Eurographics Workshop on 3D Object Retrieval (2019)
S. Biasotti and G. Lavoué (Editors)

SHREC'19 Track: Extended 2D Scene Image-Based 3D Scene Retrieval

Hameed Abdul-Rashid[†‡1], Juefei Yuan[†‡1], Bo Li[†‡*1], Yijuan Lu[†‡2], Tobias Schreck[†‡3], Tianyang Wang[†‡4], Ngoc-Minh Bui[‡6,7], Trong-Le Do[‡6,7], Mike Holenderski[‡5], Dmitri Jarnikov[‡5], Khiem T. Le[‡6], Vlado Menkovski[‡5], Khac-Tuan Nguyen[‡6,7], Thanh-An Nguyen[‡7], Vinh-Tiep Nguyen[‡8], Tu V. Ninh[‡6], Perez Rey[‡5], Minh-Triet Tran[‡6,7]

[1] School of Computer Science and Computer Engineering, University of Southern Mississippi, USA
[2] Department of Computer Science, Texas State University, USA
[3] Institue of Computer Graphics and Knowledge Visualization, Graz University of Technology, Austria
[4] Department of Computer Science and Information Technology, Austin Peay State University, USA
[5] Department of Mathematics and Computer Science, Eindhoven University of Technology, Netherlands
[6] Faculty of Information Technology, Vietnam National University - Ho Chi Minh City, Vietnam
[7] Software Engineering Lab, Vietnam National University - Ho Chi Minh City, Vietnam
[8] University of Information Technology, Vietnam National University - Ho Chi Minh City, Vietnam

Abstract

In the months following our SHREC 2018 - 2D Scene Image-Based 3D Scene Retrieval ***(SceneIBR2018)*** *track [ARYL*18] [ARYLL18], we have extended the number of the scene categories from the initial 10 classes in the* ***SceneIBR2018*** *benchmark to 30 classes [ARYLL19], resulting in a new benchmark* ***SceneIBR2019*** *which has 30,000 scene images and 3,000 3D scene models. For that reason, we seek to further evaluate the performance of existing and new 2D scene image-based 3D scene retrieval algorithms using this extended and more comprehensive new benchmark. Three groups from the Netherlands, the United States and Vietnam participated and collectively submitted eight runs. This report documents the evaluation of each method based on seven performance metrics, offers an in-depth discussion as well as analysis on the methods employed and discusses future directions that have the potential to address this task. To further enrich the current state of 3D scene understanding and retrieval, our evaluation toolkit, all participating methods' results and the comprehensive 2D/3D benchmark have all been made publicly available.*

1. Introduction

2D scene image-based 3D scene model retrieval is to retrieve 3D scene models given an input 2D scene image. It has many important related applications, including highly capable autonomous vehicles like the Renault SYMBIOZ [Ren18] [Tip18], multi-view 3D scene reconstruction, VR/AR scene content generation, and consumer electronics apps, among others. However, this task is far from trivial and lacks substantial research due to the challenges involved as well as a lack of related retrieval benchmarks. Consequently, existing 3D model retrieval algorithms have been limited to focus on single object retrieval. 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.

To promote this interesting yet challenging research, we organized a 2018 Eurographics Shape Retrieval Contest (SHREC) track [ARYLL18] titled "2D Scene Image-Based 3D Scene Retrieval", by building the first 2D scene image-based 3D scene retrieval benchmark **SceneIBR2018**, comprising 10,000 2D scene images and 1,000 3D scene models. All the images and models are equally classified into 10 indoor as well as outdoor classes.

However, as can be seen, **SceneIBR2018** contains only 10 distinct scene classes, and this is one of the reasons that all the three deep learning-based participating methods have achieved excellent performance on it. Considering this, after the track we have tripled the size of **SceneIBR2018** , resulting in an extended benchmark **SceneIBR2019** [ARYLL19], which has 30,000 2D scene images and 3,000 3D scene models. Similarly, all the 2D images and 3D scene models are equally classified into 30 classes. We have kept the same set of 2D scene images and 3D scene models belonging to the initial 10 classes of **SceneIBR2018**.

Hence, this track seeks participants who will provide new contri-

† Track organizers
‡ Track participants
* Corresponding author. For any question related to the track, please contact Bo Li. bo.li@usm.edu. or li.bo.ntu0@gmail.com

butions to further advance 2D scene image-based 3D scene retrieval for evaluation and comparison, especially in terms of scalability to a larger number of scene categories, based on the new benchmark **SceneIBR2019**. Similarly, 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. Benchmark

2.1. Overview

Building process. Scene categories were selected from the Places scene recognition database [ZLK*17] and with the criteria of selection being *popularity*, in terms of the degree to which they are commonly seen. Through a three-person voting mechanism we selected the most popular 30 scene classes (including the initial 10 classes in **SceneIBR2018**) from the 88 scene classes of Places88 dataset [ZLK*18], which are shared by ImageNet [DDS*09], SUN [XHE*10], and Places [ZLK*17]. Instances for the additional 20 classes, were sourced from Flickr [Fli18] as well as Google Images [Goo18] for images and downloaded via 3D Warehouse [Tri18] for scene models.

Benchmark details. Our extended 2D scene image-based 3D scene retrieval benchmark **SceneIBR2019** expands the initial 10 classes of **SceneIBR2018** with 20 new classes totaling a more comprehensive dataset of 30 classes. **SceneIBR2019** contains a complete dataset of 30,000 2D scene images (1,000 per class) and 3,000 3D scene models (100 per class). Examples for each class are demonstrated in both **Fig. 1** and **Fig. 2**.

In the same manner as the **SceneIBR2018** track, we randomly pull 700 images and 70 models out from each class for training and the remaining 300 images and 30 models are used for testing, as shown in Table 1. If a method involves a learning-based approach, results for both the training and testing datasets need to be submitted. Otherwise, retrieval results based on the complete dataset are needed.

Table 1: Training and testing datasets information of our **SceneIBR2019** benchmark.

Datasets	Images	Models
Training (per class)	700	70
Testing (per class)	300	30
Total (per class)	1000	100
Total (all 30 class)	30,000	3,000

2.2. 2D Scene Image Dataset

The 2D scene image query set is composed of 30,000 scene images (30 classes, each with 1,000 images) that are all from the Flicker and Google Image websites. One example per class is demonstrated in **Fig. 1**.

2.3. 3D Scene Model Dataset

The 3D scene model 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. 2**.

Figure 1: Example 2D scene images (one example per class) in our **SceneIBR2019** benchmark.

2.4. Evaluation Method

To have a comprehensive evaluation of the retrieval algorithm, we employ seven commonly adopted performance metrics in the 3D model retrieval community: Precision-Recall (PR) diagram, Nearest Neighbor (NN), First Tier (FT), Second Tier (ST), E-Measure (E), Discounted Cumulated Gain (DCG) and Average Precision (AP) [LLL*15a]. We have developed the related code to compute these metrics [1].

3. Participants

Of the six groups (two from China, one from Japan, one from the Netherlands, one from the United States and one from Vietnam) who initially registered, only three were able to submit methods by the deadline. Each group was given one month to complete the contest and submit method results and description. In total, there are eight runs for the three different methods submitted by the three groups.

The participants and their runs are listed as follows:

- ***CVAE*-{1, 2, 3, 4}** and ***CVAE-VGG*** submitted by Perez Rey, Mike Holenderski, Dmitri Jarnikov and Vlado Menkovski from Eindhoven University of Technology in the Netherlands (Section 4.1);

[1] http://orca.st.usm.edu/~bli/SceneIBR2019.

Figure 2: Example 3D scene models (one example per class, shown in one view) in our **SceneIBR2019** benchmark.

- ***RNIRAP*-{1, 2}** submitted by Ngoc-Minh Bui, Trong-Le Do, Khac-Tuan Nguyen, Tu V. Ninh, Khiem T. Le, Thanh-An Nguyen, Minh-Triet Tran and Vinh-Tiep Nguyen from Vietnam National University - Ho Chi Minh City (Section 4.3);
- ***VMV-VGG*** submitted by Juefei Yuan, Hameed Abdul-Rashid, Bo Li, Tianyang Wang, Yijuan Lu from the University of Southern Mississippi, Austin Peay State University, and Texas State University (Section 4.2).

4. Methods

4.1. CVAE: Conditional Variational Autoencoders for Image Based Scene Retrieval, by P. Rey, M. Holenderski, D. Jarnikov and V. Menkovski

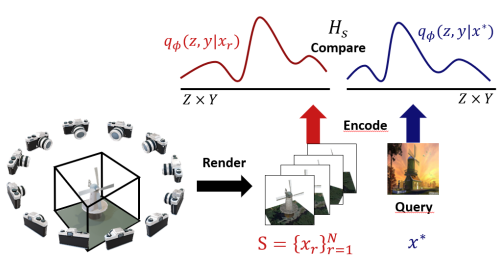

Figure 3: Overview of Scene Sampling and CVAE Distribution Learning

4.1.1. Overview

The proposed approach consists of image to image comparison with conditional variational autoencoders (CVAE) [KMRW14], as shown **Fig.** 3. The CVAE is a semi-supervised method for approximating the underlying generative model that produces a set of images and their corresponding class labels in terms of the so-called unobserved latent variables. Each of the input images is described in terms of a probability distribution over the latent variables and the classes.

Their approach consists of using the probability distributions calculated by the CVAE for each image as a descriptor. The comparison between an image query and the 3D scene renderings is with respect to the probability distributions obtained from the CVAE. The method consists of data pre-processing, training and retrieval described in the following subsections.

4.1.2. Data Preprocessing

Thirteen renderings are obtained for each of the 3D scenes. Each of the 3D scenes has a predefined view when loaded into the SketchUp software. This view is saved as a 2D view together with twelve views at different angles around the scene as in [SMKLM15]. The training data set consists of the 3D scene renders together with the training images. All images are resized to a resolution of 64×64 and all pixel values are normalized to the interval $[0,1]$. Image data augmentation is carried out by performing a horizontal flip to all images. The corresponding data space is $X = [0,1]^{64\times64\times3}$, while the 3 represents the color space.

4.1.3. Training

The CVAE consists of an encoder and a decoder neural network. The encoder network calculates from an image $x \in X$ the parameters of a probability distribution over the latent space $Z = \mathbb{R}^d$ and over the thirty class values in $Y = \{1,2,3,\ldots,30\}$. The decoder network calculates from a latent variable $z \in Z$ and a class $y \in Y$, the parameters of a distribution over the data space X .

The distributions for the encoder correspond to a normal distribution over Z and a categorical distribution over Y. A normal distribution over X is chosen for the decoder. The probabilistic model used corresponds to the M2 model described in the article [KMRW14]. Both the encoding and decoding neural networks are convolutional.

The CVAE is fed with batches of labeled images during training. The loss function is the sum of the negative Evidence Lower Bound (ELBO) and a classification loss. The ELBO is approximated by means of the parametrization trick described in [KMRW14, KW13] and represents the variational inference objective. The classification loss for their encoding distributions over Y corresponds to the cross entropy between the probability distribution over Y with respect to the input label.

4.1.4. Retrieval

After training, an image $x \in X$ can be described as a conditional joint distribution over $Z \times Y$. The density $q_\phi(z|x)$ corresponds to a normal distribution and $q_\phi(y|x)$ to a categorical distribution over Y,

where ϕ represents the weights of the encoder neural network. The joint density corresponds to $q_\phi(z,y|x) = q_\phi(z|x)q_\phi(y|x)$.

The similarity D between an input query image $x^* \in X$ and a 3D scene in terms of its N rendered images $S = \{x_r\}_{r=1}^N$ is given by the minimum symmetrized cross entropy H_s between the query and the rendered images' probability distributions (see **Fig.** 3).

$$D(x^*,S) \min_{r\in\{1,2,\ldots,13\}} H_s(q_\phi(z|x^*),q_\phi(z|x_r)) + \alpha H_s(q_\phi(y|x^*),q_\phi(y|x_r)). \quad (1)$$

They have used the parameter $\alpha = 64 \times 64 \times 3$ to increase the importance of label matching. A ranking of 3D scenes is obtained for each query according to this similarity.

4.1.5. Five Runs

They have sent five submissions corresponding to methods who differ only on the architecture of the encoding and decoding neural networks. These are described as follows:

1. **CVAE-(1,2,3,4)**: CVAE with different CNN architectures for the encoder and decoder.
2. **CVAE-VGG**: CVAE with features from pre-trained VGG [Kal17] on the Places data set [ZLK*18] as part of the encoder.

4.2. VMV-VGG: View and Majority Vote Based 3D Scene Retrieval Algorithm, by J. Yuan, H. Abdul-Rashid, B. Li, T. Wang, Y. Lu

The View and Majority Vote based 3D scene retrieval algorithm (VMV) utilizes the VGG-16 architecture, as illustrated in **Fig.** 4.

4.2.1. 3D Scene View Sampling

Each 3D scene model is in a 3D sphere observable by an automated QMacro that captures 13 scene views. Of these 13 unique perspectives, 12 are uniformly sampled along the equator of the sphere while the last view is from a top-down perspective as shown in **Fig.** 5.

4.2.2. Data Augmentation

They implemented several augmentations (e.g rotations, translations and reflections) on the dataset to avoid overfitting [YLL16]. These augmentations extended the dataset to be 500 times its initial size.

4.2.3. Pre-training and Fine-tuning

They preformed domain adaption with VGG2 on the Places scene image dataset [ZLK*17] for 100 epochs. After this adaption phase, another phase of domain adaption is performed on VGG2 with the 2D scene views training dataset, respectively.

4.2.4. Image/ View Classification and Majority Vote-Based Label Matching

Probability distributions of classifications were obtained from the trained VGG2 with the target 2D scene views testing dataset. A query image and each model's 13 scene views are used to generate a rank list for the query by using a majority vote-based label matching method.

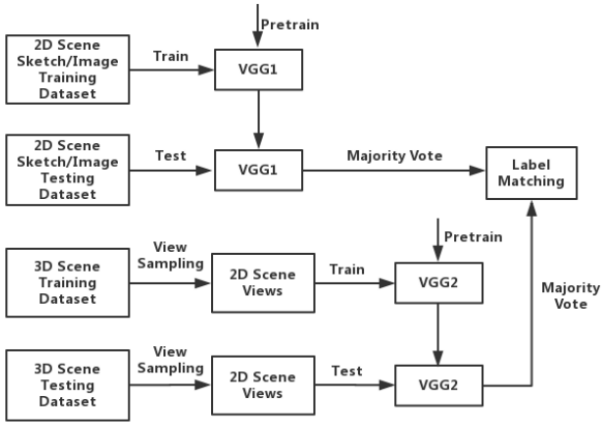

Figure 4: VMV-VGG architecture [ARYLL19].

Figure 5: A 13 sampled scene view images example of an apartment scene model [ARYLL19].

4.3. RNIRAP: ResNet18-Based 2D Scene Image Recognition with Scene Attributes and Adapting Place Classification for 3D Models Using Adversarial Training, by [CLARIFY DISCREPNCY OF 1ST AUTHOR] N. Bui, T. Do, K. Nguyen, T. Ninh, K. Le, T. Nguyen, M. Tran and V. Nguyen

4.3.1. 2D Scene Image Classification with Scenes' Deep Features

To classify an image into one of the 30 scene categories in this track, they apply their method (used in **SceneIBR2018** [ARYLL18]) to extract scenes' deep features using MIT Places API [ZLK*17]. They train a simple network with the extracted features from Places API and use this network to classify an input image with 30 labels.

In their first step, an input image is represented as a feature vector in Places API domain vector space using a pre-trained ResNet-50 [HZRS15] model on the MIT Places API scene recognition network. Instead of using 102 scene attributes as in their previous **SceneIBR2018** competition, they use a 512-dimensional deep feature representation which is generated before being processed through dense layers for classification.

Next, they utilize the extracted features to train a neural network classification with 30 labels. Different from their method used in the **SceneIBR2018**, the input feature is processed through two dense hidden layers with a dimension of 1024 for each layer, instead of a small network of $100 \leq K \leq 200$ dimensions as stated in their previous method. The visualization of their network configuration is demonstrated in **Fig.** 6. The network is trained on a server with 1 $\times$ NVIDIA Tesla K80 GPU. An Adam optimizer with learning rate at 0.0001 being hyperparameters. Three models were trained using this network configuration. The final label prediction of an image is outputted by using a majority voting scheme from these three models.

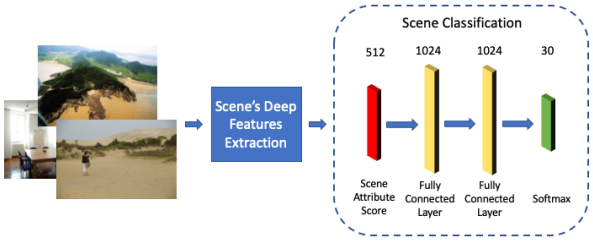

Figure 6: 2D scene classification with scenes' deep features.

4.3.2. 3D Scene Classification with Multiple Screenshots, Domain Adaptation, and Concept Augmentation

They suggest two steps for 3D scene classification as shown in **Fig.** 7. In the first step, they use a mixture of multiple classification models.

First, they employ ResNet-50 [HZRS15] model pretrained on the ImageNet [DDS*09] and Places365 [ZLK*17] datasets to extract a feature vector for each sampled scene view. Then they implement

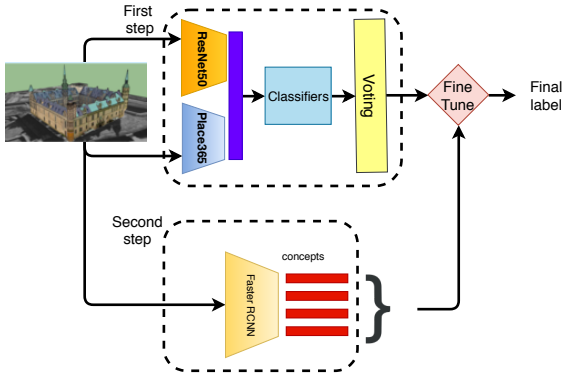

Figure 7: Two-step process of the 3D scene classification method.

different neural network architectures to train for the classification task. In order to find the best architecture, they try several configurations of a fully-connected neural network, with the number of hidden layers ranging from one to two while the number of neurons in each layer can be 128, 192, 256 or 320. The architecture that achieves the best accuracy is chosen for the voting scheme.

To utilize the scene attribute information more efficiently, they extract the 365-dimensional features from Places365 and directly concatenate with features extracted by ResNet-50. Some of the scene attribute features are useful and informative for the classification task, such as the attributes of "outdoor" and "swimming" can relate to the "beach" category. However, concatenating the two feature vectors may cause the model to overfit data and slow down the training process. Therefore, an additional step of normalizing the features and reducing the dimension to 512 using Principal Component Analysis (PCA) is applied. Finally, they continue to classify on this feature set.

They also collect images from the same set of 30 categories of the Places365 dataset and from the Internet, each category contains 1,000 images. Then they train a model using this customized dataset and obtain the weights to initialize the weights of a model when trained on the sampled views dataset.

Following their **SceneIBR2018** method, they apply the adversarial adaptive method to minimize the distance between the representation of the 3D model and the representation of the corresponding image. Their method contains two main components: the Adversarial Adaptation component, and the Place Classification component. In the adversarial adaptation component, a source representation model M_s will process a natural image into a feature vector and a target representation model M_t will process a screenshot of a 3D model into a second feature vector. The two encoded vectors are then fed into a discriminator to distinguish the two domains. They train the target representation M_t to fool the discriminator via a basic adversarial loss. In the place classification component, they train a classifier whose input is the learned representation of the 3D model. Multiple 2D scene views are sampled from the 3D model and processed by the trained classifier. The final label of the 3D model is selected from the votes of its sampled views. In order to further improve the accuracy, a number of classifiers that share the

same architecture are trained to predict the final label. The results of the classifiers are assembled via a voting scheme.

Because of the wide variation in the design of a 3D scene, it is not enough to classify the category of a scene simply by extracting the feature (from ResNet50) or from the features of scene attributes (from MIT Place, even after domain adaptation). This motivated them to employ object/entity detectors to identify entities based on hierarchical semantics present in each sampled view.

In the second step of the proposed method, they reuse the dataset of natural images collected from the Internet to train object detectors with Faster RCNN [RHGS15] for entities that might appear in a scene, such as "book" (in a library), "umbrella" (in a beach), etc. Using this list of scene semantics detected in sampled views, they further refine their results.

5. Results

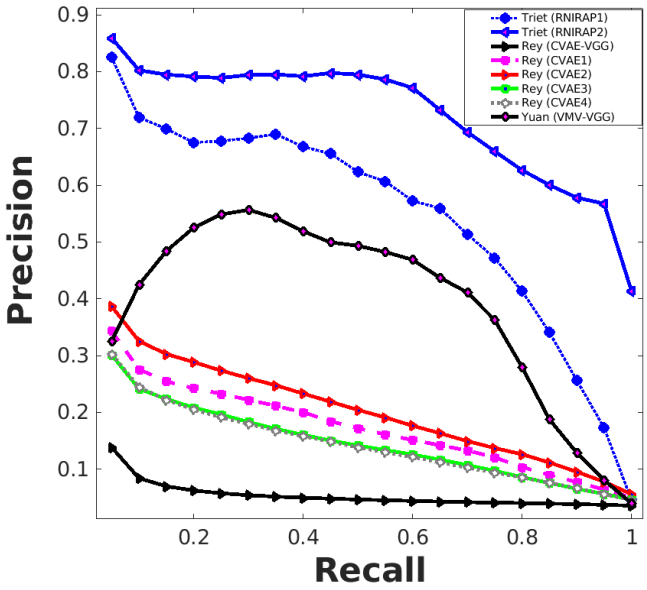

Figure 8: Precision-Recall diagram performance comparisons on testing dataset of of our **SceneIBR19** benchmark for the three learning-based participating methods.

A comparative evaluation has been performed on all methods. The measured retrieval performance is based on the seven metrics mentioned in Section 2.4: PR, NN, FT, ST, E, DCG and AP. **Fig.** 8 and **Table** 2 compare the three learning-based participating methods on the testing dataset.

As can be seen in the aforementioned figure and table, Tran's RNIRAP algorithm (run 2) performs the best, followed by the baseline method VMV-VGG and the CVAE method (CVAE2). More details about the retrieval performance of each individual query of every participating method are available on the **SceneIBR2019** track website [ARYL*19] .

Firstly, during this year's track all the three methods submitted by the three participating groups are leaning-based methods, while there is no submission involving a non-learning based approach. In addition, all of the three methods have employed a deep neural networks based learning approach.

Secondly, we could further classify the submitted approaches at a finer granularity. Both RNIRAP and VMV-VGG utilize CNN models and a classification-based approach, which contribute a lot to their better accuracies. While, the CVAE-based method uses a conditional VAE generative model and resulted latent features to measure the 2D-3D similarities.

Therefore, according to these two years' SHREC tracks (SHREC'19 and SHREC'18) on this topic, deep learning-based techniques are still the most promising and popular approach in tackling this new and challenging research direction.

In direct comparison to the results from **SceneIBR2018**, **SceneIBR2019** results do not preform as well. This is to be expected since the 10 scene categories in the **SceneIBR2018** benchmark were distinct and have few correlations. As explored by Yuan, J. et al [ARYLL19], the significant drop in performance can be attributed to the introduction of many correlating scene categories.

Finally, we would like to compare the performance of SHREC'19 two related tracks on the topic of 3D scene retrieval. Similarly, this year in a parallel way we also have organized another SHREC'19 track on "Extended 2D Sketch-Based 3D Scene Retrieval" [ARYL*19], based on the same target 3D scene dataset and a different query dataset which contains 25 sketches for each of the 30 classes. Except CVAE, these two tracks share other two participating methods (with minor differences). It is the second time that we have found that the performance achieved in this extended "Image-Based 3D Scene Retrieval (IBR)" track is significantly better, compared with that achieved on the back to back extended "Sketch-Based 3D Scene Retrieval (IBR)" track. This should be attributed to the same reasons as we have concluded in [ARYL*18]: IBR has a much larger query training dataset which contains images, instead of sketches, that have much more details and color information as well, which makes the semantic gap between the 2D query image and 3D target scenes much smaller.

6. Conclusions and Future Work

6.1. Conclusions

This track provided participants with the most diverse and comprehensive 2D/3D scene dataset to date, in hopes to advance 3D scene retrieval. Participating groups have explored many different approaches to solve the intractable task of 2D to 3D scene understanding.

Considering the importance of this research direction and its large amount of applications, we built the first 2D scene image-based 3D scene retrieval benchmark in SHREC'18 [ARYL*18]. This year, we have further extended the number of categories from 10 to 30, which further extends the line of our SHREC related research work on image-based 3D shape retrieval: SHREC'12 [LSG*12, LLG*14], SHREC'13 [LLG*13, LLG*14], SHREC'14 [LLL*14,LLL*15b], SHREC'16 [LLD*16], SHREC'18 [ARYLL18] and this year's SHREC'19 [ARYL*19].

Though even more challenging than last year, we still have three

Table 2: Performance metrics comparison on the SHREC'19 SceneIBR Track Benchmark.

Participant	Method	NN	FT	ST	E	DCG	AP
Complete benchmark							
Triet	RNIRAP1	0.845	0.620	0.674	0.618	0.791	0.5436
	RNIRAP2	**0.865**	**0.749**	**0.792**	**0.745**	**0.863**	**0.7221**
Rey	CVAE-VGG	0.071	0.054	0.099	0.055	0.405	0.0535
	CVAE1	0.235	0.187	0.295	0.189	0.532	0.1717
	CVAE2	0.272	0.217	0.331	0.219	0.560	0.2013
	CVAE3	0.199	0.154	0.251	0.157	0.507	0.1445
	CVAE4	0.211	0.149	0.246	0.152	0.505	0.1424
Yuan	VMV-VGG	0.122	0.458	0.573	0.452	0.644	0.3899

groups who have successfully participated in the track and contributed eight runs of three methods. Based on the number of (six) registrations, we also have found that it seems that this image-based retrieval track has attracted more potential contributors, compared to our sketch-based retrieval track. We believe this should be partially related to its relatively fewer difficulties and more broad applications as well. Extended from SHREC'18 [ARYLL18], this track, together with its benchmark and retrieval results, will become an even more useful resource for the researchers that are interested in this topic as well as many related applications.

6.2. Future Work

This track not only provides us with a common platform to solicit the retrieval performance (including scalability) from current 2D image-based 3D scene retrieval algorithms, but also offers us an opportunity to further identify state-of-the-art approaches as well as future research directions for this research area.

- **Large-scale benchmarks.** Our **SceneIBR2019**, even as the largest benchmark for 2D scene image-based 3D scene retrieval, has only thirty scene categories, which is far from large-scale. This again can partially explain the still relatively good performance that has been achieved by the top deep learning-based participating methods. However, we did see an apparent drop in the overall performance. Therefore, testing the scalability of a retrieval algorithm with respect to a large-scale retrieval scenario and various 2D/3D data formats is very important for many practical applications. Therefore, our next target is to build a large-scale benchmark which supports multiple modalities of 2D queries (i.e. images and sketches) and/or 3D target models (i.e. meshes, RGB-D, LIDAR, and range scans).
- **Semantics-driven retrieval approaches.** A lot of semantic information exists in both the 2D query images and the 3D target scenes in our current **SceneIBR19** benchmark. However, we have found again that there is no participating group that has directly utilized it during retrieval. Therefore, in the hope of developing a practical retrieval algorithm which is scalable to the size of the benchmark, we should prioritize this in our future work list.
- **Classification-based retrieval.** Again, we have found that classification/recognition-based 3D model retrieval (i.e. Tran's RNIRAP and Yuan's VMV-VGG) has great potential in achieving better performance.

7. Acknowledgements

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

References

[ARYL*18] ABDUL-RASHID H., YUAN J., LI B., LU Y., BAI S., BAI X., BUI N.-M., DO M. N., DO T.-L., DUONG A.-D., ET AL.: 2d image-based 3d scene retrieval. In *Proceedings of the 11th Eurographics Workshop on 3D Object Retrieval* (2018), Eurographics Association, pp. 37–44. 1, 6

[ARYL*19] ABDUL-RASHID H., YUAN J., LI B., SCHRECK T., LU Y.: SHREC'19 2D Scene Image-Based 3D Scene Retrieval Track Website. http://orca.st.usm.edu/~bli/SceneIBR2019/, 2019. 6

[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, 5, 6, 7

[ARYLL19] ABDUL-RASHID H., YUAN J., LI B., LU Y.: Sketch/image-based 3d scene retrieval: Benchmark, algorithm, evaluation. In *MIPR* (2019), IEEE. 1, 4, 6

[DDS*09] DENG J., DONG W., SOCHER R., LI L.-J., LI K., FEI-FEI L.: ImageNet: A large-scale hierarchical image database. In *CVPR* (2009), pp. 248–255. 2, 5

[Fli18] FLICKR: Flickr. https://www.flickr.com/, 2018. 2

[Goo18] GOOGLE: Google images. https://www.google.com/imghp?hl=EN, 2018. 2

[HZRS15] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. *CoRR abs/1512.03385* (2015). URL: http://arxiv.org/abs/1512.03385, arXiv:1512.03385. 5

[Kal17] KALLIATAKIS G.: Keras-vgg16-places365. https://github.com/GKalliatakis/Keras-VGG16-places365, 2017. 4

[KMRW14] KINGMA D. P., MOHAMED S., REZENDE D. J., WELLING M.: Semi-supervised learning with deep generative models. In *Advances in neural information processing systems* (2014), pp. 3581–3589. 3

[KW13] KINGMA D. P., WELLING M.: Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013). 3

[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. 6

[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*15a] LI B., LU Y., LI C., GODIL A., SCHRECK T., AONO M., BURTSCHER M., CHEN Q., CHOWDHURY N. K., FANG B., ET AL.: A comparison of 3d shape retrieval methods based on a large-scale benchmark supporting multimodal queries. *Computer Vision and Image Understanding 131* (2015), 1–27. 2

[LLL*15b] 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. 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

[Ren18] RENAULT: Renault SYMBOIZ Concept. http://www.renault.co.uk/vehicles/concept-cars/symbioz-concept.html, 2018. 1

[RHGS15] REN S., HE K., GIRSHICK R. B., SUN J.: Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR abs/1506.01497* (2015). URL: http://arxiv.org/abs/1506.01497, arXiv:1506.01497. 6

[SMKLM15] SU H., MAJI S., KALOGERAKIS E., LEARNED-MILLER E.: Multi-view Convolutional Neural Networks for 3D Shape Recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)* (12 2015), vol. 2015 Inter, IEEE, pp. 945–953. doi:10.1109/ICCV.2015.114. 3

[Tip18] TIPS L. T.: Driving a multi-million dollar autonomous car. http://www.youtube.com/watch?v=vlIJfVlu2hM&feature=youtu.be, 2018. 1

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

[XHE*10] XIAO J., HAYS J., EHINGER K. A., OLIVA A., TORRALBA A.: Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2010), IEEE, pp. 3485–3492. 2

[YLL16] YE Y., LI B., LU Y.: 3d sketch-based 3d model retrieval with convolutional neural network. In *2016 23rd International Conference on Pattern Recognition (ICPR)* (2016), IEEE, pp. 2936–2941. 4

[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). 2, 4, 5

[ZLK*18] ZHOU B., LAPEDRIZA A., KHOSLA A., OLIVA A., TORRALBA A.: Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 40*, 6 (2018), 1452–1464. doi:10.1109/TPAMI.2017.2723009. 2, 4