhichao Zhou[‡3]

[1] School of Computing, University of Southern Mississippi, USA
[2] Department of Computer Science, Texas State University, San Marcos, USA
[3] School of Electronic Information and Communications, Huazhong University of Science and Technology, China
[4] University of Science, Vietnam National University, Vietnam [5] University of Illinois at Urbana-Champaign, USA
[6] University of Information Technology, Vietnam National University, Vietnam
[7] School of Electrical and Information Engineering, Tianjin University, China

Abstract

Sketch-based 3D model retrieval has the intuitiveness advantage over other types of retrieval schemes. Currently, there is a lot of research in sketch-based 3D model retrieval, which usually targets the problem of retrieving a list of candidate 3D models using a single sketch as input. 2D scene sketch-based 3D scene retrieval is a brand new research topic in the field of 3D object retrieval. Unlike traditional sketch-based 3D model retrieval which ideally assumes that a query sketch contains only a single object, this is a new 3D model retrieval topic within the context of a 2D scene sketch which contains several objects that may overlap with each other and thus be occluded and also have relative location configurations. It is challenging due to the semantic gap existing between the iconic 2D representation of sketches and more accurate 3D representation of 3D models. But it also has vast applications such as 3D scene reconstruction, autonomous driving cars, 3D geometry video retrieval, and 3D AR/VR Entertainment. Therefore, this research topic deserves our further exploration.

To promote this interesting research, we organize this SHREC track and build the first 2D scene sketch-based 3D scene retrieval benchmark by collecting 3D scenes from Google 3D Warehouse and utilizing our previously proposed 2D scene sketch dataset Scene250. The objective of this track is to evaluate the performance of different 2D scene sketch-based 3D scene retrieval algorithms using a 2D sketch query dataset and a 3D Warehouse model dataset. The benchmark contains 250 scene sketches and 1000 3D scene models, and both are equally classified into 10 classes. In this track, six groups from five countries (China, Chile, USA, UK, and Vietnam) have registered for the track, while due to many challenges involved, only 3 groups have successfully submitted 8 runs. The retrieval performance of submitted results has been evaluated using 7 commonly used retrieval performance metrics. We also conduct a thorough analysis and discussion on those methods, and suggest several future research directions to tackle this research problem. We wish this publicly available [YLL18] benchmark, comparative evaluation results and corresponding evaluation code, will further enrich and advance the research of 2D scene sketch-based 3D scene retrieval and its applications.

Categories and Subject Descriptors (according to ACM CCS): H.3.3 [Computer Graphics]: Information Systems—Information Search and Retrieval

1. Introduction

2D scene sketch-based 3D scene retrieval is to retrieve relevant 3D scenes (in either .OBJ or .SKP format) using a 2D scene sketch as input. This scheme is intuitive and convenient for users to learn and search for 3D scenes. It is also very promising and has great potentials in many applications such as autonomous driving cars, 3D scene reconstruction, 3D geometry video retrieval, virtual reality (VR) and augmented reality (AR) in 3D Entertainment like Disney World's Avatar Flight of Passage Ride [Wik18, Att18, tM18].

However, although there are many existing 2D sketch-based 3D shape retrieval systems, there is little existing research work on 2D scene sketch-based 3D scene retrieval due to two major reasons 1) It is challenging to collect a large-scale 3D scene dataset and there

† Track organizers. *Corresponding author. For any question related to the track, please contact Bo Li. E-mail: bo.li@usm.edu.
‡ Track participants.

exists a very limited number of available 3D scene shape benchmarks. 2) Like 2D sketch-based 3D shape retrieval, there is a big semantic gap between the iconic representation of 2D scene sketches and the accurate 3D coordinate representations of 3D scenes. All of above reasons make the task of retrieving 3D scene models using 2D scene sketch queries a challenging, although interesting and promising, research direction.

Ye et al. [YLJ16] collected the Scene250 benchmark comprising 250 2D scene sketches of 10 classes, each with 25 sketches. It avoids the bias issue since they collected the same number of sketches for every class, while the sketch variation within one class is also adequate enough. However, there are no 3D scene dataset corresponding to this 2D scene sketch dataset.

Motivated by above obstacles, 100 3D scene models have been selected for each of the ten classes in Scene250 from 3D Warehouse [Tri18], an open source library which allows SketchUp users to upload 3D models to share and download needed 3D models. The SketchUp (.SKP) type of the online 3D scene models can be transformed into many other formats, such as OBJ, PLY, and OFF.

We organize this track to foster this challenging research direction by soliciting retrieval results from current state-of-the-art 3D scene retrieval methods for comparison, especially in terms of scalability to 2D scene sketch queries. We also provide corresponding evaluation code for computing a set of performance metrics similar to those used in the Query-by-Model retrieval technique.

2. SceneSBR Benchmark

2.1. Overview

Our 2D scene sketch-based 3D scene retrieval benchmark **SceneSBR** utilizes the 250 2D scene sketches in Scene250 [YLJ16] as its 2D scene sketch dataset and 1000 SketchUp 3D scene models (.OBJ and .SKP format) as its 3D scene dataset. Each of the ten classes has the same number of 2D scene sketches (25) and 3D scene models (100).

To facilitate learning-based retrieval, we randomly select 18 sketches and 70 models from each class for training and use the remaining 7 sketches and 30 models for testing, as indicated in **Table** 1. Participants are required to submit results on the testing dataset only if they use learning in their approach(es). Otherwise, the retrieval results based on the complete (250 sketches, 1000 models) dataset is needed.

Table 1: *Training and testing datasets (per class) of our **SceneSBR** benchmark.*

SceneSBR Benchmark	Sketch	Model
Training	18	70
Testing	7	30
Total (per class)	25	100

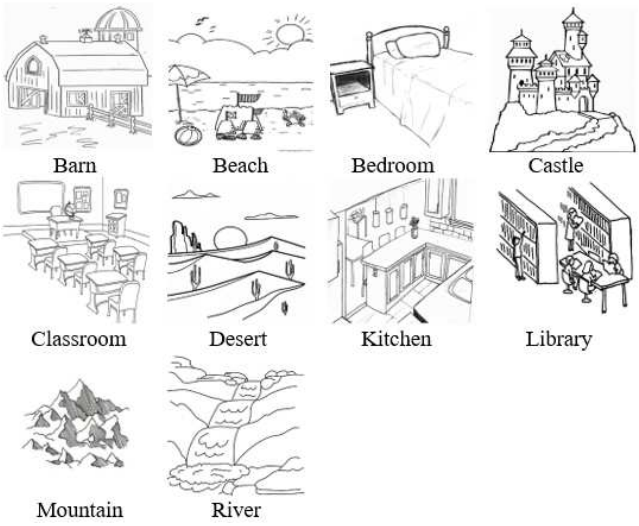


Figure 1: *Example 2D scene sketches (one example per class) [YLJ16] in our **SceneSBR** benchmark.*

2.2. 2D Scene Sketch Dataset

The 2D scene sketch query set comprises 250 2D scene sketches (10 classes, each with 25 sketches), while all the classes have relevant models in the target 3D scene dataset which are downloaded from the 3D Warehouse. One example per class is demonstrated in **Fig.** 1.

2.3. 3D Scene Dataset

The 3D scene dataset is built on the selected 1000 3D scene models downloaded from 3D Warehouse. Each class has 100 3D scene models. One example per class is shown in **Fig.** 2.

2.4. Evaluation Method

The objective of this track is to evaluate the performance of different 2D scene sketch-based 3D scene retrieval algorithms using a 2D sketch query dataset and a 3D Warehouse model dataset. While, each algorithm targets retrieving 3D scene models that belong to the same class as that of each query 2D scene sketch. To have a comprehensive evaluation of the retrieval algorithm, we employ seven commonly adopted performance metrics in the 3D model retrieval community [LLL*15, LLG*14]. They are 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 code to compute them.

3. Participants

There were six groups who registered for the track. Two groups come from China, and one group each comes from Chile, USA, UK, and Vietnam. Each group was given three weeks to complete

Figure 2: *Example 3D scene models (one example per class) in our **SceneSBR** benchmark.*

Figure 3: *Several example views of scene models.*

the contest. They were registered to submit both their results and methods description.

Three groups have finally participated in the SHREC'18 track on 2D Scene Sketch-Based 3D Scene Retrieval. Eight (8) rank list results (runs) for four (4) different methods developed by three (3) groups have been submitted. The participants and their runs are listed as follows:

- *VGG* and *MMD-VGG* submitted by Wenhui Li, Shu Xiang, Heyu Zhou, Weizhi Nie, Anan Liu, and Yuting Su from Tianjin University, China (Section 4.1);
- *TCL1*, *TCL2*, and *TCL3* submitted by Xiaolong Liu, Xinwei He, Zhichao Zhou, Yang Zhou, Song Bai and Xiang Bai from Huazhong University of Science and Technology, China (Section 4.2);
- *RNSRAP1*, *RNSRAP2*, and *RNSRAP3* submitted by Minh-Triet Tran, Tu-Khiem Le, Van-Tu Ninh, Khac-Tuan Nguyen, Ngoc-Minh Bui, Vinh Ton-That, Trong-Le Do, Vinh-Tiep Nguyen, Minh N. Do and Anh-Duc Duong from Vietnam National University, Vietnam and the University of Illinois at Urbana-Champaign, USA (Section 4.3).

4. Methods

4.1. MMD-VGG: Maximum Mean Discrepancy Domain Adaptation on the VGG-Net, by W. Li, S. Xiang, H. Zhou, W. Nie, A. Liu, and Y. Su

They proposed the Maximum Mean Discrepancy based on the VGG model (MMD-VGG) to tackle sketch-based 3D scene retrieval task. The query data are 2D scene sketches and the target data are 3D scene models. Obviously, those two types of data have diverse data distribution. They address this task from two settings, learning-based setting and non-learning based setting. As the query data and target data have different types, the first step of the algorithm is data preprocessing. Then, they use the processed data to learn the feature representation of sketch images and scene models.

4.1.1. Data preprocessing

As the scene models are 3D format, they use the 3D design software SketchUp to automatically extract the views of all the 3D models. The input data type of SketchUp is the SKP and the output of SketchUp is the 480*480 view image. **Fig.** 3 shows several example images resulting from transforming 3D scene models to 2D images.

4.1.2. Feature representation

After using SketchUp to extract the images of a 3D model, the 2D-to-3D retrieval problem can be transformed into a 2D-to-2D task. For the feature representation, they use the training data to learn the feature representation model for the testing dataset, that is, a learning-based setting or a non-learning based setting to represent the features of the complete dataset.

4.1.2.1. Learning-based setting

They employ the deep networks VGG [SZ14] pretrained on the Places dataset [ZLK*17] as the initial network parameters and fine-tune the network on the 180 sketch images and 700 images of 3D models. Then, they use the output of the last but one (fc7) layer as the feature representation for each image. It is obvious that the divergence between the sketch image and the scene image is quite huge even though they depict the same category. The divergence makes it difficult for cross-domain similarity measurement. In this algorithm, they adopt Maximum Mean Discrepancy [LWD*13] to compute the difference in both marginal and conditional distributions from two different domains and construct a feature representation by using the Principal Component Analysis (PCA) method. After the above steps, the features of two domains have been projected into a common space and then can be measured by using the Euclidean distance.

4.1.2.2. Non-Learning based setting

For the complete data, they directly use the VGG [SZ14] model which is pretrained on the Places dataset [ZLK*17] to extract the

features of sketch images and model images. Then, they use the Euclidean distances between the scene sketch images and the images of 3D scene models as their similarities to generate the retrieval results.

4.2. TCL: Triplet center loss, by X. Liu, X. He, Z. Zhou, Y. Zhou, S. Bai and X. Bai

Their method is based on a two-stream CNN which processes samples from either domain with a corresponding CNN stream. Based on triplet center loss [HZZ*18] and softmax loss supervision, the network is trained to learn a unified feature embedding for each sample, which is then used for similarity measurement in the following retrieval procedure. Below is the detailed description of the method.

4.2.1. View Rendering

Their approach exploits the view-based representations of 3D scene models. For each 3D scene model (with color texture), they render it into multiple color images from N_v (N_v = 12 in their experiments) view directions. Each view image is of size 256×256. To fit the pre-defined CNNs during training, images of size 224×224 are randomly cropped as input from these rendered view images. While for testing, they only take the center crop of the same size from each view image.

4.2.2. Network Architectures

An overview of the feature learning network is depicted in **Fig.** 4. Considering the huge semantic gap between images and 3D scene models, they adopt two separate CNN streams for samples from the two different domains. A normal CNN (Stream 1) is used to extract the features of sketches. While the MVCNN [SMKLM16] framework (Stream 2) is adopted to obtain features from the rendered view images. In their experiment, these two streams are based on the same backbone (e.g. VGG11-bn [SZ14]). But note that their parameters are not shared.

4.2.3. Triplet Center Loss

In order to increase the discrimination of the features, they adopt triplet center loss (TCL) [HZZ*18] for feature learning. Given a batch of training data with M samples, they define TCL as

$$L_{tc} = \sum_{i=1}^{M} \max\left(D(f_i, c_{y_i}) + m - \min_{j \in C \setminus \{y_i\}} D(f_i, c_j), 0\right) \quad (1)$$

where $D(\cdot)$ represents the squared Euclidean distance function. y_i and f_i and are the ground-truth label and the embedding for sample i respectively. C is the label set. c_{y_i} (or c_j) is the center of embedding vectors for class c_{y_i} (or j). Intuitively, TCL is to enforce the distances between the samples and their corresponding center c_{y_i} (called *positive center*) smaller than the distances between the samples and their nearest *negative center* (i.e. centers of other classes $C \setminus \{y_i\}$) by a margin m. For a better performance, softmax loss is also employed.

4.2.4. Retrieval

After training, they first extract features for all the testing samples, including sketches and 3D scene models. Then they calculate the similarity matrix between the sketches and 3D scene models using Euclidean distance metric. To improve the final retrieval performance, the re-ranking algorithm they use is the same as that of GIFT [BBZ*16]. They experiment with three runs: *Run 1* only uses a single VGG11-bn model, while *Run 2* and *Run 3* use the ensemble results of different models including VGG11-bn, ResNet50 [HZRS16] and ResNet101 [HZRS16] but have different re-ranking parameters.

4.3. RNSRAP: ResNet50-Based Sketch Recognition and Adapting Place Classification for 3D Models Using Adversarial Training, by M. Tran, T. Le, V. Ninh, K. Nguyen, N. Bui, V. Ton-That, T. Do, V. Nguyen, M. Do, and A. Duong

4.3.1. Sketch Recognition with ResNet50 Encoding

In sketch classification task, they employ the output of ResNet50 [HZRS16] to encode a sketch image into a feature vector of 2048 elements. Due to the extremely small-scale data in sketch data, it is difficult to use only the extracted features to train their neural network model directly, so they create variant samples by data augmentation. From the original training dataset, different variations of a sketch image can be generated. Regular transformations can be applied, including flipping, rotation, translation, and cropping. From the saliency map of an image, they extract different patches with their natural boundaries corresponding to different entities in the image and synthesize other sketch images by matting these patches. By this way, they enrich the training dataset with 2000 images.

They construct two types of fully connected neural networks. The first network type contains two hidden layers to train extracted feature vectors. The number of nodes in the first and second hidden layers are 256 and 128, respectively. The second network type uses only one hidden layer with 200 nodes. Extracted features from ResNet50 of all training sketch images, including the original and synthesized extra samples, are used to train different classification models conforming the two proposed neural network structures.

Owing to the small-scale training data, Batch Gradient Descent with Adam optimizer is used to minimize the cross entropy loss function in the training process [KB14]. The output scores are processed through softmax function to provide proper predicted probability for each class.

They improve the performance and accuracy of their system by training multiple classification networks with different initializations for random variables for the two types of neural networks. They fuse the results of those models by using the majority-vote scheme to determine the label of a sketch query image.

They use ASUS-NotebookSKU X541UV, Intel(R) Core(TM) i5-6198DU CPU @ 2.30GHz, 8 GB Memory, and 1 x NVIDIA GeForce 920MX. The training time for a classification model is about 30 minutes. It takes less than 1 second to predict the category of a sketch image.

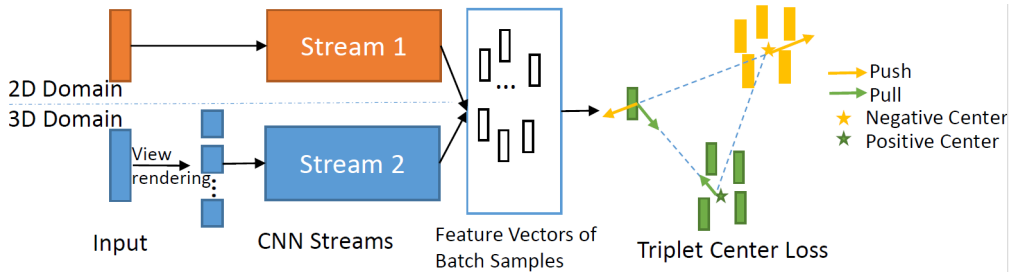


Figure 4: *An overview of the network architecture. They adopt two separate CNN streams to extract features for different domains. Triplet center loss and softmax loss (not shown here) are used as the supervision loss.*

4.3.2. Saliency-Based Selection of 2D Screenshots

For a 3D model, there exist multiple viewpoints to capture screenshots, some capture the general views of the model while others focus on a specific set of entities in the scene. They randomly generate multiple screenshots from different viewpoints at 3 different scales: general views, views on a set of entities, and views on a specific entity. Screenshots with many occlusions are removed. Then, they estimate the saliency map of a screenshot with DHSNet [LH16] to evaluate if this view has sufficient human-oriented visually attracted details. By this way, they generate a set of visually information rich screenshots for each 3D model. In this task, experimental results show that using no more than 5 appropriate views can be sufficient to classify the place of a 3D model with high accuracy.

4.3.3. Place Classification Adaptation for 3D Models

Adversarial training is a promising approach for training robust deep neural network. Adversarial approaches are also possible to unsupervised domain adaptation [THSD17, SLZ*17]. They apply the adversarial adaptive method to minimize the distance between the source and target mapping distributions. This approach aims to create an efficient target mapping model due to substantial variance between the two domains.

In this work, the source domain is a set of natural images that are used to train Places365-CNN models, while the target domain is a set of 3D place screenshots that are captured from given 3D models. Inspired by the idea of adversarial discriminative domain adaptation for face recognition [THSD17], they propose their method to train the target mapping model so as to match the source distribution for place classification. **Fig.** 5 illustrates the overview of their proposed method to adapt a place classification system from natural images to screenshots of 3D models. They first train a target representation M_t to encode a screenshot of a 3D model into a feature vector that cannot be distinguished with the feature from a natural image by the domain discriminator. Then they train a classifier C that can correctly classify target images.

In the ***Adversarial Adaptation*** step, a natural image is encoded by a source representation M_s and a screenshot of a 3D model is encoded by a target representation M_t. The goal of this step is to learn M_t so that the discriminator cannot distinguish the domain of a feature vector encoded by either M_s or M_t. They keep the source representation M_s fixed and train the target representation M_t using a basic adversarial loss until the feature maps of the two domains are indistinguishable by the discriminator. By this way, they obtain a transformation to match the target distribution (screenshots from 3D models) with the source distribution (natural images).

In the ***Classification for Target Domain*** step, they use M_t to encode screenshots of 3D models and train a classifier with data from the training dataset. The label for a 3D model is determined by voting from the results of its selected screenshots with the coefficient weights corresponding to the prediction score of each view. To further boost the overall accuracy for place classification of 3D models from 2D screenshots, they train multiple classifiers with the same network structure and assemble the output results with voting scheme. They use Google cloud machines n1-highmem-2, each with 2 vCPUs, Intel(R) Xeon(R) CPU @ 2.50GHz Intel Xeon E5 v2, 13 GB Memory, and 1 x NVIDIA Tesla K80.

4.3.4. Ranking Generation

Because of the wide variation of sketch images, for each sketch image in the test set, they consider up to the two best labels of the sketch image, then retrieve all related 3D models (via their common labels), and finally sort all retrieved items (3D models) in ascending order of dissimilarity.

- Single-labeled sketch image: they select all the 3D models corresponding to the label of a sketch image and insert them into the rank list in a descending order of confidence scores measuring the possibility that a 3D model belongs to that category.
- Multi-labeled sketch image: the similarity score between a sketch image and a 3D model is determined by the product of the confidence score of the sketch image and that of the 3D model. All 3D models in categories related to a sketch image are inserted into the rank list and sorted in descending order of similarity, i.e. ascending order of distance.

In this track, they submit 3 runs as follows:

- Run 1: they use the single label of a sketch image from one network in Type 1 and the single label of a 3D model from one place classification model.
- Run 2: they use the single label of a sketch image from the fusion of 3 networks (one Type 1 and two Type 2 networks) and the single label of a 3D model from the fusion of 5 place classification models.
- Run 3: they use the two best labels of a sketch image from one

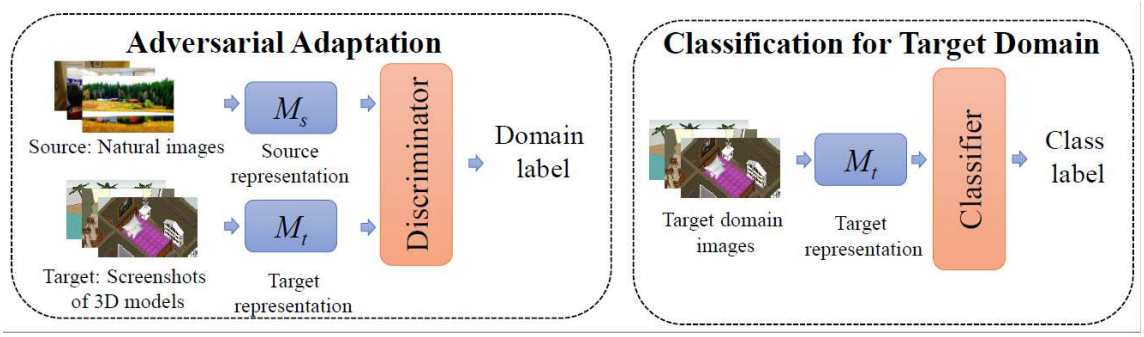


Figure 5: *Place classification for screenshots of 3D models with adversarial discriminative domain adaptation.*

network in type 1 and the single label of a 3D model from the fusion of 5 place classification models.

5. Results

In this section, we perform a comparative evaluation of the eight runs of the four methods submitted by the three groups. We measure retrieval performance based on the seven metrics mentioned in Section 2.4: PR, NN, FT, ST, E, DCG and AP. **Fig.** 6 and **Table** 2 compare three learning-based participating methods and one non-learning-based participating method on the testing and complete dataset, respectively.

As shown in the aforementioned figure and table, in the learning based category, Tran's RNSRAP algorithm (run 2) performs the best, followed by Liu's TCL method (run 3), while the overall performance of all the learning-based methods are close to each other. In the non-learning based category, there is only one participating method, whose performance is much inferior if compared with learning-based ones. More details about the retrieval performance of each individual query of every participating method can be found on the track homepage [YLL18].

Though we cannot directly compare non-learning based approaches and learning-based approaches together, we have found much more promising results in learning-based approaches. The CNNs contribute a lot to the top performance of those three learning-based approaches. Considering many latest sketch-based 3D model retrieval methods utilize deep learning techniques, we regard it as the currently most popular and promising machine learning technique for 2D/3D feature learning and related retrieval. In fact, the three methods that adopt certain deep learning models also perform well when adapted to this challenging benchmark.

Based on the same target 3D scene dataset of this Sketch-Based 3D Scene Retrieval (SBR) track, we also organized another SHREC'18 track on 2D Image-Based 3D Scene Retrieval (IBR) [ARYLL18], whose 2D query image dataset contains 1000 images for each of the 10 classes. The IBR track also has almost the same four participating methods, while as can be seen from the corresponding figures and tables, for the same method each performance metric achieved on the IBR track is significantly better than that on the SBR track. We believe at least the following three differences of IBR contribute to its better performance: (1) it has a much larger query dataset which is very helpful for the training of the deep neural networks; (2) compared with the query sketches of SBR, there is much more accurate 3D shape information in IBR's query images; and (3) each of IBR's query images has additional color information to correlate to the texture information existing in the 3D scene models. Therefore, there is a much smaller semantic gap to bridge between the query and target datasets for the IBR track, while the SBR track is much more challenging due to a big semantic gap there.

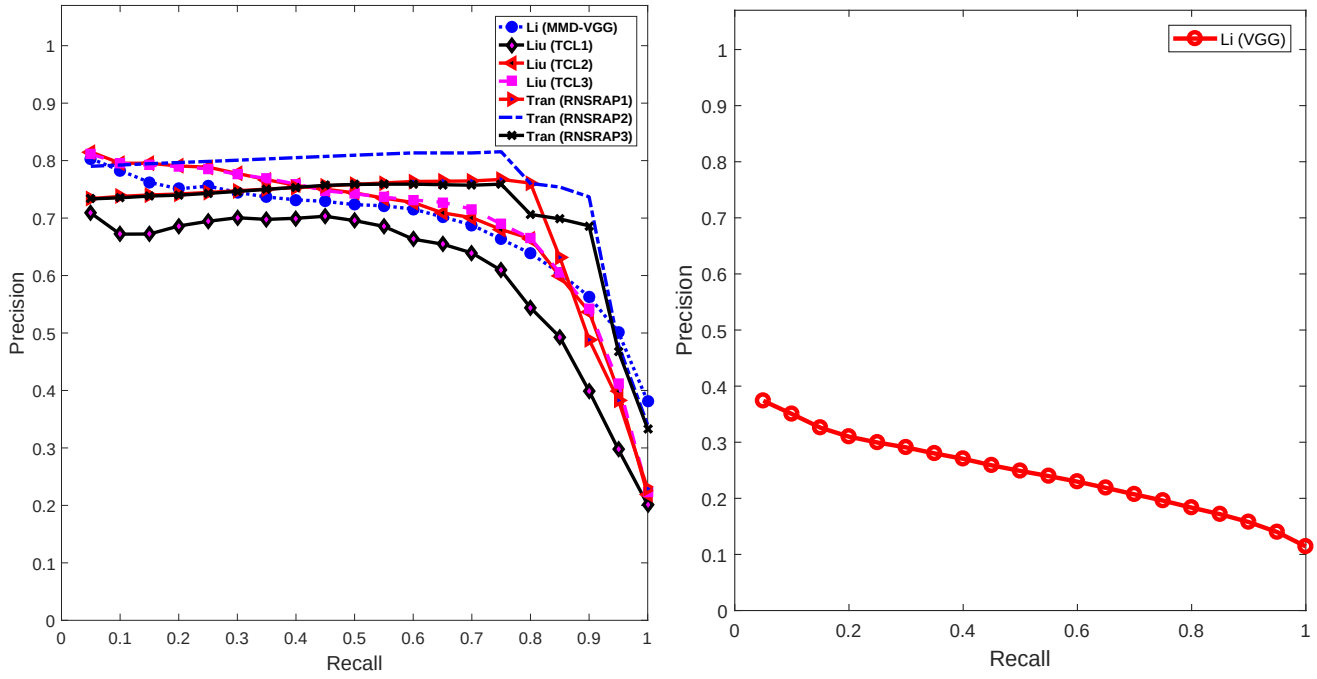
Finally, we classify all the participating methods with respect to the techniques employed: all the three participating groups (Li, Liu, Tran) utilize local features. All of the three groups (Li, Liu, Tran) employ deep learning framework to automatically learn the features. But Tran further applies regular transformations and adversarial training as well. On the other hand, Li and Liu directly compute the 2D-3D distances based on the distributions of sketches and models by using the Euclidean distance metric, while Tran conducts the retrieval based on 2D/3D classification.

6. Conclusions and Future Work

6.1. Conclusions

Due to the semantic gap existing between the inaccurate 2D scene sketch queries and more accurate 3D scene model representations for the same scene that we want to search in this scenario of 2D scene sketch-based 3D scene model retrieval, learning a deep model is potential in bridging the gap. In conclusion, this 2D scene sketch-based 3D scene model retrieval track is to further foster the challenging and interesting research direction of sketch-based 3D model retrieval, encouraged by the success of SHREC'12 [LSG*12, LLG*14], SHREC'13 [LLG*13, LLG*14], SHREC'14 [LLL*14, LLL*15] and SHREC'16 [LLD*16] sketch-based 3D shape retrieval tracks.

Though 2D scene sketch-based 3D scene retrieval is even more challenging than 2D sketch-based 3D model retrieval and 2D image-based 3D model retrieval, we still have three groups who have successfully participated in the track and contributed eight runs of four methods. This track provides a common platform to solicit current 2D scene sketch-based 3D scene model retrieval approaches in terms of this 2D sketch-based 3D scene retrieval scenario. We also hope that the **SceneSBR** benchmark, together with



(a) Learning-based approaches (testing dataset)

(b) Non-Learning based approaches (complete benchmark)

Figure 6: *Precision-Recall diagram performance comparisons on two different datasets of our **SceneSBR** benchmark for three learning-based and one non-learning based participating method.*

Table 2: *Performance metrics comparison on two different datasets of of our **SceneSBR** benchmark for three learning-based and one non-learning based participating method.*

Participant	Method	NN	FT	ST	E	DCG	AP
Testing dataset							
Li	MMD-VGG	0.771	0.630	**0.835**	0.633	0.856	0.685
Liu	TCL1	0.643	0.582	0.753	0.579	0.810	0.606
	TCL2	**0.814**	0.630	0.794	0.626	0.860	0.688
	TCL3	0.800	0.640	0.801	0.633	0.861	0.691
Tran	RNSRAP1	0.729	0.658	0.659	0.637	0.826	0.689
	RNSRAP2	0.786	**0.729**	0.734	**0.707**	**0.864**	**0.757**
	RNSRAP3	0.729	0.652	0.766	0.637	0.834	0.707
Complete benchmark							
Li	VGG	0.336	0.262	0.428	0.151	0.684	0.243

the retrieval results we have obtained in the track, will become a useful reference for researchers in this community.

6.2. Future Work

This track not only helps us identify state-of-the-art methods, but also existing problems, current challenges and future research directions for this important, new and interesting research topic.

- **Building a large-scale and/or multimodal 2D scene-based 3D scene retrieval benchmark.** Our proposed **SceneSBR** contains only ten scene classes, which is one of the reasons that all the three deep learning-based participating methods have achieved excellent performance. However, since scalability to a large scale retrieval and 2D/3D format diversities is very important for related applications, we plan to significantly extend the **SceneSBR** benchmark by incorporating much more scene categories, as well as more modalities in either 2D query (i.e. sketches and images) or 3D target (i.e. RGD-D scenes, LIDAR scene images, and other scene range scans produced by other range scanners) format. Then, we will invite people to adapt and run their algorithms on the new benchmark again to evaluate their scalability in a large-scale and/or multimodal 3D scene retrieval scenario.
- **Semantics-driven 2D scene sketch-based 3D scene retrieval.** To improve either the accuracy or efficiency of a 2D scene sketch-based 3D scene retrieval algorithm, we need to consider utilizing the semantic information existing in both a 2D scene

sketch query and all the 3D scene target models. None of the four participating methods has exploited this already available semantic information. We believe related applications (i.e. online 3D scene retrieval, 3D Entertainment contents development, and autonomous driving cars) will benefit a lot from the retrieval based on extracted semantic information in both the queries and targets.

- **Application-oriented 2D scene-based 3D scene retrieval.** Developing a 2D scene-based 3D scene retrieval dedicated for a related application, such as creating 3D scene contents for a new 4D immersive program, like Disney World's Avatar Flight of Passage Ride [Wik18, Att18, tM18], or for retrieving domain-specific 3D scenes such as indoor/outdoor scenes, sand table models for real estate applications, rainforest scenes for cartoon or movie production, and so on.
- **Developing new deep learning models specially for this research topic.** According to the evaluation, we have found promising performance achieved by deep learning techniques. However, due to limited competition time, most of the participating methods are a straight-forward application of a retrieval algorithm developed for another purpose. Therefore, we have confidence to believe that their performance will be elevated further if they consider the characteristics of this retrieval problem, or even better develop new deep learning models which fit this scenario well.
- **Interdisciplinary research directions.** We have noticed the more outstanding performance achieved by Tran's 3D scene retrieval algorithm RNSRAP which is based on sketch and model classification. According to our previous class-based or semantic information-based 3D model retrieval research experience [LJ13, LLJF17], it is a promising approach to further improve retrieval accuracy, especially for NN, and FT since we can push more 3D scene models classified into one class forward to the front part of a retrieval rank list.

Acknowledgments

This project is supported by the University of Southern Mississippi Faculty Startup Funds Award to Dr. Bo Li.

References

[ARYLL18] ABDUL-RASHID H., YUAN J., LI B., LU Y.: SHREC'18 2D Scene Image-Based 3D Scene Retrieval Track Website. http://orca.st.usm.edu/~bli/SceneIBR2018/, 2018. 6

[Att18] ATTRACTIONS W.: 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. 1, 8

[BBZ*16] BAI S., BAI X., ZHOU Z., ZHANG Z., LATECKI L. J.: GIFT: A real-time and scalable 3D shape search engine. In *CVPR* (2016), IEEE, pp. 5023–5032. 4

[HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *CVPR* (2016), pp. 770–778. 4

[HZZ*18] HE X., ZHOU Y., ZHOU Z., BAI S., BAI X.: Triplet center loss for multi-view 3D object retrieval. In *CVPR* (2018). 4

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *CoRR abs/1412.6980* (2014). 4

[LH16] LIU N., HAN J.: DHSNet: Deep hierarchical saliency network for salient object detection. In *CVPR* (2016), pp. 678–686. 5

[LJ13] LI B., JOHAN H.: 3D model retrieval using hybrid features and class information. *Multimedia Tools Appl. 62*, 3 (2013), 821–846. 8

[LLD*16] LI B., LU Y., DUAN F., DONG S., FAN Y., QIAN L., LAGA H., LI H., LI Y., LIU P., OVSJANIKOV M., TABIA H., YE Y., YIN H., XUE Z.: SHREC'16: 3D sketch-based 3D shape retrieval. In *3DOR 2016* (2016). 6

[LLG*13] LI B., LU Y., GODIL A., SCHRECK T., AONO M., JOHAN H., SAAVEDRA J. M., TASHIRO S.: SHREC'13 track: Large scale sketch-based 3D shape retrieval. In *3DOR* (2013), pp. 89–96. 6

[LLG*14] LI B., LU Y., GODIL A., SCHRECK T., BUSTOS B., FERREIRA A., FURUYA T., FONSECA M. J., JOHAN H., MATSUDA T., OHBUCHI R., PASCOAL P. B., SAAVEDRA J. M.: A comparison of methods for sketch-based 3D shape retrieval. *CVIU 119* (2014), 57–80. 2, 6

[LLJF17] LI B., LU Y., JOHAN H., FARES R.: Sketch-based 3D model retrieval utilizing adaptive view clustering and semantic information. *Multimedia Tools Appl. 76*, 24 (2017), 26603–26631. 8

[LLL*14] LI B., LU Y., LI C., GODIL A., SCHRECK T., AONO M., BURTSCHER M., FU H., FURUYA T., JOHAN H., LIU J., OHBUCHI R., TATSUMA A., ZOU C.: SHREC'14 Track: extended large scale sketch-based 3D shape retrieval. In *3DOR* (2014), pp. 121–130. 6

[LLL*15] LI B., LU Y., LI C., GODIL A., SCHRECK T., AONO M., BURTSCHER M., CHEN Q., CHOWDHURY N. K., FANG B., FU H., FURUYA T., LI H., LIU J., JOHAN H., KOSAKA R., KOYANAGI H., OHBUCHI R., TATSUMA A., WAN Y., ZHANG C., ZOU C.: A comparison of 3D shape retrieval methods based on a large-scale benchmark supporting multimodal queries. *CVIU 131* (2015), 1–27. 2, 6

[LSG*12] LI B., SCHRECK T., GODIL A., ALEXA M., BOUBEKEUR T., BUSTOS B., CHEN J., EITZ M., FURUYA T., HILDEBRAND K., HUANG S., JOHAN H., KUIJPER A., OHBUCHI R., RICHTER R., SAAVEDRA J. M., SCHERER M., YANAGIMACHI T., YOON G.-J., YOON S. M.: SHREC'12 track: Sketch-based 3D shape retrieval. In *3DOR* (2012), pp. 109–118. 6

[LWD*13] LONG M., WANG J., DING G., SUN J., YU P. S.: Transfer feature learning with joint distribution adaptation. In *ICCV* (2013), pp. 2200–2207. 3

[SLZ*17] SOHN K., LIU S., ZHONG G., YU X., YANG M., CHANDRAKER M.: Unsupervised domain adaptation for face recognition in unlabeled videos. *CoRR abs/1708.02191* (2017). 5

[SMKLM16] SU H., MAJI S., KALOGERAKIS E., LEARNED-MILLER E.: Multi-view convolutional neural networks for 3D shape recognition. In *ICCV* (2016), pp. 945–953. 4

[SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556* (2014). 3, 4

[THSD17] TZENG E., HOFFMAN J., SAENKO K., DARRELL T.: Adversarial discriminative domain adaptation. *CoRR abs/1702.05464* (2017). 5

[tM18] THE MAGIC I.: 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. 1, 8

[Tri18] TRIMBLE: 3D Warehouse. http://3dwarehouse.sketchup.com/?hl=en, 2018. 2

[Wik18] WIKIPEDIA: Avatar flight of passage. http://en.wikipedia.org/wiki/Avatar_Flight_of_Passage, 2018. [Online; accessed 1-March-2018]. 1, 8

[YLJ16] YE Y., LU Y., JIANG H.: Human's scene sketch understanding. In *ICMR '16* (2016), pp. 355–358. 2

[YLL18] YUAN J., LI B., LU Y.: SHREC'18 2D Scene Sketch-Based 3D Scene Retrieval Track Website. http://orca.st.usm.edu/~bli/SceneSBR2018/, 2018. 1, 6

[ZLK*17] ZHOU B., LAPEDRIZA A., KHOSLA A., OLIVA A., TORRALBA A.: Places: a 10 million image database for scene recognition. *IEEE Trans. on PAMI* (2017). 3