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

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

Hameed Abdul-Rashid[†1], 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

2D scene image-based 3D scene retrieval is a new research topic in the field of 3D object retrieval. Given a 2D scene image, it is to search for relevant 3D scenes from a dataset. It has an intuitive and convenient framework which allows users to learn, search, and utilize the retrieved results for vast related applications, such as automatic 3D content generation for 3D movie, game and animation production, robotic vision, and consumer electronics apps development, and autonomous vehicles. To advance this promising research, we organize this SHREC track and build the first 2D scene image-based 3D scene retrieval benchmark by collecting 2D images from ImageNet and 3D scenes from Google 3D Warehouse. The benchmark contains uniformly classified 10,000 2D scene images and 1,000 3D scene models of ten (10) categories. In this track, seven (7) groups from five countries (China, Chile, USA, UK, and Vietnam) have registered for the track, while due to many challenges involved, only three (3) groups have successfully submitted ten (10) runs of five methods. To have a comprehensive comparison, seven (7) commonly-used retrieval performance metrics have been used to evaluate their retrieval performance. We also suggest several future research directions for this research topic. We wish this publicly available [ARYLL18] benchmark, comparative evaluation results and corresponding evaluation code, will further enrich and boost the research of 2D scene image-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

Provided with a 2D scene image, a 2D scene image-based 3D scene retrieval algorithm is to search for relevant 3D scenes (i.e., .OBJ or .SKP files) from a dataset. It is an intuitive and convenient framework which allows users to learn, search, and utilize the retrieved results for related applications. For example, automatic 3D content generation based on one or a sequence of captured images for AR/VR applications, or 3D movie, game and animation production, robotic vision (i.e. path finding), and consumer electronics apps development, which facilitate users to efficiently generate a 3D scene after taking an image of a real scene. It is also very promising and has great potentials in other related applications such as 3D geometry video retrieval, and highly capable autonomous vehicles like the Renault SYMBIOZ [Ren] [Tip].

However, there is little research in 2D scene image-based 3D scene shape retrieval [MSL*11] [XKH*16] due to at least two reasons: (1) the problem itself is challenging to cope with; (2) lack of related retrieval benchmarks. Seeing the benefit of advances in retrieving 3D scene models using 2D scene image queries makes the research direction meaningful, interesting and promising.

Deng et al. [DDS*09] collected the ImageNet database initially comprising of 5,247 synsets, defined as a set of one or more synonyms in WordNet [Mil95], and 3.2 million images back in 2009. Nearly ten years after its inception, there are over 21,000 synsets

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

indexed and nearly 14.2 million images. For this track, we build a smaller and more manageable dataset comprising of 10,000 scene images across 10 classes, each with 1,000 images. It avoids the bias issue since we have collected the same number of images for every class, while the images' variation within one class is also adequate enough.

To organize another track titled "SHREC'18 2D Scene Sketch-Based 3D Scene Retrieval" [YLL18], we have collected 1,000 3D Warehouse [Tri18] scene mesh models (in original .SKP format as well transformed .OBJ format) to correspond to the uniformly classified 250 scene sketches of 10 classes in the Scene250 sketch dataset [YLJ16]. For each class, we have collected the same number (100) of 3D scene models as well. Therefore, we reuse this 3D scene target dataset for this track and only need to collect 2D scene images as query data, which are not difficult to find since we have quite a few 2D scene image benchmarks, like ImageNet [DDS*09] and SUN [XEH*16] datasets.

This track [ARYLL18] is organized to promote this challenging research direction by soliciting state-of-the-art 2D scene image-based 3D scene retrieval methods and predict the future directions on this research topic. Evaluation code for computing a set of performance metrics similar to those used in the Query-by-Model retrieval technique is also provided online on the track's website [ARYLL18].

2. SceneIBR Benchmark

2.1. Overview

Our 2D scene image-based 3D scene shape retrieval benchmark **SceneIBR** utilizes 10,000 2D scene images selected from ImageNet [DDS*09] as its 2D scene image dataset and 1,000 3D Warehouse scene models (both .SKP and .OBJ formats) as its 3D scene dataset, and both have ten classes. Each of the ten classes contains the same number of 2D scene images (1,000 per class) and 3D scene models (100 per class).

To facilitate learning-based retrieval, for each class we randomly select 700 images and 70 models for training and use the remaining 300 images and 30 models for testing, as listed in **Table 1**. Participants are required to submit results on the testing dataset if they use a learning-based approach. Otherwise, the retrieval results on the complete (10,000 images, 1,000 models) dataset are needed. To provide a complete reference for future users of our **SceneIBR** benchmark, we evaluate the participating algorithms on both the testing dataset (300 images and 30 models per query) and the complete **SceneIBR** benchmark (1,000 images and 100 models per class).

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

SceneIBR Benchmark	Image	Model
Training	700	70
Testing	300	30
Total (per class)	1,000	100

2.2. 2D Scene Image Dataset

The 2D scene image query set is composed of 10,000 scene images (10 classes, each with 1,000 images) that are all from ImageNet [DDS*09], while all the classes have relevant models in the target 3D scene dataset which are downloaded from the 3D Warehouse [Tri18]. One example per class is demonstrated in **Fig. 1**.

Figure 1: *Example 2D scene images (one example per class) in our **SceneIBR** benchmark.*

2.3. 3D Scene Dataset

The 3D scene dataset is built on the selected 1,000 3D scene models downloaded from Google 3D Warehouse. Each class has 100 3D scene models. One example per class is shown in **Fig. 2**.

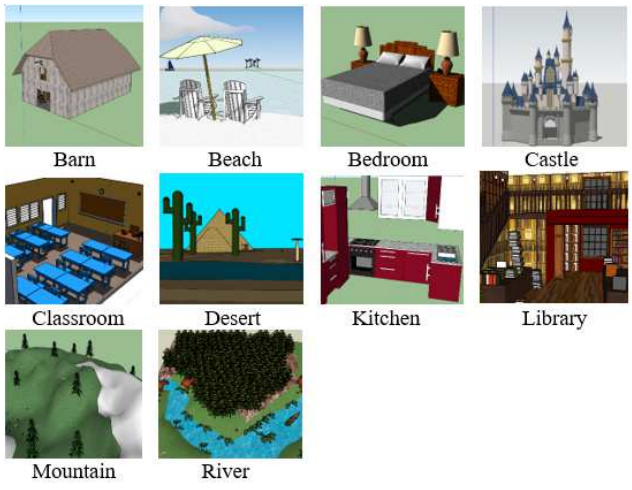


Figure 2: *Example 3D scene models (one example per class) in our **SceneIBR** benchmark.*

2.4. Evaluation Method

The objective of this track is to evaluate the performance of different 2D scene image-based 3D scene retrieval algorithms using

a 2D image query dataset and a collected 3D warehouse model dataset. While, for each algorithm, it will target retrieving target 3D scene models that share the same class as that of a query 2D scene image. To have a comprehensive evaluation of the retrieval algorithm, we employ seven commonly adopted performance metrics in 3D model retrieval technique [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, which is can be downloaded from the track's website [ARYLL18].

3. Participants

There were six groups who registered for the track. Both China and USA have two groups, while the rest two groups come from Chile and Vietnam, respectively. Each group was given three weeks to complete the competition. They were asked to submit both their results and methods description before the deadline.

However, due to a new and challenging track and limited competition time as well, only three groups have finally participated in the track. Ten (10) rank list results (runs) for five (5) different methods developed by three (3) groups have been submitted to the track. The participating groups 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);
- *RNIRAP1*, *RNIRAP2*, *RNIRAP3*, *BoW1*, and *BoW2* 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 (Sections 4.3∼ 4.4).

4. Methods

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

4.1.1. Overview

They proposed the Maximum Mean Discrepancy domain adaption based on the VGG model (MMD-VGG) to address scene image-based 3D scene retrieval problem, where the query is a 2D scene image and the target is 3D scene models. Those two types of data come from different datasets with diverse data distribution. They address this task from two settings, learning-based setting and non-learning based setting. This method mainly contains two successive steps, data preprocessing and feature representation.

4.1.2. Data preprocessing

For 3D scene data, they use SketchUp, which is a very popular and easy-to-use 3D design software, to capture the representative views of all the 3D models automatically. The format of the input model is SKP and the output of the model in SketchUp is a 480*480 image. Several example representative views are shown in **Fig. 3**.

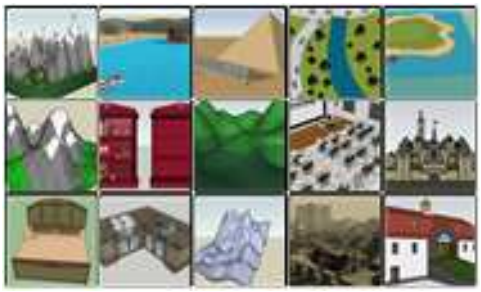


Figure 3: *Several example representative views.*

4.1.3. Feature representation

After obtaining the representative views of all the 3D models, the 2D-to-3D retrieval task can be transformed into a 2D-to-2D task. For the feature representation, they use two settings: learning-based setting and non-learning based setting.

4.1.4. Learning-based setting

Inspired by the impressive performance of deep networks, they employ the VGG [SZ14] model pretrained on the Places [ZLK*17] as the initial network parameters. Then, they combine the 7000 scene images and 700 representative views of 3D models to fine-tune the network. Finally, they use the output of last but one fully connected layer as the feature representation of each image.

It is obvious that the domain divergence between the target and the query is quite huge. A scene image dataset and a 3D scene dataset can own different visual features even though when they depict the same category, which makes it difficult for cross-domain 3D model retrieval. Since the fine-tuning operation can only moderately reduce the divergence between these two datasets, they apply a domain adaption method to help to solve the cross-domain problem. In this algorithm, they aim to find a unified transformation which learns a new common space for features from two different domains. In detail, the nonparametric Maximum Mean Discrepancy [LWD*13] is leveraged to measure the difference in both marginal and conditional distributions. Then, they unify it by Principal Component Analysis (PCA) to construct a feature representation which is robust and efficient for the domain shift reduction. After the domain adaptation, the features of two domains are projected into a common space. They measure the similarity between the query and target directly by computing their Euclidean distance.

4.1.5. Non-learning based setting

For non-learning based setting, they directly use the VGG [SZ14] model pretrained on the Places dataset to extract the features of images/views. Then, they directly compute the Euclidean distance between scene images and representative views of 3D models to measure the similarity.

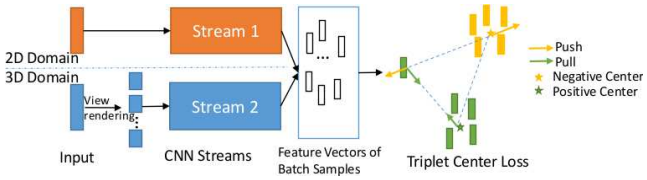


Figure 4: *Illustration of the network architecture. Two separate CNN streams are used to extract features for the two domains. Triplet center loss along with softmax loss (not depicted here) is used to optimize the whole network.*

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

They adopt a deep neural network composed of two CNN streams (see **Fig. 4**) that process samples from 2D scene image samples and 3D scene samples respectively. In order to learn a discriminative and robust representation from either domain, triplet center loss (TCL) [HZZ*18] and softmax loss are employed. The final learned embeddings are used for the retrieval task. In the following sections, a brief description of their approach is presented.

4.2.1. Network Architectures

Because of the great difference between 3D scene models and 2D scene images, two independent CNN streams are used to handle the two kinds of samples respectively. Stream 1 is a normal CNN that extracts features of 2D scene images. While Stream 2 is adapted from Stream 1 to the MVCNN [SMKL15] framework, which takes N_v ($N_v = 12$ in their experiment) view images of a 3D scene model as input. The last fully connected layer of each stream outputs a N_c-dimension embedding vector, where N_c is the number of categories.

4.2.2. Learning

For its good performance in 3D shape retrieval, triplet center loss (TCL) [HZZ*18] is adopted for feature learning. Given a set of M training samples $\{(x^i, y^i)|x^i \in X, y_i \in C\}_{i=1}^{M}$, TCL is defined 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 f_i and y_i are the embedding and label for the i-th sample respectively. $D(\cdot)$ denotes the Euclidean distance function. c_j is the center (average) of embeddings for samples with the label j. Intuitively, TCL is to push the distances between the samples and their nearest *negative center* (namely the center of any other class $C \setminus \{y_i\}$) larger than the samples and the *positive centers* c_{y_i} by a margin m. To achieve a better performance, they also use softmax loss.

4.2.3. Retrieval

In the testing stage, the two CNN streams are employed to extract the feature embeddings of both the 2D scene images and the 3D scene models, respectively. Euclidean distance is adopted as the distance metric. To further improve the retrieval performance, an efficient re-ranking algorithm utilized in GIFT [BBZ*16] is taken as a post-processing step. Three runs with different experimental settings are provided, they are, *Run1* with a single VGG11-bn model as the backbone network, *Run2* and *Run3* which are the ensemble results computed using different backbone models including VGG11-bn [SZ14], ResNet50 [HZRS16] and ResNet101 [HZRS16] and different re-ranking parameter settings.

4.3. RNIRAP: ResNet18-Based 2D Scene Image Recognition with Scene Attributes and Adapting Place Classification for 3D Models Using Adversarial Training, M. Tran, V. Ninh, T. Le, K. Nguyen, V. Ton-That, N. Bui, T. Do, V. Nguyen, M. N. Do, A. Duong

4.3.1. 2D Scene Classification

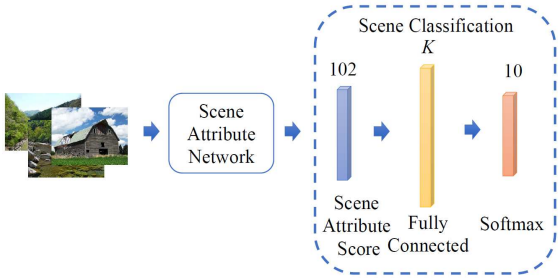


Figure 5: *2D scene classification with scene attributes.*

A 2D scene image can be classified into one of the ten categories by using the scene attributes of that image, such as open area, indoor lighting, natural light, wood, etc. Thus, they employ the output of Places365-CNNs [ZLK*17] as the input feature vector for their neural network. They choose the ResNet18 model in the core of Place365 network and extract the scores of its scene attributes which yield a vector of 102 elements. By feeding the model with 7000 training 2D scene images, they obtain a training data with a dimension of 7000×102 used as the input vector for the 2D scene classification task.

Their classification model is a fully connected neural network having one hidden layer with K nodes, $100 \leq K \leq 200$ (see **Fig. 5**). A training algorithm called Batch Gradient Descent with Adam optimizer [KB14] is used to minimize the cross entropy loss function in training process. The output scores are processed through softmax function to provide the predicted probability for each class. It should be noticed that some query images may be classified into more than one categories. For example, some images contain a river but also has a mountain in the background. Thus, they assign up to two best predicted classes to each 2D scene query image.

They also improve the performance and accuracy of the retrieval system by training multiple classification networks with different numbers of nodes K in the hidden layer and different initializations for random variables. Finally, they obtain five classification models with the same structure and fuse the results of those models with the voting scheme to determine the label of a 2D scene query image. It takes about one hour to train each classification model on an ASUS X541UV Notebook with an Intel(R) Core(TM) i5-6198DU CPU @ 2.30GHz, 8 GB Memory, and 1 x NVIDIA GeForce 920MX.

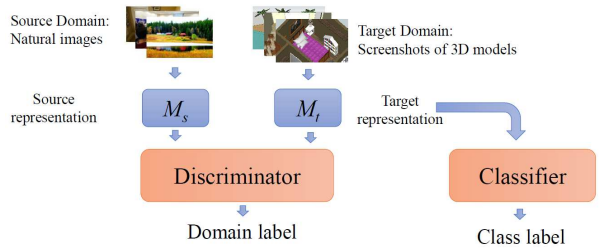


Figure 6: *Adversarial adaptation for place classification on 3D model views.*

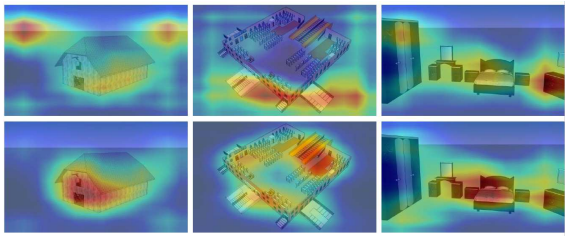


Figure 7: *Heatmaps of informative regions for place prediction on three 3D model views before adaptation (the first row) and after adaptation (the second row).*

4.3.2. Saliency-Based 2D Views Selection

One common approach to classify or recognize a 3D object is to capture multiple views of the object and then apply appropriate methods for 2D images to process those views. Instead of using all the views, they propose a saliency-based method to obtain some views with high level of human-oriented visually attracted details.

For each 3D model, they randomly capture multiple views at 3 different levels of details: general views, views focusing on a set of entities, and detailed views on a specific entity. They use DHSNet [LH16] to generate the saliency map of each view and select promising views of each 3D model for the place classification task. Their experiments on the dataset of this track demonstrate that a 3D model can be classified with high accuracy (more than 92%) with no more than 5 information-rich views.

4.3.3. Place Classification Adaptation for 3D Models

If a 3D model is carefully designed and illuminated, its views may look realistic and indistinguishable to photos captured from natural environments. However, the views of regular 3D models are usually not realistic enough to be processed directly with existing methods for natural images, including Place365-CNN. Therefore, it is necessary to train a transformation model to make the views of 3D models be adaptive to the scene analysis and place classification of natural images.

The adversarial network exhibits adaptability with many different domains [THSD17, SLZ*17]. In this work, they apply the adversarial adaptive method to minimize the distance between the natural image domain mapping distribution and the 3D view domain mapping distribution. **Fig. 6** illustrates their proposed method for adversarial adaptation for place classification on 3D model views with two main components:

- **Adversarial Adaptation component** aims to reduce the distance between the two domains. They use the adversarial network structure which is almost similar to the regular generative adversarial network. In this case, the target domain plays the same role as latent space and the target representation is the generator. The discriminator is responsible for distinguishing between source images and target images given the mapping features. An adversarial loss is applied to train the robust generator to fool the discriminator. Until target images are indistinguishable by discriminator, the distance between the two domains is expected to be reduced.
- **Place Classification component** aims to train a classifier corresponding to target labels using traditional cross entropy loss.

Fig. 7 demonstrates the heatmaps of informative regions for place prediction on 3D model views before adaptation (the first row) and after adaptation (the second row). These examples show that by adversarial adaptation, the system can better focus on appropriate areas in a view to give correct place prediction.

4.3.4. Rank List Regeneration

They assign one or two best labels for each query image, and retrieve all 3D models having such labels. The similarity between a query image and a 3D model is determined by the product of the prediction score of the query image and that of the 3D model on the same label. For a single classification model, they use the prediction score directly from the network. For a fusion result from multiple classification models, they use the voting ratio as the prediction score.

There are 3 runs submitted with this proposed method.

- Run1: they use the result of a single model for 2D scene classification and a single model for 3D place classification.
- Run2: they use the fusion result from 5 models for 2D scene classification and the fusion result of the other 5 models for 3D place classification.
- sRun3: they use two best labels of a single model for 2D scene classification and the fusion result of the other 5 models for 3D place classification.

After retrieving all relevant 3D models into a rank list, all other 3D models which are considered irrelevant are inserted in the tail of that rank list with the distance of infinity.

4.4. BoW: Bag-of-Words Framework Based Retrieval, M. Tran, V. Ninh, T. Le, K. Nguyen, V. Ton-That, N. Bui, T. Do, V. Nguyen, M. N. Do, A. Duong

The same participating group as that of Section 4.3 contributed another two runs based on the Bag-of-Words framework. In this approach, they do not train a model to classify a 2D scene image or a 3D model. Instead, their non-learning based method takes advantage of their framework on Bag-of-Word retrieval [NNT*15, LWA*17] to determine the category of a 2D scene (query) and a 3D model (target). They also employ the same method to generate a set of useful views for each 3D model (see Section 4.3).

For both 2D scene images and 3D model views, they follow the same retrieval process. First, they apply the Hessian Affine detector to detect the interest points N in each image, either a 2D scene image or a 2D view of a 3D model. They use RootSIFT without angle for keypoint descriptors and train the codebook using Approximate K-Means algorithm with 1 million codewords. They perform the quantization on all training images with KDTree data-structure to calculate the BoW representation of each frame. They also perform soft assignment with 3 nearest neighbors, L2 asymmetric distance measurement [ZJS13], TF-IDF weighting, and average pooling for each representation.

For each unlabeled 2D scene image, they retrieve a rank list of relevant images. Then they determine the top-M most voted labels from those of the retrieved images and assign these candidate labels to the input image. In this task, they choose $M = 1$ or 2. Similarly, they also determine the top M most voted labels for each 2D view, then assign the most reliable label to the corresponding 3D model.

The codebook training module using Python 2.7 is deployed on a computer with a Ubuntu 14.04 OS and 2.4 GHz Intel Xeon E5-2620 v3 CPU, and 64 GB RAM. It takes 2 hours to create a codebook with 1 million visual words from 15 million features. The retrieval process in Matlab R2012b with feature quantization and dissimilarity matrix calculation is performed on a computer with a Windows Server 2008 R2 OS, a 2.2 GHz Intel Xeon E5-2660 CPU, and 12 GB RAM. It takes less than 1 second to perform the retrieval for each image.

There are two runs in this method. In this first run, they determine only one label for each scene image and only one label for each 3D model. In the second one, they determine up to two labels for each scene image and up to two labels for each 3D model.

5. Results

In this section, we perform a comparative evaluation of the ten runs of the five 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. 8** and **Table 2** compare three learning-based participating methods on the testing dataset and two non-learning-based participating methods on the complete dataset, respectively.

As shown in the aforementioned figure and table, in the learning-based category, Tran's RNIRAP algorithm (run 3) performs the best, closely followed by Li's MMD-VGG and Liu's TCL method (run 3), which are close to each other as well. That is, the performance of all the three learning-based methods are similar to each other. In the non-learning based category, Li's VGG algorithm outperforms Tran's BoW method. For each participating method, more details about the rank list and evaluated retrieval performance of each query can be found on the track website [ARYLL18].

Although it is not fair to compare non-learning based approaches with learning-based approaches, it is easy to find that the learning-based approaches have produced much more appealing accuracies. In Tran's top-performing learning based approach RNIRAP, in terms of automatically learning the features, the deep learning approach Place365-CNN [ZLK*17] contributes a lot to its better accuracy among the learning based approaches. Except BoW, all the other four participating methods are almost exactly the same as the four participating methods in our organized "SHREC'18 2D Scene Sketch-Based 3D Scene Retrieval" track [YLL18]. Therefore, similarly, we think deep learning is also currently the best candidate to handle this image-based 3D scene retrieval problem. Compared with the performance achieved in our SHREC'18 Sketch-Based 3D Scene Retrieval (SBR) track [YLL18], the performance achieved in this Image-Based 3D Scene Retrieval (IBR) track is significantly better. We conclude that IBR's following three factors contribute to its superior performance: (1) a 40 times larger query dataset for better training of the deep networks; (2) compared with sketches, images contain much more accurate 3D shape information; and (3) images also have additional color information to correspond to texture information of 3D scene models. In a word, the semantic gap between the query and target datasets is much smaller, compared with that of SBR. This difference also makes the SBR track an even more challenging task for us to have further explorations.

Finally, all the five participating methods are categorized according to the techniques they employed. All the three learning-based methods (MMD-VGG, TCL, RNIRAP) from three participating groups (Li, Liu, Tran) utilize deep learning techniques to automatically learn local features. Therefore, all of the three groups have considered the deep learning framework for feature learning. In the non-learning based category, Tran's BoW method employs the Bag-of-Words, while Li's VGG method uses a pre-trained model VGG to directly extract local features. Only Tran utilizes a classification-based 3D model retrieval framework.

6. Conclusions and Future Work

6.1. Conclusions

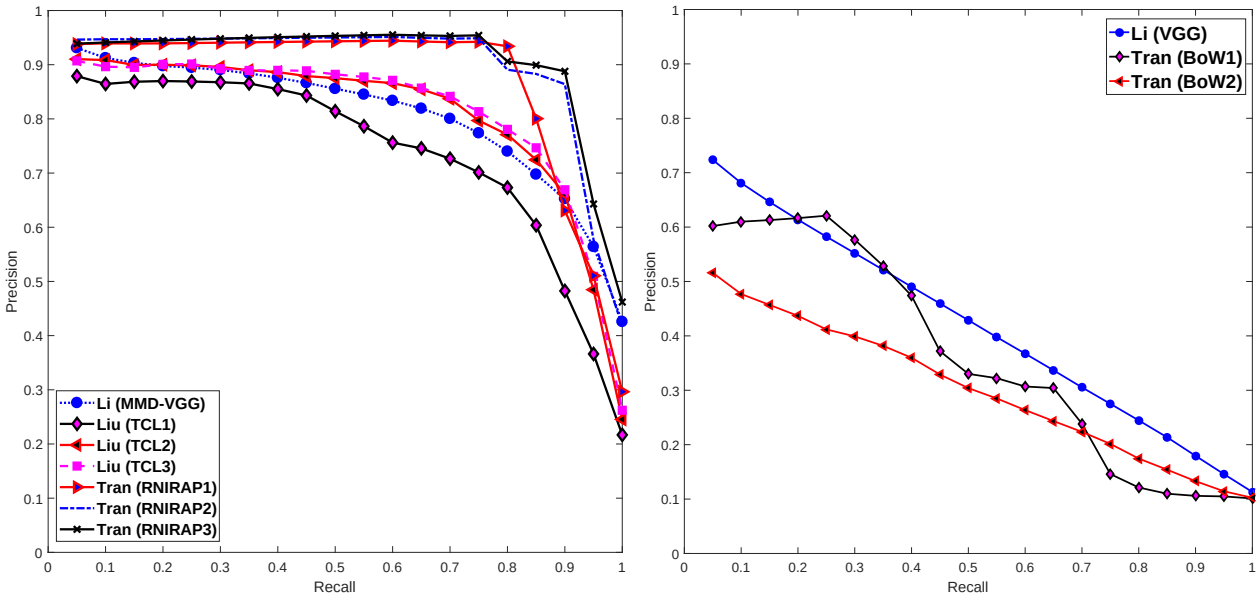
2D scene image-based 3D scene retrieval is a new, challenging but also important research direction in the research field of 3D object retrieval. To promote the development of this research direction, we build the first 2D scene image based 3D scene retrieval benchmark **SceneIBR** and organize this SHREC track. It is a direct extension upon our successfully organized 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 more challenging than before, we still have three groups who have successfully participated in the track and contributed ten runs of five methods.

This track provides a common platform to solicit current 3D model retrieval approaches in terms of this 2D image-based 3D scene retrieval scenario. It also helps us identify state-of-the-art methods as well as future research directions for this research area. We also hope that the 2D image-based retrieval benchmark, together with 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 issues, current challenges and future research directions for this important, new and interesting research problem.

- **Large-scale 2D scene-based 3D scene retrieval benchmarks**



(a) Learning approaches (testing dataset) (b) Non-Learning-based approaches (complete benchmark)

Figure 8: *Precision-Recall diagram performance comparisons on two different datasets of our **SceneIBR** benchmark for the three learning-based and two non-learning based participating methods.*

Table 2: *Performance metrics comparison on two different datasets of our **SceneIBR** benchmark for the three learning-based and two non-learning based participating methods.*

Participant	Method	NN	FT	ST	E	DCG	AP
Testing dataset							
Li	MMD-VGG	0.910	0.750	0.899	0.750	0.929	0.8032
Liu	TCL1	0.823	0.689	0.856	0.687	0.900	0.7327
	TCL2	0.871	0.751	0.888	0.759	0.927	0.8028
	TCL3	0.864	0.760	0.893	0.762	0.927	0.8086
Tran	RNIRAP1	0.864	0.760	0.893	0.762	0.927	0.8086
	RNIRAP2	**0.944**	**0.882**	0.890	**0.854**	0.954	0.8931
	RNIRAP3	0.936	0.875	**0.941**	0.850	**0.958**	**0.9018**
Complete benchmark							
Li	VGG	**0.719**	**0.416**	**0.585**	**0.291**	**0.803**	**0.4139**
Tran	BoW1	0.575	0.316	0.396	0.272	0.735	0.3602
	BoW2	0.501	0.311	0.469	0.196	0.719	0.2984

supporting multimodal queries and targets. There are only ten scene categories in our proposed **SceneIBR** benchmark, and this also partially explains the excellent performance that has been achieved by the three deep learning-based participating methods. An algorithm's scalability to a large-scale retrieval scenario or diverse 2D/3D data formats is crucial for many practical applications. Therefore, we will build a large-scale 2D based 3D scene retrieval benchmark which has substantially more representative scene classes and also supports diverse types of 3D scene models, like RGB-D, LIDAR or other types of 3D range scans produced by certain range scanners. The query dataset will also contain both 2D images and 2D sketches.

- **Semantic 2D scene image-based 3D scene retrieval.** There is a lot of semantic information in both the 2D query images and the 3D target scene models in our current **SceneIBR** benchmark. This type of information is special for this research problem and will be highly important to further improve the retrieval performance or for related real-time applications which requires a high-level (i.e. semantic-level) retrieval due to their higher standard on response time. However, we find that there is no participating group that has considered this, probably due to limited time for the competition. Therefore, we can expect even better performance if they also incorporate the semantic information into their methods.
- **Classification-based retrieval.** It can be found that class-based or classification-based 3D model retrieval (i.e. Tran's RNIRAP) is potential to achieve even better performance compared to other algorithms which utilize a more traditional 3D model retrieval

pipeline. This also coincides with our prior findings related to class-based 3D model retrieval [LJ13] or semantic information-based 3D model retrieval [LLJF17]. This relatively new framework contributes to better NN, FT and the overall performance metrics such as DCG and AP.

- **VR/AR applications.** It seems promising and interesting to apply 2D scene-based 3D scene retrieval algorithms to develop some VR/AR apps. For example, automatic 3D Entertainment contents generation based on the video clips captured by consumer cameras (see example applications listed in Section 1), or retrieving domain-specific 3D scenes based on a captured image is also among our considerations for the next work.

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. 1, 2, 3, 6

[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

[DDS*09] DENG J., DONG W., SOCHER R., LI L., LI K., LI F.: ImageNet: A large-scale hierarchical image database. In *CVPR* (2009), pp. 248–255. 1, 2

[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. 3, 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. 3, 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

[LWA*17] LIMBERGER F. A., WILSON R. C., AONO M., AUDEBERT N., BOULCH A., BUSTOS B., GIACHETTI A., GODIL A., SAUX B. L., LI B., LU Y., NGUYEN H. D., NGUYEN V., PHAM V., SIPIRAN I., TATSUMA A., TRAN M., VELASCO-FORERO S.: SHREC'17: Point-cloud shape retrieval of non-rigid toys. In *3DOR* (2017). 5

[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

[Mil95] MILLER G. A.: WordNet: A lexical database for english. *Commun. ACM 38*, 11 (1995), 39–41. 1

[MSL*11] MERRELL P., SCHKUFZA E., LI Z., AGRAWALA M., KOLTUN V.: Interactive furniture layout using interior design guidelines. *ACM Transactions on Graphics (TOG) 30*, 4 (2011), 87. 1

[NNT*15] NGUYEN V., NGO T. D., TRAN M., LE D., DUONG D. A.: A combination of spatial pyramid and inverted index for large-scale image retrieval. *IJMDEM 6*, 2 (2015), 37–51. 5

[Ren] RENAULT: Renault SYMBOIZ Concept. http://www.renault.co.uk/vehicles/concept-cars/symbioz-concept.html. 1

[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

[SMKL15] SU H., MAJI S., KALOGERAKIS E., LEARNED-MILLER E. G.: Multi-view convolutional neural networks for 3D shape recognition. In *ICCV* (2015), 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. In *CVPR* (2017), pp. 2962–2971. 5

[Tip] TIPS L. T.: Driving a multi-million dollar autonomous car. http://www.youtube.com/watch?v=vlIJfV1u2hM&feature=youtu.be. 1

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

[XEH*16] XIAO J., EHINGER K. A., HAYS J., TORRALBA A., OLIVA A.: SUN database: exploring a large collection of scene categories. *IJCV 119*, 1 (2016), 3–22. 2

[XKH*16] XU K., KIM V. G., HUANG Q., MITRA N., KALOGERAKIS E.: Data-driven shape analysis and processing. In *SIGGRAPH ASIA 2016 Courses* (2016), ACM, p. 4. 1

[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. 2, 6

[ZJS13] ZHU C., JEGOU H., SATOH S.: Query-adaptive asymmetrical dissimilarities for visual object retrieval. In *ICCV* (2013), pp. 1705–1712. 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, 4, 6