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], Perez Rey[§4], Mike Holenderski[§4], Dmitri Jarnikov[§4], Vlado Menkovski[§4], Ngoc-Minh Bui[§5,6], Trong-Le Do[§5,6], Khac-Tuan Nguyen[§5,6], Tu V. Ninh[§5], Khiem T. Le[§5], Thanh-An Nguyen[§6], Minh-Triet Tran[§5,6], Vinh-Tiep Nguyen[§7],

[1] School of Computing, University of Southern Mississippi, USA
[2] Department of Computer Science, Texas State University, USA
[3] Graz University of Technology, Austria
[4] Eindhoven University of Technology, Netherlands
[5] Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City, Vietnam
[6] Software Engineering Lab, University of Science, Vietnam National University - Ho Chi Minh City Vietnam
[7] University of Information Technology, University of Science, 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 [ARYLL18], we have extended the number of the scene categories from the initial 10 classes in the* ***SceneIBR2018*** *benchmark to 30 classes [Yea19], 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.*

1. Introduction

2D scene image-based 3D scene model retrieval is to retrieve 3D scene models given an input 2D scene image. It has vast related applications, including highly capable autonomous vehicles like the Renault SYMBIOZ [Ren] [Tip], 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**, 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 contributions to further advance 2D scene images-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 will also provide corresponding evaluation code for computing a set of performance metrics similar to those used in the Query-by-Model retrieval technique.

[†] Track organizers
[*] Corresponding author
[§] Track participants.
For any question related to the track, please contact Bo Li.
bo.li@usm.edu.

2. Benchmark

2.1. Overview

Building process. Scene categories were selected from Places [ZLK*17] and SUN [Xea10] with the criteria of selection being popularity. Through a three-person voting mechanism we selected the most popular 30 scene classes (including the initial 10 classes in **SceneIBR2018**) from the Places88 dataset [ZLK*18]. Many of Places88's scene categories are shared by ImageNet [Dea09], SUN [Xea10], and Places [ZLK*17].

Instances for the additional 20 classes, were sourced from Flickr [Fli18] as well as Google Images [Goo18b] for images and downloaded SketchUp 3D [Goo18a] 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 datasets 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 Dataset

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

2.4. Evaluation Method

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

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

3. Participants

Of the six groups who initial 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 rank list results for the four different methods submitted by the three group.

The participants and their runs are listed as follows:

- ***CVAE*** and ***CVAE-VGG*** submitted by Perez Rey, Mike Holenderski, Dmitri Jarnikov and Vlado Menkovski from Eindhoven University of Tecnology in the Netherlands (Section 4)
- ***VMV-VGG*** submitted by Juefei Yuan, Hameed Abdul-Rashid, Bo Li, Yijuan Lu from the University of Southern Mississippi and Texas State University (Section 4.6)
- ***HCMUS*** submitted by Ngoc-Minh Bui, Trong-Le Do, Khac-Tuan Nguyen, Tu V. Ninh, Khiem T. Le, Thanh-An Nguyen,

Code for Evaluation: http://orca.st.usm.edu/~bli/SceneIBR2019/SceneIBR2019_Evaluation_Toolkit.zip

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

Minh-Triet Tran and Vinh-Tiep Nguyen from the Vietnam National University - Ho Chi Minh City (Section 4.7)

4. Conditional Variational Autoencoders for Image Based Scene Retrieval

4.1. Overview

The proposed approach consists of image to image comparison with conditional variational autoencoders (CVAE) [KRMW14]. The CVAE is a semi-supervised method for approximating the underlying generative model that produced 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 renders 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.

4.2. Data Preprocessing

Thirteen renders 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 render 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 augmentation is carried out by performing a horizontal flip to all images. The corresponding data space is $X = [0,1]^{64\times64\times3}$.

4.3. 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 [KRMW14]. 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 [KRMW14,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.

4.4. 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 render probability distributions, see Figure 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.

4.5. Submissions

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.

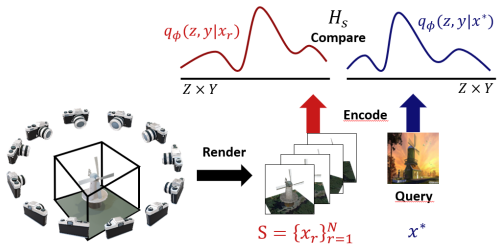

Figure 3: Overview of Scene Sampling and CVAE Distribution Learning

4.6. VMV: View and Majority Vote based 3D scene retrieval algorithm

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

4.6.1. 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 sample along the equator of the sphere while the last view is from a top-down perspective as shown in **Fig.** 5.

4.6.2. Data augmentation

They implemented several augmentations on the dataset to avoid over fitting (e.g rotations, translations and reflections) these augmentations extended the dataset 500 times its initial sizes [YLL16].

4.6.3. 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.

4.6.4. 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.

The classification probability distribution query image and each model's 13 scene-views are target 3D scene's 13 are used to generate a ranklist for each image query by using this majority vote-based label matching method.

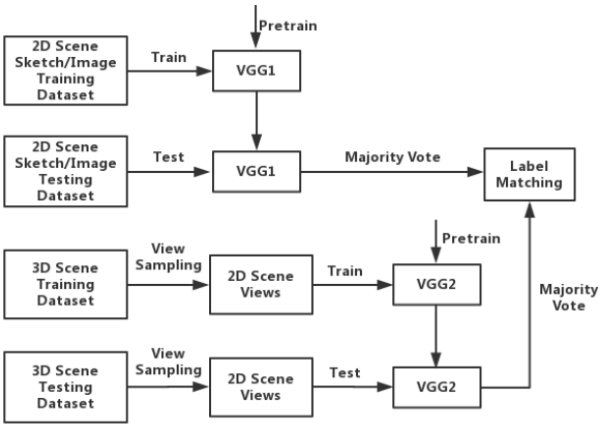

Figure 4: VMV architecture.

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

4.7. HCMUS

To classify an image into one of the 30 scene categories in this track, they apply their method (used in SHREC 2018) to extract scene attributes using MIT Place API. They train a simple network with the extracted features from Place API and use this network to classify an input image with 30 labels.

4.8. 3D scene classification with multiple screenshots, domain adaptation, and concept augmentation

In this track, they perform two-step process for 3D scene classification with multiple screenshots.

In the first step, they train multiple classification models and use the voting scheme to ensemble the classification result. They apply

their proposal of domain adaptation (used in SHREC 2018) to classify a 2D screenshots of a 3D scene. They also try to train simple networks (with one to two hidden layers) to classify a 2D screenshot using features from ResNet50.

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 motivates their proposal to employ object/entity detectors to identify entities related to certain concepts existing a screen shot.

In version 2 of the proposed method, they first collect a dataset of natural images from Internet corresponding to concepts that are related to the 30 scene categories. For example, they use the query terms such as "cactus", "camel", etc to serve the scene classification for "desert". They train their set of object detectors from this dataset of natural images with Faster RCNN. Then they apply their detectors to identify entities that might appear in a scene, such as "book" (in a library), "umbrella" (in a beach), etc. By this way, they further refine their retrieval results.

5. Results

A comparative evaluation of the is performed on all methods. Retrieval performance measurements are based on the seven metrics mentioned in Section 2.4: PR, NN, FT, ST, E, DCG and AP. **Fig.** 6 and **Table** 2 compare the four learning-based participating methods on the testing dataset.

6. Conclusions and Future Work

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.

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.

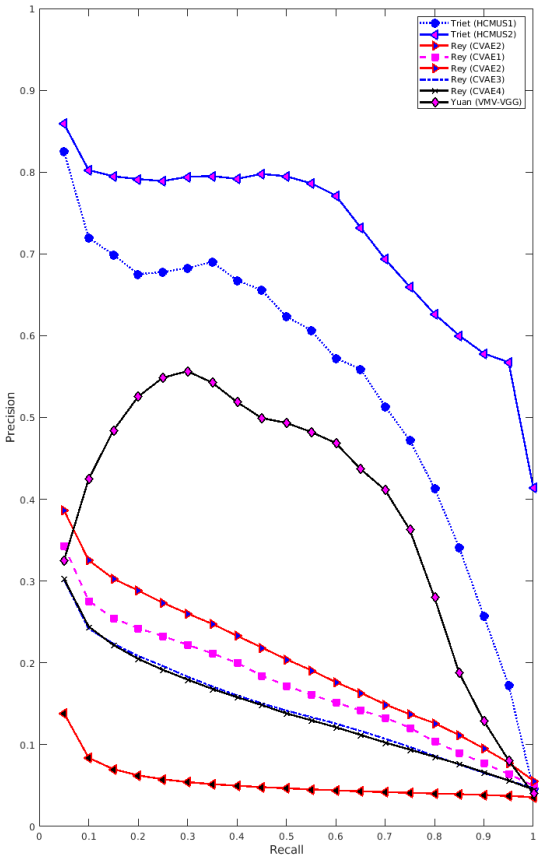

Figure 6: Precision-Recall diagram performance comparisons on testing dataset of of our **SceneIBR19** benchmark for the four learning-based participating methods.

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

[Dea09] DENG J., ET AL: ImageNet: A large-scale hierarchical image database. In *CVPR* (2009), pp. 248–255. 2

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

[Goo18a] GOOGLE: 3D Warehouse. http://3dwarehouse.sketchup.com/?hl=en, 2018. 2

[Goo18b] GOOGLE: Google images. https://www.google.com/imghp?hl=EN, 2018. 2

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

[KRMW14] KINGMA D. P., REZENDE D. J., MOHAMED S., WELLING M.: Semi-Supervised Learning with Deep Generative Models. 1–9. 3

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

[LLL*15] 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

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

[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)*

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

Participant	Method	NN	FT	ST	E	DCG	AP
Complete benchmark							
Rey	CVAE-VGG	0.071	0.054	0.099	0.055	0.405	0.0535
Rey	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
Triet	HCMUS1	0.845	0.620	0.674	0.618	0.791	0.5436
	HCMUS2	**0.865**	**0.749**	**0.792**	**0.745**	**0.863**	**0.7221**
Yuan	VMV-VGG	0.122	0.458	0.573	0.452	0.644	0.3899

(12 2015), vol. 2015 Inter, IEEE, pp. 945–953. doi:10.1109/ICCV.2015.114. 3

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

[Xea10] XIAO J., ET AL: SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR* (2010), IEEE Computer Society, pp. 3485–3492. 2

[Yea19] YUAN J., ET AL: Sketch/image-based 3d scene retrieval: Benchmark, algorithm, evaluation. In *MIPR* (2019), IEEE. 1

[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

[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