Single Image 3D Scene Reconstruction Based on ShapeNet Models

Xueyang Chen*, Yifan Ren*, Yaoxu Song*
*Zhiyuan College, Shanghai Jiao Tong University, Shanghai 200240, People's Republic of China

ABSTRACT

The 3D scene reconstruction task is the basis for implementing mixed reality, but traditional single-image scene reconstruction algorithms are difficult to generate regularized models. It is believed that this situation is caused by a lack of prior knowledge, so we try to introduce the model collection ShapeNet [2] to solve this problem. Besides, our approach incorporates traditional model generation algorithms. The predicted artificial indoor objects as indicators will match models in ShapeNet. The refined models selected from ShapeNet will then replace the rough ones to produce the final 3D scene. These selected models from the model library will greatly improve the aesthetics of the reconstructed 3D scene. We test our method on the NYU-v2 [11] dataset and achieve pleasing results.

Keywords: ShapeNet, 3D reconstruction, matching, point cloud

1. INTRODUCTION

How fascinating it would be if what you can see can be digitized immediately into 3d scenes or models for later interaction. Reconstructing the image captured by the camera into a 3D scene is one way to accomplish Mixed Reality. Besides cameras, there are far more single images stored in various places and devices, which can also be used in 3D scene reconstruction. However, the method of 3d scene reconstruction using only a single image may not produce pleasing results of regularized models. In past research, many irregular depressions or protrusions are encountered on the reconstructed models. As you can imagine, with merely a single image, there is too little information to determine the shape of the obscured object. It can be improved, though, if other prior knowledge is introduced, such as the Application Scenarios. In particular, for indoor artificial objects, the task to generate corresponding models can be accomplished more easily.

In this paper, the ShapeNet model collection [2] is introduced as the required prior knowledge. Models of interior objects of a large number are included in the ShapeNet model collection. Our insight is to use these regular models to design an end-to-end process that can reconstruct regularized 3D models using only a single image.

What we have tried is to match the rough model with the precise ones from the ShapeNet model collection. The regular model from this collection that matches the primary result best will directly replace the rough model with a bit of scaling and rotation adjustment. The main challenges are to choose a traditional model generation algorithm with stable performance and to design an efficient model matching algorithm. For the first part, which is the model generating algorithm, the powerful Total3DUnderstanding [7] is utilized as support. As for the second part, we sample point clouds from the generated models, as well as from the ShapeNet. Then the problem is reduced to the point cloud matching algorithm, where there exist many algorithms that give satisfying performances, such as ICP[1].

We validate the feasibility of our process on the NYU-v2 [11] dataset, which is a dataset of indoor scene images with semantic segmentation labels. With the support of ShapeNet, our reconstructed 3D scenes contain smoother and more reasonable internal objects.

2. RELATED WORKS

Object Shape Detection. Recent methods reconstruct 3D shapes. For instance, [4] estimates scene structure from a single image, and [5, 8] estimates 3D shapes and poses at the instance level. However, these approaches merely sketch the coarse spatial layout, all of which highly depend on preprocessed or post-processed detailed object information.
Object Representation. Single object shapes are represented as point clouds [6], patches [14], voxel grids [3, 13], and triangle meshes [10, 15]. A hybrid representation of point clouds and triangle meshes is used in our approach due to the tradeoff between time efficiency and representative flexibility.

Classification from 3D Inputs. A growing number of works tend to focus on predicting object classes from 3D inputs, for example, point clouds. [12] exploits the efficient sparse bilateral convolutional networks, while [9] applies hierarchical networks to capture local features from multiple scales. In our method, we leverage these algorithms to finetune the primitive meshes that will be embedded into 3D bounding boxes.

Conclusion. Different from all the above works, our method reconstructs a 3D scene with fine-grained objects from a single image. Particularly, with the prior knowledge from ShapeNet, our method produces 3D object bounding boxes and meshes from this image.

3. METHOD

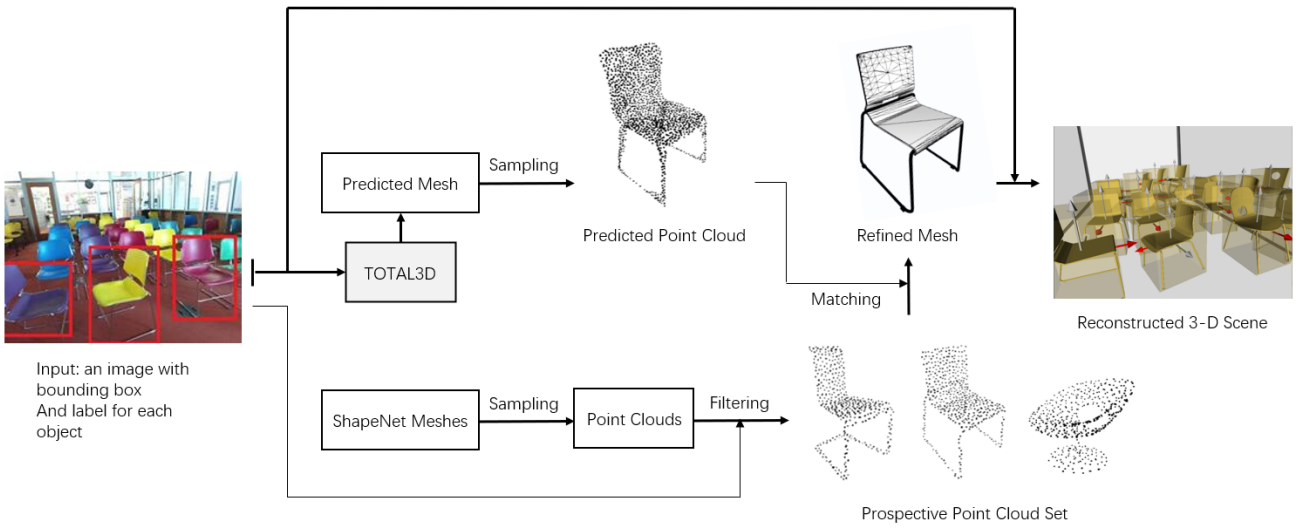

Figure 1. Overview of our approach.

As illustrated in **Figure 1**, we propose an end-to-end method given a single image as input and the 3-D reconstruction result from the image as output. The intermediate networks we utilize were first proposed in Total3D [7]. When given a single image with 2D bounding boxes as its input, those networks can construct a roughly 3-D bounding box and mesh for each object. These meshes are then sampled to produce point clouds of the corresponding objects. On the other side of our work, we sample a part of the model in ShapeNet to get their point clouds. Then labels of the objects being matched would serve as indicators of how to split the original point cloud set into smaller sets, each corresponding to a label. Finally, with a refined model in ShapeNet for each object, we can embed the 3-D model of each object into the 3-D scene and complete the reconstruction work.

3.1 Point Cloud Construction

We mainly use the following three networks, which are first proposed in Total3DUnderstanding [7]. First, by the Layout Estimation Network, we can produce the layout bounding box of the input single image. Then, the Object Detection Network helps to detect the 3-D bounding box of each detected object. In those bounding boxes, we utilize the Mesh Generation Network to generate object meshes. This method bridges the gap between mesh reconstruction and scene understanding. After this work, a 3-D scene constructed by meshes of each object is generated successfully. Then by the mesh sampling process, we can sample thousands of points uniformly from the mesh to generate the point cloud.

3.2 Point Cloud Matching

In this part, we first sample a part of the model in ShapeNet [2] to get their point clouds. This work helps to match point clouds sampled from any object. Then the size of possible matching point cloud sets is reduced, which is achieved by filtering the original sets to much smaller sets with specific labels of certain objects as indicators, for instance, all point clouds with the label “chair”. Then the point clouds that were constructed before are utilized to match those prospective

matching point clouds. By figuring out those precise models, it is much more possible to match the model to the highest fitness and improve accuracy when reconstructing the 3D module.

We apply the point-to-point ICP algorithm [1] provided by Open3D [16] to register the point cloud of each object with those potential matching point clouds generated by sampling uniformly from the models in ShapeNet. The ICP algorithm iterates over the following two steps. Firstly, find correspondence set $K = \{(p, q)\}$ from the target point cloud P, and source point cloud Q transformed with current transformation matrix T. Secondly, update the transformation matrix T by minimizing an objective function $E(T)$ defined over the correspondence set K. Specifically the point-to-point ICP uses the objective

$$E(T) = \sum_{(p,q)\in K} ||p - Tq||^2$$

Before applying the ICP algorithm, we first find the best coordinate and size for the source point cloud, and an initial transformation matrix T to improve the performance of our registration process. Then, after it runs, whether two point clouds match with each other is determined by checking whether there are enough point pairs with the distance between them less than a prescribed threshold.

3.3 3D Scene Reconstruction

So far, we have obtained the best matching model for each object. The only thing that is left is to reconstruct the 3D scene. The location where they should be placed is given in the Layout Estimation Network in Total3DUnderstanding [7]. The work left is thus to replace those objects with the corresponding model.

4. EXPERIMENTS

Figure 2. The 1st column is original images (without drawing 2D-bounding boxes). The 2nd column is the 3D-bounding box predicted by Total3d[1]. The 3rd is predicted meshes generated by Total3D. The 4th column is the results of our matching and final reconstructed scene. Comparing the 3rd with the 4th, our method does achieve better performance.

4.1 Dataset and Toolbox

We use the NYU-v2 [11] dataset for evaluation. Taking the performance of our equipment into account, the *misc_part* in the dataset is selected, which contains 1449 images captured from 12 indoor scenes. In particular, each image has depth tags and object segmentation labels. With the help of nyuv2-python-toolbox, we can easily generate the bounding box for each object for subsequent processing. As to point cloud matching, we use the Point-to-point ICP matching algorithm provided by Open3D [16].

4.2 Comparison with Total3DUnderstanding

Our method is an external version of Total3DUnderstanding. As shown in **Figure 2**, a threshold is set for the matching algorithm, and only if the fitness between the predicted model and refined model was greater enough will the replacements in the final reconstruction process take place. This allows filtering out refined models with poor matching results. The models predicted by Total3DUnderstanding are considered as a baseline.

4.3 Transition Matrix of Matching

Since our point cloud matching algorithm needs iteration over an initial transition matrix, the initialization of this transition matrix will play a key role in matching performance. The transition process includes scaling, translation, and rotation. We tried to set the matrix randomly, which yet did not perform well. The fact helps here is that most common indoor objects are not upside down, and vary not much in size as well. In the case of translation, we can calibrate them according to their centers of gravity. Inspired by these facts, we could merely rotate and randomly scale the initial matrix to a mild extent, which can greatly reduce the number of match attempts.

Figure 3. The 1st, 3rd column is the original image, The 2nd, 4th column is our 3D reconstructed result.

5. LIMITATION & FUTURE WORK

Unfortunately, the generated model by matching from ShapeNet does not always work as well as it is expected. As is shown in the experiments, two parts can be improved in the future.

5.1 Better Accuracy

We ended up with a result that in a sense already met our requirements. For models that fit well with the models in ShapeNet, we can indeed find a model that performed much better than the generated one. However, it was not always able to find the right model to match. Moreover, the performance of the model generation network we use, Total3DUnderstanding, is also not satisfactory, thus causing a large noise interference in our subsequent matching work, which is one of the reasons why our final refined models are not quite satisfying.
Since Total3D is trained on the NYU37 dataset, there must be higher quality for the predicted mesh, instead of the NYU-v2 dataset. We have tried our method with the best demonstrations from NYU37 that were provided by Total3D, which are shown in **Figure 2**. We believe that better-predicted mesh will also make our refined mesh better.
Another drawback is that we evaluate the difference between meshes by evaluating the sampled point cloud. The message will be lost for sure. So we are wondering if there are better sampling algorithms or if we can evaluate meshes directly.

5.2 Better Efficiency

In this paper, we sample the point cloud from each mesh and use the ICP algorithm for evaluation. In a picture, i.e. a scene, there are many detected objects. Meanwhile, it is unaffordable to totally scan the models in ShapeNet for every round of matching due to a large number of models. As a result, we choose to reuse the object class label generated by Total3D. Such improvements are far from enough. There are hundreds of given models in a category of objects in ShapeNet. The algorithm in each point cloud matching process also requires several thousand iterations to find the best matching transition matrix. Thus there is still space for improvement for the time efficiency of our model generation.

6. CONCLUSION

We have proposed a novel method for the 3D scene reconstruction task of a single image, which is based on ShapeNet model collections and can mostly ensure the smoothness of the generated model mesh. The idea is to introduce prior knowledge and then we can match the result generated from other methods with it, to achieve better and smoother results.

REFERENCE

[1] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 14, no. 2, pp. 239-256, Feb. 1992.

[2] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, F. Yu, “ShapeNet: An Information-Rich 3D Model Repository”, 2015.

[3] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, “3D-R2N2: A unified approach for single and multi-view 3d object reconstruction”, In ECCV, 2016.

[4] D. Hoiem, A. A. Efros, and M. Hebert, “Geometric context from a single image”, In ICCV, 2005.

[5] A. Kundu, Y. Li, and J. M. Rehg, “3d-rcnn: Instance-level 3d object reconstruction via render-and-compare”, In CVPR, 2018.

[6] P. Mandikal, Navaneet KL, and R. V. Babu, “3d-psrnet: Part segmented 3d point cloud reconstruction from a single image”, In ECCV, 2018.

[7] Nie, Y. Han, X. Guo, S. Zheng, Y. Chang, J. Zhang, J. Jun, “Total3DUnderstading: Joint Layout, Object Pose and Mesh Reconstruction for Indoor Scenes from a Single Image”, In CVPR, 2020.

[8] G. Pavlakos, X. Zhou, A. Chan, K. G. Derpanis, and K. Dani-ilidis, “6-dof object pose from semantic keypoints”, In ICRA, 2017.

[9] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space”, In NeurIPS, 2017.

[10] J. Rock, T. Gupta, J. Thorsen, J. Gwak, D. Shin, and D. Hoiem, “Completing 3d object shape from one depth image”, In CVPR, 2015.

[11] N. Silberman, P. Kohli, D. Hoiem, R. Fergus, “Indoor Segmentation and Support Inference from RGBD Inputs”, In ECCV, 2012.

[12] H. Su, V. Jampani, D. Sun, S. Maji, E. Kalogerakis, M. H. Yang, and J. Kautz, “SPLATNet: Sparse lattice networks for point cloud processing”, In CVPR, 2018.

[13] B. Wallace and B. Hariharan, “Few-shot generalization for single-image 3d reconstruction via priors”, In Proceedings of the IEEE International Conference on Computer Vision, pages 3818–3827, 2019

[14] P. Wang, C. Sun, Y. Liu, and X. Tong, “Adaptive o-cnn: a patch-based deep representation of 3d shapes”, In SIGGRAPH Asia 2018 Technical Papers, page217. ACM, 2018.

[15] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, “Pixel2Mesh: Generating 3D mesh models from single RGB images”, In ECCV, 2018.

[16] Qian-Yi Zhou, Jaesik Park, Vladlen Koltun: “Open3D: A Modern Library for 3D Data Processing”, 2018; [http://arxiv.org/abs/1801.09847 arXiv:1801.09847].

AUTHORS’ BACKGROUND

Your Name	Email	Title*	Research Field	Personal website
Yifan Ren	ivan-ren@sjtu.edu.cn	undergraduate student	Computer Science	
Xueyang Chen	anoxiacxy@sjtu.edu.cn	undergraduate student	Computer Science	
Yaoxu Song	Richard_K@sjtu.edu.cn	undergraduate student	Computer Science	