Multimedia Tools and Applications (2021) 80:19491–19511
https://doi.org/10.1007/s11042-021-10615-7

A survey of recent 3D scene analysis and processing methods

Juefei Yuan[1] · **Hameed Abdul-Rashid**[1] · **Bo Li**[1]

Received: 20 January 2020 / Revised: 2 December 2020 / Accepted: 25 January 2021 /
Published online: 27 February 2021

Abstract
With ubiquitous cameras and popular 3D scanning and capturing devices to help us capture 2D/3D scene data, there are many scene understanding related applications, as well as quite a few important and interesting research problems in processing, analyzing, and understanding the available scene data. During the recent several years, there is a significant advancement in different research directions in this field and quite a few novel 3D scene analysis and processing methods have been proposed correspondingly in each direction. This paper provides a review and critical evaluation on the most recent (i.e., within five recent years) and novel data-driven or semantics-driven 3D scene analysis and processing methods, as well as several involved 3D scene datasets. For each method, its advantage(s) and disadvantage(s) are discussed, after an overview and/or analysis of the approach. Finally, based on the review, we propose several promising future research directions in this field.

Keywords 3D Scenes · Survey · Scene analysis · Scene processing · Semantics-driven approaches · Data-driven approaches

1 Introduction

Nowadays, more and more different types of 3D sensing devices could help us capture 3D scene data, such as Acuity Laser [49], Light Detection and Ranging (LIDAR) [68], and Leap Motion [67]. Those captured 3D scenes include not only indoor scenes, but also outdoor scenes. In addition, in order to deal with different situations or meet different research requirements, researchers have built different benchmarks [14, 26, 56, 58, 80]. Cordts et al. [14], Straub et al. [56], and Vasiljevic et al. [58] are built for 3D indoor and/or outdoor scene

✉ Bo Li
bo.li@usm.edu

Juefei Yuan
juefei.yuan@usm.edu

Hameed Abdul-Rashid
hameedabdulrashid@gmail.com

[1] School of Computing Sciences and Computer Engineering, University of Southern Mississippi, Long Beach, MS, USA

research, while Zheng et al. [80] and Gupta et al. [26] are examples that are more accurate and more comprehensive than previous benchmarks.

In 2018 and 2019, we have successfully organized four sketch/image-base3d 3D scene shape retrieval contest (SHREC) tracks [1, 2, 73, 75, 76] which have attracted many interest from researchers who regarded the topic of 3D scene retrieval as an important and promising research direction. For these four tracks, we have built two benchmarks which have 10 and 30 scene categories respectively, while each category has 100 3D scene models. According to the track reports [75], we have found that usually a 3D scene retrieval algorithm will employ the following 3D scene processing and analysis techniques: classification, recognition, reconstruction/generation, view sampling, and semantics learning. This motivates us to conduct a survey on related, among others, 3D scene processing and analysis methods to advance the 3D scene related research, especially 3D scene retrieval.

Compared to 3D objects, 3D scenes are more directly related to our daily life. There are a large amount of real life relevant application scenarios, such as autonomous driving cars, 3D geometry video retrieval, and 3D AR/VR Entertainment. Therefore, recently researchers have proposed many 3D scene analysis and processing methods and significantly contributed to this research area. 3D scene retrieval is one part of this area. The research directions within this area include: **a) 3D scene classification**, which is to classify 3D scene models into different certain categories based on a training dataset containing labeled examples from known categories; **b) 3D scene recognition**, which is to recognize the category of a given 3D scene with/without a training dataset; **c) 3D scene retrieval**, which is to retrieve 3D scene models given an input query (i.e., a 2D scene sketch/image) provided by the users; **d) 3D scene reconstruction**, which is to reconstruct three-dimensional scene models from multiple 2D projected scene images, whose depth information may be missing; **e) 3D scene generation**, which is to generate 3D scene models from 2D images or nature languages (e.g., "a person besides a table").

These research directions may involve either data-driven or semantics-driven based techniques: a) **Data-driven methods** are the methods that are based on the original raw data or the data preprocessed by some techniques like redundant data points reduction, error data removal [25], and GPU parallel calculating. b) **Semantics-driven methods** are the techniques that are not only based on the data, but also incorporate semantic information of the objects or the context in the 3D scenes. For examples, Rangel et al. [44] proposed a 3D scene classification method based on semantic labels extracted from 3D scenes. Akase et al. [6] presented a web-based 3D room layout generation system, which utilizes the semantic information of each furniture in a 3D room and each furniture's related objects. In addition, to improve reconstruction accuracy, Vineet et al. [59] reconstructed a 3D scene by fusing the 3D map with the semantic information of each objects in the scene, etc.

Section 2 provides an overview by defining several typical related terminologies, and summarizing the papers to be reviewed in the survey. Sections 3 ∼7 introduce and review each direction individually. Finally, after a conclusion, several promising future work directions are proposed in Section 8.

2 Overview

In this section, we first provide a definition for the most commonly used terminologies in 3D scene analysis and processing techniques.

3D scene In computer world, we define a 3D scene as an arrangement of scenery objects and properties to represent a recognizable place, where the objects that appear, and their

shapes, sizes, and spatial relationships, as well as the background (i.e., ground, and sky) are important features to characterize the place.

3D scene shape representations 3D scene contains a list of objects, which are entitled independent representations to represent their shapes and textures. To represent and easily maintain the semantic relationship between the objects in a scene, a scene graph data structure is often used. People have developed quite a few 3D object representations to meet the needs of practical applications, for example, (1) meshes; (2) point sets; (3) Spline surfaces; (4) Volumetric representations (i.e., voxels, particle systems, and finite element method (FEM); (5) Subdivision surfaces (i.e., Loop subdivision surface [36]); (6) Constructive solid geometry (CSG) (a shape defined based on boolean operations on simple objects); and (7) Implicit surfaces (a surface defined by a mathematical equation).

Besides the above representations, RGB-D is a popular 3D scene representation to represent 3D scenes captured by various 3D capturing and sensing devices.

3D scene features We can divide 3D scene features into low-level 3D scene features and high-level 3D scene features. **Low-level 3D scene features:** characterize a 3D scene at a lower level, e.g., pixel-level, by focusing on details like colors, textures, shapes (e.g., lines, dots), and spatial locations. **High-level 3D scene features:** represent a scene at a higher level, e.g., object-level or object-group level, by examining the spatial and semantic relationships between the objects in the scene.

3D scene semantics information Semantics information is used to interpret a special entity. There are a lot of semantic information (i.e. objects, object parts and object groups) existing in 3D scene models. To improve 3D scene analysis and processing accuracy, we could incorporate such semantic information into the learning process.

3D scene datasets A 3D scene dataset is a collection of 3D scene data spanning over different categories, and often contains both training and testing subsets. Different 3D scene datasets are built for different purpose, e.g., Cordts et al. [14] released a Cityscapes dataset for urban street 3D scene analysis, while Vasiljevic et al. [58] curated a Dense Indoor and Outdoor DEpth (DIODE) dataset for both indoor and outdoor 3D scene analysis.

In this paper, we review very recently (i.e., within five recent years) published thirty-five (35) papers related to the five research directions (3D scene classification/recognition/retrieval/reconstruction/generation). We further group them based on two different inputs (2D and 3D), as well as two types of approaches (data-driven and semantics-driven). Table 1 gives the overview of the above information. In the following five sections, we will review each of the five research directions individually.

Table 1 Overview of the thirty-five (35) 3D scene analysis and processing research papers reviewed in this paper w.r.t different research directions, inputs, and approaches

Tasks	Input (2D)	Input (3D)	Data-driven	Semantics-driven
classification	[37, 39, 49, 50, 58]	[11, 31]	[31, 37, 49, 50, 58]	[11, 39]
recognition	[5, 67]	[36, 44, 65, 69]	[5, 44, 69]	[36, 65], [67]
retrieval	[43]	[63](methods 1∼3), [64]	[43, 63](methods 1∼2), [64]	[63](method 3)
reconstruction		[7, 14, 17, 21]	[7, 14, 17]	[34, 42], [54]
		[22, 34, 38, 40, 41]	[21, 22, 38, 40]	
		[42, 46, 54, 62]	[41, 46, 62]	
generation	[4, 30, 33]	[20, 57, 66]	[66]	[4, 20, 30, 33, 57]

3 3D scene classification

Given a 3D scene model, 3D scene classification is to classify this scene model into one of the candidate categories.

3.1 Data-driven 3D scene classification

A variety of data-driven based methods have been proposed and many of them work well under certain circumstances.

Steinhauser et al. [55] proposed a scene classification method based on the data collected from a LIDAR laser scanner. It can be used to collect and classify the raw data of the real time surrounding environment of the vehicle into safe condition for driving road and unsafe obstacles (static obstacles or moving obstacles). They tested the method on the university campus and forest tracks, and the approach generates good results in estimating safe road (e.g., untarred road). However, it still has some places to be improved: (a) need to reduce the time cost of the method so as to run in real time, (b) make the method be able to deal with the LIDAR failure issues (e.g., a scanner may fail in a small degree range, like 10 degrees), (c) the algorithm may not work well if there are quite a few moving cars around the vehicle, or when only a small number of landmarks are visible or the trees beside the road are dense and hard to distinguish them from each other.

Ramezani and Ebrahimnezhad [43] presented a geometric features-based algorithm for 3D model classification. They extracted geometric features from the faces and vertices of a 3D model and utilized a histogram of the features for classification. The histogram comprises two sets of features for each vertex: (1) the deviation angle of the vertex's normal vector from the center-to-vertex vector [43]; (2) the distance between the vertex and the model's center. They also adopted mutual Euclidean distance histogram to improve the classification accuracy, and compared their classification accuracy and efficiency with respect to two different classifiers which are Probabilistic Neural Network (PNN) and Support Vector Machine (SVM).

Lin et al. [34] proposed a method for indoor scene understanding based on RGB-D data. They utilized the Constrained Parametric Min-Cuts (CPMC) [11] framework to generate candidate cuboids for the 3D objects in a 3D scene, and then classify these cuboids. With 2D segmentation information, 3D geometry properties, and the contexture relationship between objects and scenes integrated in this method, the 3D object and 3D scene classification can be solved together. Compared to the part-based model DPM [22], their method achieved a good performance improvement on the NYU v2 dataset [51]: the F1-score accuracy, which is the harmonic mean value of the precision and recall [18], has been increased considerably.

Wang et al. [63] proposed two contributions to solve the two issues existing in scene recognition/classification: (a) large intra-class variations; (b) label ambiguity. Firstly, they proposed a multi-resolution CNN architecture, which consists two parts: (a) coarse-resolution CNNs, which deal with global features and large objects in the scene; (b) fine-resolution CNNs, which deal with local features and small objects in the scene. They are complementary to each other. Secondly, for the label ambiguity issue, they adopted two ways to deal with it: (i) utilizing a confusion matrix technique (by computing the similarity between any two categories), which can merge those ambiguous scene categories into one super category (e.g., outdoor athletic, and outdoor track scenes); (ii) using other networks to predict the label of each scene, which is called soft label. Then, train the model with the guidance of super category labels and soft labels. However, there still exist some failure

examples: some scene categories still cannot be easily distinguished with each other, e.g., supermarkets and shops are similar if looked from outside.

Aiger et al. [4] proposed a multi-view based CNN model, which has a good accuracy in classifying water and trees. Compared with the state-of-the-art model Inception-V3 [57], the related accuracy has been increased from 79% to 96%. The method requires neither fully segmented labels, nor marked object class boundaries in a scene image, while it only requires sparsely labeled pixels.

Muller-Budack et al. [41] treated the geolocalization (subdividing the earth into multiple geographical cells) of a photo as a scene classification problem. To incorporate the hierarchical knowledge of different spatial resolutions, they adopted a multi-partition CNN model, which can be used to compute geolocalization loss. Moreover, they extracted the scene label information from different scene types (indoor, nature, urban, etc) by using the ResNet model [29], and incorporated the information into the multi-partition CNN model as well. They ran their method on two benchmarks Im2GPS [28] and Im2GPS3k [61], and compared with the PlaNet [64] approach and demonstrated that their method has improved the classification accuracy. This CNN model requires a small number of training images and does not rely on the retrieval results from any dataset for verification. To further improve the geolocalization, they could also incorporate other contextual information into the CNN model, such as specific landmarks, and image styles.

3.2 Semantics-driven 3D scene classification

Unlike data-driven 3D scene classification that only focuses on the scene data itself, semantics-driven 3D scene classification also considers the semantic relatedness between objects, or between objects and scenes.

Since it is challenging for robotics to achieve a high accuracy in 3D indoor scene classification due to a large number of scene categories in related datasets, Chen et al. [13] proposed a word vector (a.k.a word embedding) based algorithm for the 3D indoor scene classification task. This algorithm first uses GPS to locate a robot's rough area, e.g., a school, or a shopping mall. Then it just needs to search the objects belonging to this area instead of searching all the object categories. They employed different CNN models for different purposes in their approach, which consists of four modules. The first is a typical CNN-based scene classification module to obtain the top-5 prediction labels. The second is a CNN-based scene parsing module which is to detect the objects, background and foreground in a scene. Next, the third module word embedding is to compute the vector for the objects in a scene image and the vector for the top-5 prediction labels. Finally, the fourth module refines the rank list of the top-5 labels based on the comparison of the above two vectors. They adopted ResNet50 as the CNN model. After incorporating the word vector information into the CNN model, they further increased its classification accuracies on both the Places365 dataset and their selected indoor scenes dataset, which is composed of the school, home and shopping mall scenes selected from the original Places365 dataset.

Rangel et al. [44] proposed a scene classifier based on the semantic labels recognized by the Clarifai [54] descriptor. This paper compares the Clarifai-based approach with other descriptors (i.e., GIST, ESF), and shows that the Clarifai-based descriptor is competitive if compared with those state-of-the-art ones. Moreover, the Clarifai-based approach performs the best when dealing with general scenes. For example, after this approach is trained on the semantic sequences of one type of building scenes, it can obtain good classification results on the semantic sequences of another type of building scenes.

4 3D scene recognition

Similar to 3D scene classification, 3D scene recognition can be categorized into data-driven and semantics-driven approaches.

4.1 Data-driven 3D scene recognition

Behl et al. [7] proposed a new system for estimating 3D scene traffic flow for autonomous driving. This system addresses the large displacement or local ambiguity (due to lack of texture or surface reflection) problems which can fail the estimation in existing methods. It is a recognition-based approach instead of like existing ones relying on local features. They conducted experiments on 2D bounding boxes calculation, 2D instance segmentations, and 3D object part predictions. The results demonstrated that the approach improves the performance by a lot when dealing with large displacement or local ambiguities.

Zhong et al. [81] proposed a method for 3D text recognition in 3D scenes. It helps in shadow detection and removal. This method segments shadow pixels from background and text pixels by utilizing the Gabor kernel, then removes their depth information, and finally converts the 3D texts into a 2D text image. Since it is the first attempt in 3D text recognition, there are still some room for improvement. For example, the thresholds determined by the Gabor kernel for shadow detection cannot achieve good performance where there are low contrast, small fonts, non-uniform illumination effects, and so on.

Shi et al. [50] proposed a variational denoising recursive autoencoder (VDRAE) system to predict the 3D scene layout of a 3D point cloud indoor scene, as demonstrated in Fig. 1. This system generates and denoises the predicted 3D object proposals by incorporating the hierarchical context information of 3D objects. The denoised indoor scenes can improve the 3D scene recognition accuracy. However, this system is not an end-to-end system. For example, the hierarchical proposals prediction and denoising steps are done separately.

4.2 Semantics-driven 3D scene recognition

Zhao et al. [79] proposed a framework that can parse scene images at both pixel level and word concept level. They jointly embedded them into a high-dimensional positive vector space, as demonstrated in Fig. 2. At the word concept level, their framework incorporates the semantic word-word relations, i.e., using a hypernym/hyponym based on WordNet [21]. They made rules for the space construction process: making the pixel level features close to their annotated labels and keeping the semantic relations unchanged. In general, their framework includes two streams: (a) **Concept stream**, which is to incorporate the semantic relationship information into the embedding space; (b) **Image stream**, which is to segment the image by using a fully convolutional network. Then, their framework combines the two streams by a joint loss function to measure the similarity in their image features and word concept hierarchies, while the weights of the two streams in the loss function are predefined. They selected 150 object categories from the ADE20K dataset [83] to train and test their framework based on certain evaluation measures, e.g., using weighted intersection-over-union (IoU) [66] as a baseline flat metric. They also compared their jointly embedding framework with other models, such as Word2Vec [39]. The results show that their framework has achieved better performance and demonstrated two main advantages: (a) It has more freedom for the user to label an object at different grained levels (e.g., Husky and dog categories) without sacrificing the training accuracy. (b) The system is end-to-end, thus the semantic relationship information can be extended easily in the system. Nevertheless, it

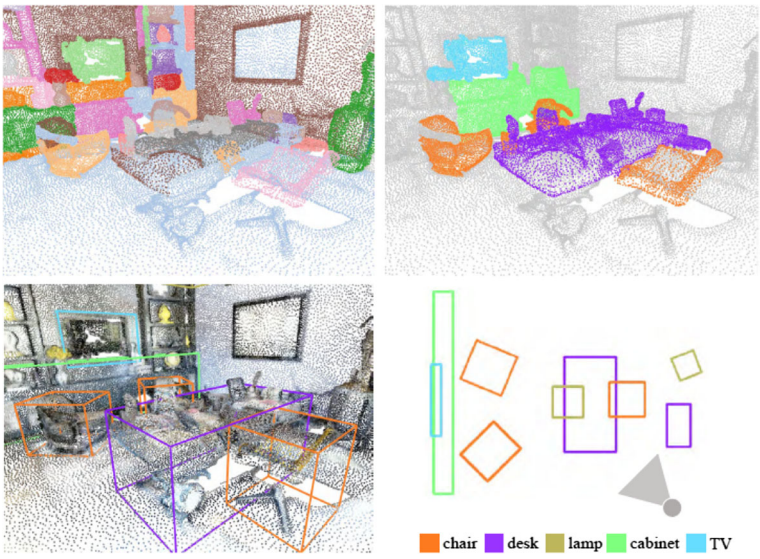

Fig. 1 A VDRAE-based 3D object layout prediction example. Segmented 3D point cloud as input (top left), processed by VDRAE system (top right, make the objects in the same category have the same color), and make fully objects contained in 3D bounding boxes (bottom) [50]

also has some limitations that may affect its performance, such as: (a) the training data and target data are very different from each other; (b) compared to the label set, the size of the image dataset is too small.

Miksik et al. [40] presented an augmented reality system for 3D outdoor scene recognition and reconstruction. This system simulates the 3D outdoor scene map in real-time and allows users to segment objects manually. With a machine learning model learned from the existing 3D object/scene datasets and objects drawn by the users, the system can recognize the scene in a more accurate way. The limitations of this system are in three-fold: (a) the computational load is heavy; (b) it needs powerful GPUs, which limits the laptop usage for outdoor scenes; and (c) the learning and prediction processes require users' voice commands to switch, and these two functions cannot be used at the same time, while in the mean time the feedback of the two processes could amplify the errors and decrease the accuracy.

Yuan et al. [77] proposed a semantic tree-based framework for 3D scene model recognition. Firstly, this framework builds a scene semantic tree based on the semantic ontology in WordNet [21]. Secondly, the framework can identify the semantic attributes (e.g., object labels contained in the scenes) that the 2D query image contains via a deep learning-based recognition approach. And Finally, by measuring the semantic similarity between the 2D image's semantic attributes and the nodes in the semantic tree, the framework could recognize the target 3D scene categories. In this framework, the scene semantics of a particular scene category contain three probability distributions: (a) object occurrence probability, it is the conditional probability that an object class appears in the scene category, (b) object co-occurrence probability, it is the conditional probability that both of two object classes appear simultaneously in the scene category, and (c) spatial relation probability, it is the conditional probability that two object classes have a certain spatial relation in the scene category.

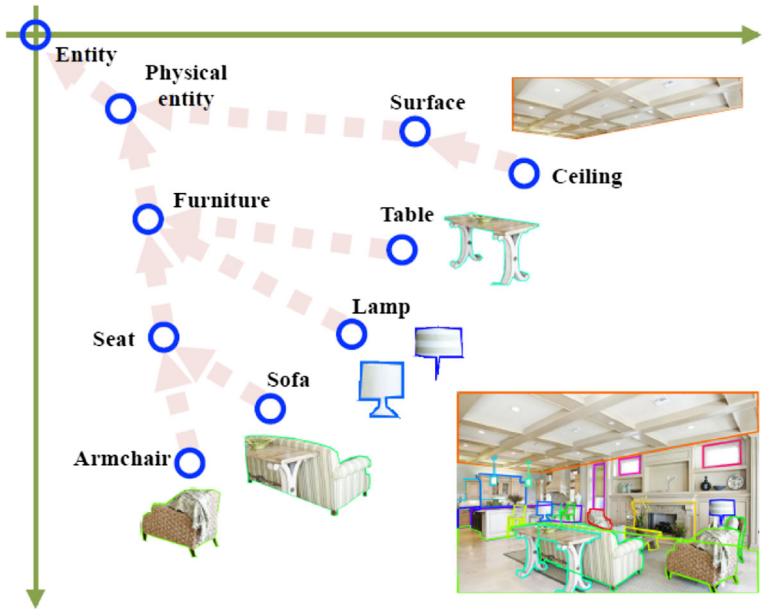

Fig. 2 Embed image pixel features and word concepts jointly [79]

5 3D scene retrieval

3D scene retrieval is to retrieve 3D scene models given an input query provided by the users. This research topic has vast applications such as 3D scene reconstruction, 3D geometry video retrieval, and 3D AR/VR entertainment.

5.1 Data-driven 3D scene retrieval

Savva et al. [48] advised a system to design and help retrieve 3D indoor scenes. This system is based on a large-scale learned 3D priors set which is extracted from existing 3D scenes. These priors are related to static support, position, and orientation. Moreover, by using those priors, this system provides suggestions for 3D object placement and assembles 3D objects with regard to desired scene category. However, this system does not consider collision detection between two objects, which may lead to incorrect placement.

Yuan et al. [74] proposed a sketch/image-based 3D scene retrieval algorithm. The input query of the approach is a user's hand-drawn 2D scene sketch or a 2D scene image. This method represents a 3D scene model by multiple 2D view images sampled from different viewpoints. Then, they train two CNN models separately on the 2D scene sketches/images, and the scene view images, as shown in Fig. 3. Finally, the ranking is based on the two CNN classification results on the corresponding testing datasets.

Li et al, one of the participant groups in two Shape Retrieval Contest 2019 (SHREC'19) tracks on 3D scene retrieval tracks [2, 73], presented the Maximum Mean Discrepancy domain adaption method based on the VGG model (MMD-VGG) to tackle 3D scene retrieval task. The query is a 2D scene sketch/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.

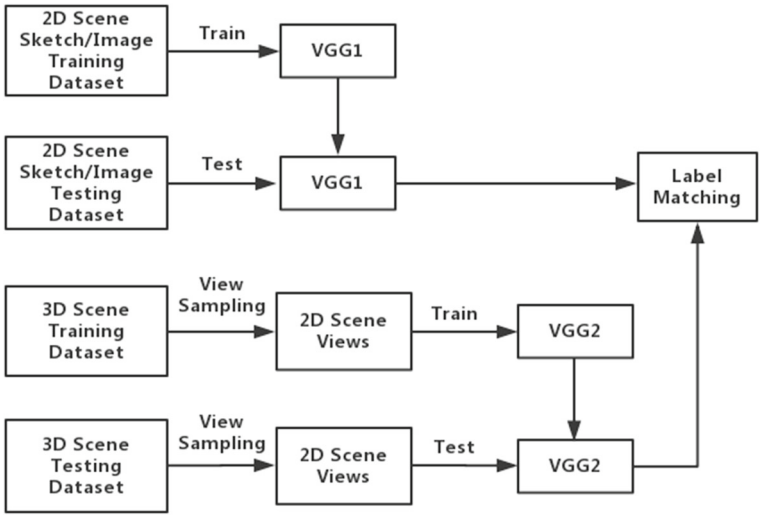

Fig. 3 VBV-VGG architecture [74]

Liu et al, one more participant group in the two SHREC tracks, proposed a two-stream CNN-based method. In their method, the 2D scene sketch/images dataset is regarded as the source domain, and the 3D scene models dataset is regarded as the target domain. It processes samples from either domain with a corresponding CNN stream. They adopted triplet center loss [30] and softmax loss for training supervision, the network is trained to learn a unified feature embedding for each sample, which is then used for similarity measurement for the retrieval process.

5.2 Semantics-driven 3D scene retrieval

Minh-Triet Tran et al, another participant group in the two SHREC'19 tracks on 3D scene retrieval, proposed a domain adaptation based method named ResNet50-Based Sketch Recognition and Adapting Place Classification for 3D Models Using Adversarial Training (RNSRAP). In addition to the training dataset provided, they performed data augmentation by adding semantic related sketches/images (e.g., add camel, cactus sketches/images to the desert category). Due to the substantial variance exists in the two domains (source domain and target domain), the adversarial adaptive method they utilized is to minimize the variance between the source and target domains. As a result, the trained domain adaptation model can be used for classification in both the source and target domains.

6 3D scene reconstruction

Similar to 3D object reconstruction, 3D scene reconstruction is to reconstruct a three-dimensional scene model from multiple 2D projected scene view images, whose depth information needs to be recovered.

6.1 Data-driven 3D scene reconstruction

Ebrahimnezhad and Ghassemian [20] proposed a space curve-based method to reconstruct a moving 3D object from stereo rigs which capture image sequences. The space curves are extracted from the stereo images. This method ensures accurate geometry, and minimizes the number of outliers. In addition, photometric information is not required after adopting the new space curve extraction method. Last but not least, by utilizing perpendicular stereo, the method can estimate the motion of the 3D object more accurately. Based on the estimated motion, they construct multiple virtual cameras to obtain multiple views and extract the finest visual hull of the 3D object, which is useful for reconstructing poorly-textured objects.

Song et al. [52] presented an end-to-end system named semantic scene completion network (SSCNet), which is based on convolutional neural network techniques. It is able to reconstruct a 3D indoor scene by using 3D voxel representations and predict semantic segmentation labels with a 2D depth image as input. This system takes both 3D scene reconstruction and semantic labels into consideration simultaneously, which were handled individually in previous work. This system solves two issues: a) extend the receptive field of the network to effectively capture 3D volume data context information; b) manually build a 3D scene dataset named SUNCG, which provides complete labeled 3D objects information.

Bobenrieth et al. [9] proposed a 3D indoor scene reconstruction method. Due to the reason that some applications require a complete scanning data captured by some scanning devices like Kinect, which is a time-consuming process, their method only requires a few shots of the 3D scene, and also no overlapping requirement is required to generate a seamless scene. This method aligns these shots by looking for a group of transformations, and constructs an alignment graph which is used to find a global solution for all the transformations. However, since their method searches all the possible solutions, the time cost is highly dependent on the provided number of shots, e.g., it only takes a few seconds for simple cases, but the time may increase rapidly if the provided number of shots increases sharply.

Penner and Zhang [42] proposed a method to perform soft (keeping depth uncertainty) 3D scene reconstruction and view synthesis. During each stage, their method keeps the depth uncertainty, which can help to refine the depth estimates of object boundaries during the 3D reconstruction step. It also helps to adjust view rays and texture mapping rays during the view synthesis step. Their approach accepts a variety of inputs, which include not only structured images and wide-baseline captures, but also unstructured images and narrow-baseline captures.

Dai et al. [16] presented a data-driven based system named ScanComplete, which can reconstruct a high-resolution 3D scene from an incomplete RGB-D 3D scene scan. This system utilizes fully-convolutional neural network techniques to train on small subvolumes of the 3D scene and test on either small or large 3D scenes. In addition, in order to obtain high-resolution outputs with regard to the 3D scene size, the system adopts a coarse-to-fine strategy to predict small details and global structure simultaneously. The results show that it improves the quality of the 3D scene reconstruction with incomplete RGB-D 3D scan input as well as the semantic segmentation performance when compared with other methods.

Xu et al. [72] presented a system of reconstructing unknown 3D indoor scenes automatically with a single robot. This system enables the robot to scan and reconstruct the scene simultaneously, while taking care of both exploration efficiency and high quality scans. The system utilizes a time-varing 2D tensor field, a 2D image computed over the partial scanned

scene, to guide the movement and camera control of the robot along its movement path. The system flexibly guides the camera's movement instead of using a fixed camera.

In 2018, Guo and Guo [25] presented a method to improve the reconstruction of urban scenes with buildings based on multi-view images. This method fuses the reconstructed dense points and line segments. According to the fusion process, it helps remove error line segments, sample the correct line segments with points, and finally determine and fuse the corrected line segments and points for the 3D scene. The results show that the approach provides more accurate edge information in some parts with rare point features to represent the 3D urban scenes, such as windows and walls, and the time cost is acceptable.

Ritchie et al. [46] presented a 3D indoor scene synthesis model that can solve the following three limitations existing in previous work: (1) cannot place reasonable objects in the scene; (2) fail to take the size of an object into consideration; and (3) time-consuming. This model is a deep convolutional generative model, which can generate data distributions. It utilizes a top-down scene image, extracted from a 3D scene and fed into the model to iteratively synthesize new objects into the 3D scene. The synthesis process involves the decisions of objects' categories, positions, orientations, and sizes.

Rematas et al. [45] developed an end-to-end system to reconstruct a 3D soccer field with moving players from a soccer game video. It can detect the players in the video and estimate the depth map for each player. Compared to other methods that need to set up many synchronized cameras in a real soccer field, this system can reduce the cost. However, this system also has some limitations, e.g., if the system fails to detect the player(s), the player(s) will not be presented in the reconstruction result, and the overlap between the players may cause incorrect depth estimation, etc.

In 2019, Dong et al. [19] presented an end-to-end system that allows multiple robots to collaboratively scan unknown 3D indoor scenes for 3D scene reconstruction. This system utilizes an approach, named Optimal Mass Transport (OMT), to solve the resource distribution problem for the robots scanning the 3D indoor scenes. It adopts a divide-and-conquer scheme to assign tasks to the robots and optimize their paths. The timing and statistics performance information can be found in Fig. 4. However, this system is greedy-based, thus, may fall into local minimum.

Scene	Area	#R	#I	PT	IT	TT	TD
SunCG#1	110 m^2	3	16	0.9 sec	22 sec	6 min	48 m
Matterport3D#1	125 m^2	6	17	2.2 sec	26 sec	8 min	55 m
Office	60 m^2	3	10	0.8 sec	17 sec	3 min	32 m
Sitting_room	85 m^2	4	9	1.1 sec	25 sec	4 min	21 m
Classroom	120 m^2	5	5	1.2 sec	40 sec	5 min	41 m
Meeting_room	80 m^2	3	4	0.8 sec	46 sec	4 min	18 m
Dorm	35 m^2	2	4	0.5 sec	40 sec	3 min	17 m
Lab	300 m^2	6	40	2.3 sec	6 sec	5 min	156 m

Fig. 4 Performance on both synthetic (rows 2 $\sim$ 3)and real scenes (rows 4 $\sim$ 9). Each rows contains the scene area, # of robots (#R), # of planning intervals (#I), planning time for each planning interval (PT), time of each planning interval (IT), total scanning time (TT), and all robots' total movement distance (TD) [19]

Flynn et al. [24] proposed a 3D scene view synthesis method, which first generates a multiplane image (MPI), a kind of representation that can model the exterior effects of light fields such as transparency, and then uses it for view synthesis, based on a sparse set of views of the 3D scene. This method utilizes and improves the learned gradient descent-based method (LGD) [3], which is to update the prediction model parameters, by replacing its update rule with a deep neural network parameters update. This method can deal with depth complexity, object boundaries, light reflections, and thin structures as well, and the results demonstrate the state-of-the-art performance.

6.2 Semantics-driven 3D scene reconstruction

Vineet et al. [59] proposed an end-to-end 3D scene reconstruction system. This system can efficiently perform dense, large-scale semantic 3D scene reconstruction. This system can also deal with moving objects in the 3D scenes by fusing the semantic information of the objects with the 3D map. The core of the system is that they adopt a hash-based fusion approach and a volumetric mean-field (a technique that can gradually refine the edges of each voxel in iterations [60]) based optimization approach for 3D scene reconstruction and object labeling separately.

To improve the 3D scene reconstruction accuracy, Blaha et al. [8] presented a 3D scene reconstruction method that takes both 3D scene reconstruction and semantic labeling into consideration at the same time, because these two themes can affect each other. This method is adaptive, which means it only reconstructs the necessary regions (near to the predicted surfaces) of the 3D scenes. This can save much memory and time, and as a result, it can reconstruct large-scale scenes.

Savinov et al. [47] proposed an approach for dense semantic 3D reconstruction. It utilizes two schemes: one is continuous regularization and the other one is ray potentials. While ray potentials means that a ray is composed of voxels, and its information is contained in the pixels of the observed images. Therefore, by using correct ray potentials, it can achieve more accurate reconstructions. While, continuous regularization is performed to handle the noise in the input data. Particularly, this approach can also reconstruct thin objects due to the accurate representation of the input data, which is optimized by continuous regularization on the surfaces.

Ma et al. [38] presented a hybrid framework to reconstruct semantic 3D dense models from monocular images. This framework utilizes the conditional random fields (CRFs)-based [32] method as the baseline method. It considers the correlation between 3D space points and image pixels, which helps to obtain consistent object segmentation from multi-view images. With those semantic information from images, it can remove the noisy points in the 3D space, correct wrongly-labeled voxels, and fill the space where points are difficult to recover during the reconstruction process.

7 3D scene generation

3D scene generation is to generate 3D scene models guided by the purpose of the generation method. Since many professionals such as autonomous vehicle designers, game developers, VR/AR engineers, and architects are increasingly using virtual 3D scenes for prototyping as well as end products design, the demand for related 3D scene data is high, which triggers the need of 3D scene generation.

7.1 Data-driven 3D scene generation

Zhang et al. [78] proposed a 3D indoor scene modification framework to help users to enrich 3D indoor scenes with many small objects with regard to three scheduling rules related to: (a) object category, (b) object placement, and (c) object arrangement. The modification process could make the scenes more realistic based on the users' preferences. It adopts a cost function that integrates both the constraints proposed by the framework and the user-specified scheduling rules. However, it fails to involve the occurrence information of small objects. For example, the laptop and mouse objects normally appear in the same place, while the framework may separate them far away from each other.

7.2 Semantics-driven 3D scene generation

Akase and Okada [6] proposed a web-based system that deals with 3D room layout according to users' preferences. Their method is based on the Interactive Evolutionary Computation (IEC) method [5], and they used a predictive approach to narrow down the search space and adopted a multi-screen interface to reduce the fatigue of each user by using IEC. They created a semantic database which hosts the information of each single furniture object. This helps us to know each furniture's related objects. In addition, by computing the feature elements' importance, they performed a conjoint analysis on user preferences to generate satisfactory 3D scenes, and it achieved high user satisfaction.

Fisher et al. [23] proposed an activity-centric scene generation technique. It first anchors an observed 3D scene, then scans the activities supported by the 3D scene environment. Based on those activities, they finally determined semantically reasonable arrangements of the retrieved objects from an object database. The limitation of this technique is that it is expected to support a more general class of activities.

Walczak and Flotynski [62] presented a scene generation method by first creating semantically described 3D meta-scenes (3D content representations), and then generating customized 3D scenes based on those 3D meta-scenes. The advantages of using the semantics include: (1) making 3D scene customization simple; (2) supporting high-level abstraction operation and complex content customization.

Ma et al. [37] proposed a sub-scene level framework that can generate 3D indoor scenes by using natural language commands. It contains two steps: (1) retrieve related sub-scene(s) from a 3D scene database; (2) synthesize a new 3D scene by using the sub-scenes and the current 3D environment. To bridge the gap between user language commands and scene modeling operations, they adopted a representation named Semantic Scene Graph (SSG), which contains objects' information, attributes and relationships to encode geometric and semantic scene information. To demonstrate the scalability of their framework, they need to train their model on a larger set of group relations and natural language commands.

Li et al. [33] presented a non-convolutional generative recursive neural network (RvNN) which also focuses on indoor 3D scenes. This network can learn hierarchical scene structures by utilizing a variational autoencoder (VAE) [10]. Figure 5 shows an example scene hierarchy. Besides the semantic object-object relations, they also proposed three grouping operations (support, surround, and co-occurrence), and utilized object co-occurrences during the generation process. However, global scene hierarchies have some limitations due to certain reasons like imperfect training data, and unsatisfactory performance on complicated scenes (i.e. messy offices), so a network that can learn sub-scene level structures by itself may address this issue.

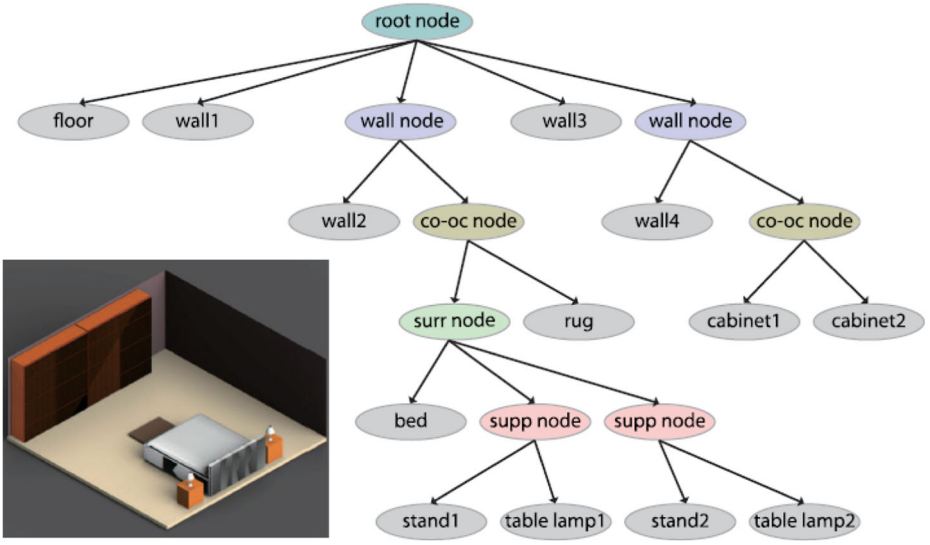

Fig. 5 A training set bedroom example with the corresponding scene hierarchy. The root node have five children which are one floor node and four wall nodes. Then each wall has its own subtree with more detailed object-object relations [33]

8 3D scene datasets

8.1 Cordts et al.'s Cityscapes dataset (2016)

Cordts et al. [14] built a 2D scene image dataset, named Cityscapes, which contains urban street scenes recorded by stereo video for 50 cities. This dataset is composed of finely-annotated and coarsely-annotated images. 5,000 finely-annotated images are manually selected from 27 cities, which contain highly diverse objects, background and layouts. 20,000 coarsely-annotated images are automatically selected from the videos. In order to increase the annotation speed, the object boundaries are not as accurate as finely-annotated images, but they still have a 97% segmentation accuracy.

8.2 Hua et al.'s SceneNN dataset (2016)

Hua et al. [31] created a richly annotated RGB-D indoor scene dataset named SceneNN. It contains 100 scenes categories annotated at vertex, mesh and pixel level, respectively. This multi-level annotation was designed to promote its usage in diverse related applications.

8.3 Xiang et al.'s ObjectNet3D dataset (2016)

Xiang et al. [70] released a large scale dataset called ObjectNet3D containing 100 categories of scene data. There are 90,127 scene images comprising 201,888 objects, and 44,147 3D objects in the dataset. It has performed 2D images-3D shapes alignment, and also provides pose and shape annotations for the 3D shapes.

8.4 Handa et al.'s SceneNet network and dataset (2016)

Handa et al. [27] designed an automatic 3D scene data synthesis framework to generate synthetic 3D scenes by utilizing existing CAD repositories, and generated about 10,000 synthetic views for five different types of 3D indoor scenes.

8.5 Song et al.'s SUNCG dataset (2017)

Song et al. [53] constructed a SUNCG dataset, a synthetic 3D scene database with manually labeled voxel occupancy and semantic labels. This dataset has 84 categories, and 45,622 different scenes and 2,644 objects across those categories.

8.6 Yuan et al.'s SceneSBR2019 and SceneIBR2019 dataset (2019)

Yuan et al. [73] and Abdul-Rashid et al. [2] compiled two 3D scene retrieval benchmarks, named SceneSBR2019 and SceneIBR2019. SceneSBR2019 is using 2D scene sketches as the input query while SceneIBR2019 is using 2D scene images as the input query. Both benchmarks contain 30 categories, which were selected from the Places88 dataset [82] scene labels. The 88 categories of the Place88 dataset are also shared by the ImageNet [17] and SUN datasets [71]. SceneSBR2019 contains 25 scene sketches for each category, while SceneSBR2019 contains 1,000 scene images for each category. Both SceneSBR2019 and SceneIBR2019 share the same 3,000 3D scene models, which is the target dataset. It is currently the first and largest benchmark for 2D scene sketch/image-based 3D scene retrieval.

8.7 Zheng et al.'s Structured3D dataset (2019)

Zheng et al. [80] built a synthetic dataset, named Structured3D, to meet the increasing demand of symmetries (e.g., lines, cuboids, surfaces) for 3D indoor scene reconstruction and recognition. They first collected a lot of 3D indoor scenes designed by professional specialists. Then, they extracted 3D structures (ceiling, floor, wall, etc) annotations as ground truth from those 3D scenes. Finally, based on the extracted 3D structures, they synthesized and generated high-quality (photo-realistic) 2D scene images.

8.8 Straub et al.'s Replica dataset (2019)

Straub et al. [56] created a dataset, named Replica, which contains 18 different indoor scenes. Compared to other 3D scene datasets such as [15] or [12], the Replica dataset is more realistic because it captures the full indoor scenes and has no missing surfaces. In addition, for each mesh primitive, Replica introduces high dynamic range (HDR) textures by changing the settings of the RGB texture camera. Moreover, Replica also contains glass and mirror reflectors surface information, which also can be rendered and make the 3D scenes appear more realistic.

8.9 Vasiljevic et al.'s DEpth dataset (DIODE) dataset (2019)

Vasiljevic et al. [58] curated a RGB-D 2D scene image dataset, named Dense Indoor and Outdoor DEpth Dataset (DIODE), which contains both indoor and outdoor scene categories. Most existing datasets only contain one domain (either indoor or outdoor) since due to

different scene types of the two domains, indoor and outdoor scene images are obtained with different types of sensor suites. As a result, it is difficult to obtain a good accuracy for related cross-domain problems. This dataset adopts one sensor type, thus making the indoor and outdoor scenes have the same scene type.

8.10 Gupta et al.'s large vocabulary instance segmentation dataset (LVIS) dataset (2019)

Gupta et al. [26] constructed a 2D scene image dataset, named Large Vocabulary Instance Segmentation dataset (LVIS), which contains about 2 million object segmentation masks for more than 1,000 object categories, and about 164K 2D scene images in total. Compared to some related datasets, e.g., COCO [35], LVIS provides a more accurate mask for each segmented object instance, thus will be more beneficial in improving the accuracy of a learning method for scene image object detection or segmentation.

9 Conclusions and future work

9.1 Conclusions

3D scene analysis and processing is important for many applications such as autonomous driving cars and AR/VR industries. Recently, it has received more and more attentions. To improve the performance of related deep neural network models, a large amount of 3D scene data are required. With the increasing popularity and power of 3D scene sensing and capturing devices, it is more and more convenient to obtain more accurate 3D scene data.

This paper aims to provide a comprehensive survey of most recent state-of-the-art 3D scene analysis and processing research methods. We summarize this research area from five directions: (1) 3D scene classification; (2) 3D scene recognition; (3) 3D scene retrieval; (4) 3D scene reconstruction; and (5) 3D scene generation. For each direction, we further classify the involved methods into data-driven and semantics-driven methods. In addition, we also review several most recent and popular 3D scene datasets in this research area. Each dataset meets the needs of one or more research directions in this area.

9.2 Challenges and future work

9.2.1 Challenges

– **Accuracy improvement in 3D scene analysis and processing.** So far, the scholars and researchers have made great progress in the analysis and processing of single 3D object. However, the accuracy of the 3D scene analysis and processing is not as good as expected. Compared to 3D objects, 3D scenes are more complicated. 3D scenes usually contain multiple 3D objects, each having spatial and semantic relationships with others. For example, in a 3D kitchen scene model, if there is a bowl in a sink, then a spatial relationship has been established between the bowl and the sink. On the other hand, it is much more likely to find that both a table and a chair will simultaneously appear in the same (kitchen) scene model than that for both a table and an elephant due to their closer semantic relationship in the context of 3D scenes. Due to the high level of complexities existing in 3D scenes, it is still a challenging and open task to significantly improve the accuracy in analyzing and processing 3D scenes.

– **Lack of a large-scale and/or multimodal 3D scene benchmark dataset.** As we know, the size of a training dataset has great influence on the generalization performance of a machine learning algorithm, especially for a deep learning algorithm. According to our knowledge, at present there is no such widely-used large-scale 3D dataset that can be considered as big in terms of either number of categorical classes or number of variations in each category. 3D scenes are basically stemmed from our daily life, but in quite different forms. For example, we could have more than one thousand of settings for our offices and bedrooms. Considering much less 3D scene data available online if compared with single 3D object data as well as several different 3D scene representations (i.e., 3D graphical models, RGB-D videos, and range scans), it will be a much more challenging task to collect a large-scale and/or multimodal 3D scene benchmark dataset. In fact, this has become a bottleneck in the development of 3D scene analysis and processing research direction.

9.2.2 Future work

The review entitles us to identify current obstacles, as well as next trends. Based on them, we propose several important and challenging future research directions.

– **Developing a semantics-driven machine learning model specifically for 3D scene classification and recognition.** Since a 3D scene model is composed of one or more 3D objects, the semantic information existing in the 3D scene model encodes both the relationship between objects and that between the scene objects and the corresponding scene category, and is thus very useful for the 3D scene classification and recognition problems. For instance, in 2020, we proposed a semantic tree-based 3D scene recognition framework [77] which can effectively capture the scene semantics information and thus significantly improves the scene recognition accuracy. Therefore, to improve either the accuracy or efficiency of a 3D scene classification or recognition algorithm, utilizing semantic information of 3D scenes deserves more attentions
– **Application-oriented 2D scene-based 3D scene reconstruction and generation.** Reconstructing a 3D scene based on single or multiple images and automatically generating synthetic 3D scene data based on a certain type of input (e.g., sketches, text, and natural languages) have a lot of application potentials in our daily lives. For instance, creating 3D scene contents for a new 4D immersive program, like the Disney World's Avatar Flight of Passage Ride [65], or imaginary scenes for preschool education.
– **Developing a novel machine learning model specifically for 3D scene retrieval to bridge the semantic gap between the query and target datasets.** Since either hand-drawn 2D scene query sketches or realistic 2D scene query images differ a lot from target 3D scene models or views, it makes 2D scene sketch/image-based 3D scene retrieval a challenging research direction. Our initial results [77] has demonstrated that employing the semantics existing in 3D scenes can evidently improve 3D scene recognition rate. Therefore, considering the semantic gap [69], it is promising to further enhance the 3D scene retrieval performance by designing a learning-based framework which can automatically learn the semantics, and help to conduct the retrieval at semantics level.
– **Curating a large-scale and/or multimodal 3D scene benchmark dataset.** To meet the requirements of current machine learning algorithms, a large-scale dataset is required for each of the five research directions in 3D scene analysis and processing. While, currently for most of them such datasets are stilling pending, for example

a large-scale 3D scene retrieval benchmark. In addition, most of the existing popular 3D scene datasets contain only a certain type of 3D scene data. For example, Structured3D [80] comprises only 3D indoor scenes, while Cityscapes [14] consists of only street scenes. Thus, to examine the scalability of related algorithms, it is necessary to build 3D scene benchmarks that support diverse modalities of 3D scenes as well.

– **Building an adaptive machine learning model for different kinds of scene data.** Besides building a 3D dataset with various types of 3D scene data, we can propose a new machine learning model which is versatile enough to handle different modalities of 3D scene data. This is challenging but promising since it has great potentials in related practical application scenarios which typically involve big data and cloud computing.

References

1. Abdul-Rashid H, Yuan J, Li B, Lu Y (2018) SHREC'18 2D scene image-based 3D scene retrieval track website. http://orca.st.usm.edu/~bli/SceneIBR2018/
2. Abdul-Rashid H, Yuan J, Li B, Lu Y (2019) SHREC'19 extended 2D scene image-based 3D scene retrieval track website. http://orca.st.usm.edu/~bli/SceneIBR2019/
3. Adler J (2018) O. Öktem learned primal-dual reconstruction. IEEE Trans Med Imaging 37(6):1322–1332
4. Aiger D, Allen B, Golovinskiy A (2017) Large-scale 3D scene classification with multi-view volumetric CNN. arXiv:1712.09216
5. Akase R, Okada Y (2013) Automatic 3D furniture layout based on interactive evolutionary computation. In: 2013 Seventh international conference on complex, intelligent, and software intensive systems. pp 726–731
6. Akase R, Okada Y (2014) Web-based multiuser 3D room layout system using interactive evolutionary computation with conjoint analysis. In: Proceedings of the 7th international symposium on visual information communication and interaction, VINCI '14, pages 178:178–178:187, New York, NY, USA, ACM
7. Behl A, Hosseini Jafari O, Karthik Mustikovela S, Abu Alhaija H, Rother C, Geiger A (2017) Bounding boxes, segmentations and object coordinates: How important is recognition for 3D scene flow estimation in autonomous driving scenarios? In: The IEEE international conference on computer vision (ICCV)
8. Blaha M, Vogel C, Richard A, Wegner JD, Pock T, Schindler K (2016) Large-scale semantic 3D reconstruction: An adaptive multi-resolution model for multi-class volumetric labeling. In: The IEEE conference on computer vision and pattern recognition (CVPR)
9. Bobenrieth C, Seo H, Habibi A, Cordier F (2017) Indoor scene reconstruction from a sparse set of 3D shots. In: Proceedings of the computer graphics international conference, CGI '17, pp 27:1–27:5, New York, NY, USA, ACM
10. Carl D (2016) Tutorial on variational autoencoders. arXiv:1606.05908
11. Carreira J, Sminchisescu C (2012) CPMC: Automatic object segmentation using constrained parametric Min-Cuts. IEEE Trans Pattern Anal Mach Intell 34(7):1312–1328
12. Chang A, Dai A, Funkhouser T, Halber M, Niessner M, Savva M, Song S, Zeng A, Zhang Y (2017) Matterport3D: Learning from RGB-D data in indoor environments. International Conference on 3D Vision (3DV)
13. Chen BX, Sahdev R, Wu D, Zhao X, Papagelis M, Tsotsos J (2018) Scene classification in indoor environments for robots using context based word embeddings, 05. arXiv:1908.06422
14. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2016) The Cityscapes dataset for semantic urban scene understanding. arXiv:1604.01685
15. Dai A, Chang AX, Savva M, Halber M, Funkhouser TA, Nießner M (2017) ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: 2017 IEEE conference on computer vision and pattern recognition, CVPR 2017, Honolulu, HI, USA July 21-26, 2017, pp 2432–2443
16. Dai A, Ritchie D, Bokeloh M, Reed S, Sturm J, Nießner M (2017) ScanComplete: Large-scale scene completion and semantic segmentation for 3D scans. arXiv:1712.10215
17. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) ImageNet: A large-scale hierarchical image database. In: CVPR09
18. Derczynski L (2016) Complementarity, F-score, and NLP evaluation. In: LREC

19. Dong S, Xu K, Zhou Q, Tagliasacchi A, Xin S, Nießner M, Chen B (2019) Multi-robot collaborative dense scene reconstruction. ACM Transactions on Graphics, 38(4):Article 84
20. Ebrahimnezhad H, Ghassemian H (2008) Robust motion from space curves and 3D reconstruction from multiviews using perpendicular double stereo rigs. Image Vis Comput 26(10):1397–1420
21. Fellbaum C (1998) WordNet: an electronic lexical database, Bradford Books
22. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell 32(9):1627–1645
23. Fisher M, Savva M, Li Y, Hanrahan P, Niessner M (2015) Activity-centric scene synthesis for functional 3D scene modeling. ACM Trans Graph 34(6):179:1–179:13
24. Flynn J, Broxton M, Debevec PE, DuVall M, Fyffe G, Overbeck RS, Snavely N, Tucker R (2019) DeepView: View synthesis with learned gradient descent. arXiv:1906.07316
25. Guo H, Guo F (2018) Urban scene 3D reconstruction optimization leveraged by line information. In: Proceedings of the 2nd international conference on innovation in artificial intelligence, ICIAI '18, pages 92–96, New York, NY, USA, ACM
26. Gupta A, Dollar P, Girshick R (2019) LVIS: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, pp 5356–5364
27. Handa A, Patraucean V, Badrinarayanan V, Stent S, Cipolla R (2016) Understanding realworld indoor scenes with synthetic data. In: 2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp 4077–4085
28. Hays J, Efros AA (2008) IM2GPS: Estimating geographic information from a single image. In: Proceedings of the IEEE conf on computer vision and pattern recognition (CVPR)
29. He K, Zhang X, Ren S, Sun J (2016) Identity mappings in deep residual networks. arXiv:1603.05027
30. He X, Zhou Y, Zhou Z, Bai S, Bai X (2018) Triplet center loss for multi-view 3D object retrieval. In: CVPR
31. Hua B, Pham Q, Nguyen DT, Tran M, Yu L, Yeung S (2016) SceneNN: A scene meshes dataset with annotations. In: 3DV, pages 92–101. IEEE Computer society
32. Ladický L, Russell C, Kohli P, Torr PHS (2009) Associative hierarchical CRFs for object class image segmentation. In: 2009 IEEE 12th international conference on computer vision. pp 739–746
33. Li M, Patil AG, Xu K, Chaudhuri S, Khan O, Shamir A, Tu C, Chen B, Cohen-Or D, Zhang H (2019) Grains: Generative recursive autoencoders for indoor scenes. ACM Trans Graph 38(2):12:1–12:16
34. Lin D, Fidler S, Urtasun R (2013) Holistic scene understanding for 3D object detection with RGBD cameras. In: IEEE International conference on computer vision, ICCV 2013, Sydney, Australia, December 1-8, 2013, pp 1417–1424
35. Lin T, Maire M, Belongie SJ, Bourdev LD, Girshick RB, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft COCO: common objects in context. arXiv:1405.0312
36. Loop C (1987) Smooth Subdivision Surfaces Based on Triangles. The University of Utah, Master Thesis
37. Ma R, Patil AG, Fisher M, Li M, Pirk S, Hua B-S, Yeung S-K, Tong X, Guibas L, Zhang H (2018) Language-driven synthesis of 3D scenes from scene databases. ACM Trans Graph 37(6):212:1–212:16
38. Ma Z, Shen X, Cao C (2017) A hybrid CRF framework for semantic 3D reconstruction. In: Proceedings of the 23rd ACM symposium on virtual reality software and technology, VRST '17, pages 14:1–14:4, New York, NY, USA, ACM
39. Mikolov T, Chen K, Corrado GS, Dean J (2013) Efficient estimation of word representations in vector space
40. Miksik O, Vineet V, Lidegaard M, Prasaath R, Niessner M, Golodetz S, Hicks SL, Pérez P, Izadi S, Torr PH (2015) The semantic paintbrush: Interactive 3D mapping and recognition in large outdoor spaces. In: Proceedings of the 33rd Annual ACM conference on human factors in computing systems, CHI'15, pp 3317–3326, New York, NY, USA, ACM
41. Müller-budack E, Pustu-Iren K, Ewerth R (2018) Geolocation estimation of photos using a hierarchical model and scene classification. In: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII, pp 575–592
42. Penner E, Zhang L (2017) Soft 3D reconstruction for view synthesis. ACM Trans Graph 36(6):235:1–235:11
43. Ramezani M, Ebrahimnezhad H (2011) A novel 3D object categorization and retrieval system using geometric features. Int J Inform Commun Technol Res (IJICTR) 4(1):9–20
44. Rangel JC, Cazorla M, García-varea I, Martínez-Gómez J, Fromont É, Sebban M (2016) Scene classification based on semantic labeling. Advanced Robotics 30(11-12):758–769
45. Rematas K, Kemelmacher-Shlizerman I, Curless B, Seitz S (2018) Soccer on your tabletop. In: CVPR
46. Ritchie D, Wang K, Lin Y (2018) Fast and flexible indoor scene synthesis via deep convolutional generative models. arXiv:1811.12463

47. Savinov N, Hàne C., Ladicky L, Pollefeys M (2016) Semantic 3D reconstruction with continuous regularization and ray potentials using a visibility consistency constraint. arXiv:1604.02885
48. Savva M, Chang AX, Agrawala M (2017) Scenesuggest: Context-driven 3D scene design. arXiv:1703.00061
49. Schmitt Industries Inc. (2020) Acuity Laser. https://www.acuitylaser.com/
50. Shi Y, Chang AX, Wu Z, Savva M, Xu K (2019) Hierarchy denoising recursive autoencoders for 3D scene layout prediction. arXiv:1903.03757
51. Silberman N, Hoiem D, Kohli P, Fergus R (2012) Indoor segmentation and support inference from RGBD images. In: Proceedings of the 12th european conference on computer vision - volume Part V, ECCV'12, pages 746–760, Berlin, Heidelberg, Springer-Verlag
52. Song S, Yu F, Zeng A, Chang AX, Savva M, Funkhouser TA (2016) Semantic scene completion from a single depth image. arXiv:1611.08974
53. Song S, Yu F, Zeng A, Chang AX, Savva M, Funkhouser TA (2017) Semantic scene completion from a single depth image. In: CVPR, pp 190–198. IEEE Computer Society
54. Sood G Clarifai: R Client for the Clarifai API, 2015. R package version 0.2
55. Steinhauser D, Ruepp O, Burschka D (2008) Motion segmentation and scene classification from 3D LIDAR data. In: 2008 IEEE intelligent vehicles symposium, pp 398–403
56. Straub J, Whelan T, Ma L, Chen Y, Wijmans E, Green S, Engel JJ, Mur-Artal R, Ren C, Verma S, Clarkson A, Yan M, Budge B, Yan Y, Pan X, Yon J, Zou Y, Leon K, Carter N, Briales J, Gillingham T, Mueggler E, Pesqueira L, Savva M, Batra D, Strasdat HM, Nardi RD, Goesele M, Lovegrove S, Newcombe R (2019) The Replica dataset:, A digital replica of indoor spaces. arXiv:1906.05797
57. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2015) Rethinking the inception architecture for computer vision. arXiv:1512.00567
58. Vasiljevic I, Kolkin N, Zhang S, Luo R, Wang H, Dai FZ, Daniele AF, Mostajabi M, Basart S, Walter MR, Shakhnarovich G (2019) DIODE: A dense indoor and outdoor DEpth Dataset. arXiv:1908.00463
59. Vineet V, Miksik O, Lidegaard M, Nießner M, Golodetz S, Prisacariu VA, Kähler O, Murray DW, Izadi S, Pérez P, Torr PHS (2015) Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. In: 2015 IEEE international conference on robotics and automation (ICRA), pp 75–82
60. Vineet V, Warrell J, Torr PH (2014) Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. Int J Comput Vision 110(3):290–307
61. Vo NN, Jacobs N, Hays J (2017) Revisiting IM2GPS in the deep learning era. arXiv:1705.04838
62. Walczak K, Flotyński J (2015) Semantic query-based generation of customized 3D scenes. In: Proceedings of the 20th international conference on 3D web technology, Web3D '15, pp 123–131, New York, NY, USA, ACM
63. Wang L, Guo S, Huang W, Xiong Y, Qiao Y (2016) Knowledge guided disambiguation for large-scale scene classification with multi-resolution CNNs. arXiv:1610.01119
64. Weyand T, Kostrikov I, Philbin J (2016) Planet - photo geolocation with convolutional neural networks. arXiv:1602.05314
65. Wikipedia (2019) Avatar flight of passage. http://en.wikipedia.org/wiki/Avatar_Flight_of_Passage
66. Wikipedia contributors (2020) Jaccard — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Jaccard_index
67. Wikipedia contributors (2020) Leap Motion — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Leap_Motion
68. Wikipedia contributors (2020) Lidar — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Lidar
69. Wikipedia contributors (2020) Semantic gap — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Semantic_gap
70. Xiang Y, Kim W, Chen W, Ji J, Choy CB, Su H, Mottaghi R, Guibas LJ, Savarese S (2016) ObjectNet3D: A large scale database for 3D object recognition. In: ECCV (8), volume 9912 of lecture notes in computer science, pp 160–176. Springer
71. Xiao J, Ehinger KA, Hays J, Torralba A, Oliva A (2016) SUN database: Exploring a large collection of scene categories. Int J Comput Vision 119(1):3–22
72. Xu K, Zheng L, Yan Z, Yan G, Zhang E, Niessner M, Deussen O, Cohen-Or D, Huang H (2017) Autonomous reconstruction of unknown indoor scenes guided by time-varying tensor fields. ACM Trans Graph 36(6):202:1–202:15
73. Yuan J, Abdul-Rashid H, Li B, Lu Y (2019) SHREC'19 extended 2D scene sketch-based 3D scene retrieval track website. http://orca.st.usm.edu/~bli/SceneSBR2019/
74. Yuan J, Abdul-rashid H, Li B, Lu Y (2019) Sketch/image-based 3D scene retrieval: benchmark, algorithm, evaluation. In: 2nd IEEE conference on multimedia information processing and retrieval, MIPR 2019, San Jose, CA, USA, March 28-30, 2019, pp 264–269

75. Yuan J, Abdul-Rashid H, Li B, Lu Y, Schreck T, Bai S, Bai X, Bui N-M, Do MN, Do T-L, Duong A-D, He K, He X, Holenderski M, Jarnikov D, Le T-K, Li W, Liu A, Liu X, Menkovski V, Nguyen K-T, Nguyen T-A, Nguyen V-T, Nie W, Ninh V-T, Rey P, Su Y, Ton-That V, Tran M-T, Wang T, Xiang S, Zhe S, Zhou H, Zhou Y, Zhou Z (2020) A comparison of methods for 3D scene shape retrieval. Comput Vis Image Underst 201:103070
76. Yuan J, Li B, Lu Y (2018) SHREC'18 2D scene sketch-based 3D scene retrieval track website. http://orca.st.usm.edu/~bli/SceneSBR2018/
77. Yuan J, Wang T, Zhe S, Lu Y, Li B (2020) Semantic tree-based 3D scene model recognition. In: 3rd IEEE Conference on multimedia information processing and retrieval, MIPR 2020, Shenzhen, Guangdong, China, April 9-11, 2020
78. Zhang S, Han Z, Zhang H (2016) User guided 3D scene enrichment. In: Proceedings of the 15th ACM SIGGRAPH conference on virtual-reality continuum and its applications in industry - Volume 1, VRCAI '16, pages 353–362, New York, NY, USA, ACM
79. Zhao H, Puig X, Zhou B, Fidler S, Torralba A (2017) Open vocabulary scene parsing. In: 2017 IEEE international conference on computer vision (ICCV), pp 2021–2029
80. Zheng J, Zhang J, Li J, Tang R, Gao S, Zhou Z (2019) Structured3D: a large photo-realistic dataset for structured 3D modeling. arXiv:1908.00222
81. Zhong W, Raj ANJ, Shivakumara P, Zhuang Z, Lu T, Pal U (2018) A new shadow detection and depth removal method for 3D text recognition in scene images. In: Proceedings of the 2018 2nd International conference on computer science and artificial intelligence, CSAI '18, pages 277–281, New York, NY, USA, ACM
82. Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A (2018) Placesc: a 10 million image database for scene recognition. IEEE Trans Pattern Anal Mach Intell 40(6):1452–1464
83. Zhou B, Zhao H, Puig X, Fidler S, Barriuso A, Torralba A (2016) Semantic understanding of scenes through the ADE20k dataset. arXiv:1608.05442