2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) | 978-1-6654-0505-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCWAMTIP51612.2020.9317479

3D SHAPE RECONSTRUCTION OF FURNITURE OBJECT FROM A SINGLE REAL INDOOR IMAGE

LI XI[1], KUANG PING[1], GU XIAOFENG[1], HE MINGYUN[1]

[1]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
E-MAIL: xi-li@foxmail.com, kuangping@uestc.edu.cn, guxf@uestc.edu.cn, hmyun@uestc.edu.cn

Abstract:

Most of the current researches on 3D shape reconstruction based on deep learning only focus on clean-background images. In this paper, we propose a system that segments furniture object from a single real indoor image and reconstructs their 3D meshes. This system takes Pixel2Mesh, a 3D shape reconstruction network, as a branch of the state-of-the-art instance segmentation network Mask Scoring R-CNN. We trained our system on 3D-FUTURE dataset, and the results show that our method can effectively reconstruct 3D shape of furniture objects from real indoor images by designing proper loss functions and combining instance segmentation and 3D shape reconstruction network.

Keywords:

3D shape reconstruction; Real indoor image; Furniture object; Deep learning; Instance segmentation

1. Introduction

Due to the interference of complex background, slightly occluded or partial absence, it is difficult to reconstruct 3D shapes of furniture object from a single real indoor image. At present, most of the researches about three-dimensional shape reconstruction focus on images without background.

We propose a system, which equips a 3D reconstruction branch to the instance segmentation network MS R-CNN [6]. This branch takes the region feature of each target object as the inputs of the three-dimensional shape reconstruction network Pixel2Mesh(P2M) [1], and outputs the 3D mesh of the object by deforming an original sphere. In general, the input of this system is real-world RGB images containing furniture objects, and the outputs are the category, bounding box, mask and three-dimensional mesh of furniture objects.

We trained our system on 3D-FUTURE [5], a large real indoor scenes dataset released by Alibaba. The experimental results show this system can effectively reconstruct the 3D shape of furniture objects in single real indoor scenes images.

The contributions of this paper are highlighted as follows:

• Designing a system that can reconstruct 3D mesh of furniture objects from single view real indoor RGB images quickly and cheaply.

• Designing loss functions for this system properly.

2. Related works

2.1. Traditional modeling from images

Most traditional modeling methods focus on multi-view such as using 3D reconstruction software like Meshroom, 3D models can be generated with high resolution based on the Alice Vision Photogrammetric Computer Vision framework. It infers the geometry of a scene from a set of unordered photographs or videos.

These methods need multiple images of objects and spend a lot of time for professional designers to obtain accurate 3D models. Ideally, people expect three-dimensional CAD models can be modeled from a single real-world image quickly and cheaply.

2.2. Instance segmentation

Instance segmentation network Mask R-CNN [3], a learning-based method, extends the object detection network to get regions of object instances, and then give masks for every region. MS R-CNN [6] moves further forward, it adds a MaskIoU branch for mask score based on mask head, which solves the potential problem of Mask R-CNN only using classification score to measure mask quality.

2.3. Learning-based 3D shape reconstruction

In recent years, neural-network researchers have proposed a variety of forms of 3D representations based on deep-learning methods and computer graphics.

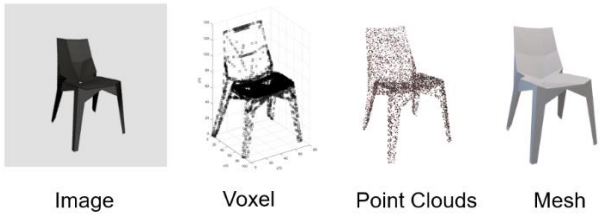

Fig.1 3D representations

3D-R2N2 [6] reconstruct 3D voxels from single or multiple pictures end-to-end; besides, there are methods based on point cloud representation, such as PSG [1], attempts to reconstruct 3D shape from RGB or RGB-D images. Moreover, Pixel2Mesh [1], based on triangular mesh representation, uses Graph Convolution Network (GCN) to represent 3D mesh information, and gradually deforms the ellipsoid mesh by using the features, which extracted from the input RGB image, to generate the desired 3D geometry shape. These methods only focus on the clean-background images.

2.4. Datasets

ShapeNet [8] is a large three-dimensional CAD model dataset, whose 3D CAD are rendered to generate synthetic pictures. Pix3D [3] is a single-image 3D shape modeling dataset, which provides more accurate 2D-3D alignment for 395 3D shapes of 9 object categories. Lately, Alibaba released 3D-FUTURE [5], which contains more than twenty thousand realistic synthetic images in 5,000+ different tidy indoor scenes. These images involve more than ten thousand unique industrial three-dimensional instances of furniture object with high-resolution information-rich textures developed by a bunch of professional designers.

3. Methods

3.1. Network architecture

The instance segmentation network MS R-CNN consists of 2 branches. R-CNN Branch is responsible for the object detection task and outputs the classification results and bounding boxes of the detected objects, while the Mask/MaskIoU Branch is responsible for the instance segmentation task and outputs the masks and mask scores of the target objects.

As shown in Figure 2, we augment MS R-CNN with P2M as a Shape Deformation Branch, which uses an ellipsoid mesh with 156 vertices and 462 edges as the initial shape of any objects, and then gradually uses the 2D image features to adjust the node state of the Graph Convolution Network in 3D mesh to deform the ellipsoid into a target object mesh with 2562 vertices and 5120 faces. Instead of using the feature maps from different layers of VGG in [4], we take ResNeXt-101 and FPN as our feature-extractor and use feature maps resulting from RoIAlign to feed the Shape Deformed Branch.

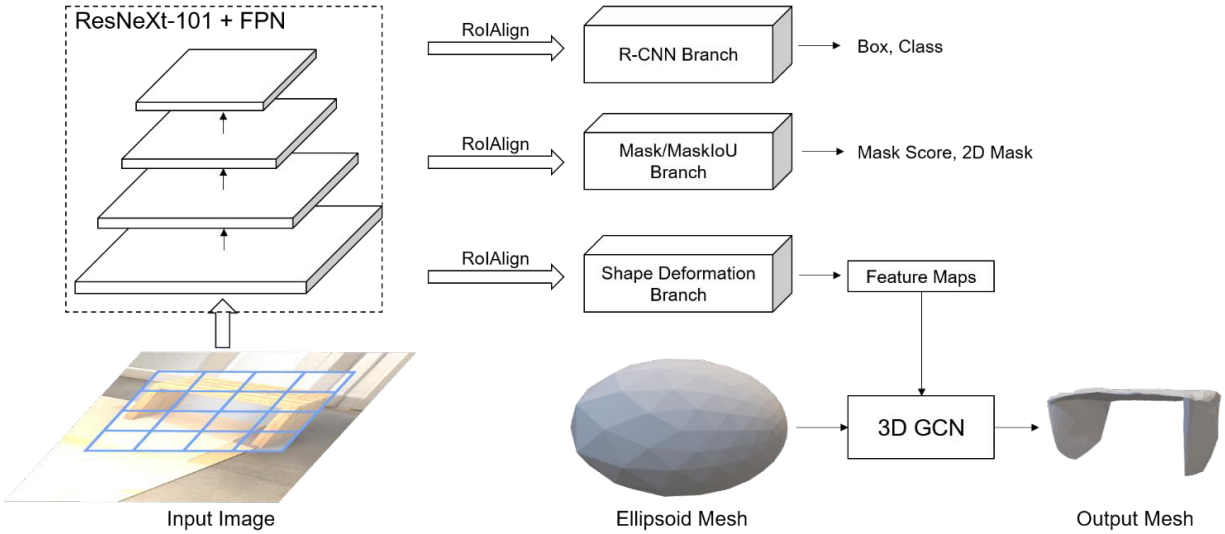

Fig.2 Network architecture

3.2. Loss functions

We properly design the loss functions for our system. X and Y are defined as the point clouds uniformly sampled from the surface of the output mesh and the ground truth mesh. The chamfer distance is used to improve the similarity between X and Y, which is given by

$$L_{cd} = \sum_{x \in X} min_{y \in Y} \|x - y\|^2 + \sum_{y \in Y} min_{x \in X} \|y - x\|^2 \quad (1)$$

A normal distance between X and Y point clouds is used to ensure the smooth surface of the output shape. We set p as the nearest neighbor point of q in Y, then let

$$\Delta = <x, y> = \{(p, q) | \, p \in X, q \in Y\} \quad (2)$$

to represents the set of pairs $<p, q>$, and let υ_x be the observed surface normal to point x from ground truth. Then the absolute normal distance is given by

$$L_{norm} = -\left(\frac{\sum_{x,y \in \Delta} \|\upsilon_x \cdot \upsilon_y\|^2}{|X|} + \frac{\sum_{y,x \in \Delta} \|\upsilon_y \cdot \upsilon_x\|^2}{|Y|}\right) \quad (3)$$

Last, we use a Laplacian regularization defined in [1] to penalizes the unsmooth edges and vertices to guarantee the superior quality mesh predictions, which we named L_{edge} in our system. All the loss functions in our system can be concluded as:

$$L_{total} = L_{cls} + L_{box} + L_{mask} + \lambda_{cd} \cdot L_{cd} + \lambda_{norm} \cdot L_{norm} + \lambda_{edge} \cdot L_{edge} \quad (4)$$

all λ_* are to ensure the balance of all loss functions of whole system. Finally, our system combines object detection losses, instance segmentation losses and mesh prediction losses for training.

4. Experiments

4.1. Datasets

For three-dimensional shape reconstruction task, 3D-FUTURE [5] provides 5203 unique precision furniture CAD models with textures.

For evaluating our system on 3D-FUTURE, we used Blender engine to render 12 view textured renderings (with gray background) and masks for each CAD model, and then we split a validation set with 1000 clean-background images and corresponding models for measuring the quality of the predict meshes, which was called D_1. The resolution of each image is 256∗256.

Besides, 3D-FUTURE offers 11676 real interior images containing furniture objects of 2851 different models, which was named D_2. For this part, we reserve 5% of the training models and images to measure the performance of our system on real pictures.

4.2. Implementation details

To train our system on 4 Tesla M60 GPUs, we set batch-size is 8 and epoch is 50. Our system optimized using Adam with weight decay 2e-4. The learning rate is initialized as 2e-3 and warm up 3 epochs and then decays to (2e-4, 2e-5) at epoch (35, 45). Hyper parameters λ_* used in Equation (4) are $\lambda_{cd} = 1$, $\lambda_{norm} = 0.1$ and $\lambda_{edge} = 1$.

4.3. Evaluation

Following [1], we also take Chamfer Distance and Fscore as our metrics to measure the quality of the 3D shape reconstruction. Lower is better for CD, and higher is better for Fscore. To compute CD, we random sample 2048 points uniformly from the meshes surface.

4.4. Results and discussion

To validate the performance of our system in real indoor images and compare with the latest method, we trained our system and Pixel2Mesh on D_1 and D_2, respectively.

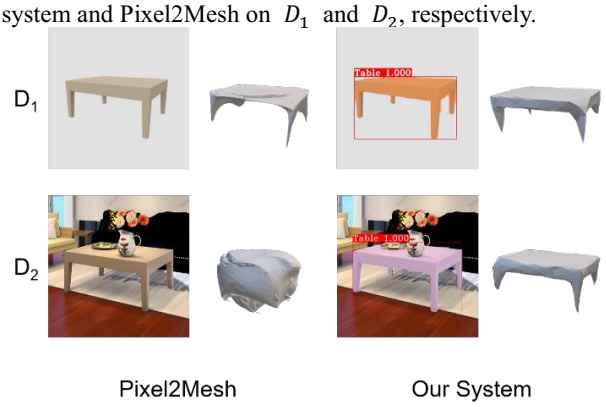

Fig.3 Comparison of results

As shown in the bottom left corner of Figure 3. Due to the interference of complex background and slight occlusion, P2M fails to find the furniture object that needs to be reconstructed. Benefit from RoIAlign proposed in [3], our system can segment and reconstructs furniture objects from real-world images correctly after adequate training.

To measure the quality of our method while compared with the latest method, we take Chamfer Distance and the harmonic mean of precision and recall (Fscore) as our metrics to measure the quality of the shape reconstruction.

The following Table 1 shows the CD and Fscore of our system and P2M on dataset D_1.

Table 1 Performance on D_1

Methods	Dataset	CD	Fscore
P2M	D_1	0.061	76.42
Ours	D_1	0.062	76.35

Due to the images of D_2 has complex backgrounds and slight occlusions, P2M can't get an effective result on D_2. The Table 2 shows the Performances of our system on dataset D_2.

Table 2 Performance on D_2

Methods	Dataset	CD	Fscore
P2M	D_2	/	/
Ours	D_2	0.079	65.87

In real indoor scenes, our system has achieved good results. The following Figure 4 shows some examples of our system's reconstruction results on real images.

Fig.4 Results on real images

5. Conclusions

In this paper, we propose an effective 3D shape reconstruction system for single-view real-world images. In contrast to most existing methods, our goal is to segment furniture objects from real indoor images, then reconstruct their 3D meshes and output their classifications, bounding boxes and masks by combining instance segmentation network MS R-CNN and 3D reconstruction network P2M. Extensive experiments have been conducted to demonstrate its effectiveness. This system can be easily applied in 3D games, AR/VR, home decoration and house rental and sale fields.

Acknowledgements

This work supported by National Key R&D Program of China (2019YFB1406202) and supported by Sichuan Science and Technology Program (2019ZDZX0009).

References

[1] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y u-Gang Jiang. Pixel2Mesh: 3D Mesh Model Generation Via Image Guided Deformation. In IEEE Transactions On Pattern Analysis and Machine Intelligence, 2020.

[2] H. Fan, H. Su, and L. J. Guibas. A Point Set Generation Network For 3D Object Reconstruction From A Single Image. In CVPR, 2017.

[3] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum, and W. T. Freeman. Pix3D: Dataset And Methods For Single-Image 3D Shape Modeling. In CVPR, 2018.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In ICCV, 2017.

[5] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. arXiv preprint arXiv:2009.09633, 2020.

[6] C. B. Choy, D. Xu, J. Y. Gwak, K. Chen, and S. Savarese. 3D-R2N2: A Unified Approach For Single And Multi-View 3D Object Reconstruction. In ECCV, Cham, 2016.

[7] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang. Mask Scoring R-CNN. In CVPR, 2020.

[8] Chang, Angel X. et al. ShapeNet: An Information-Rich 3D Model Repository. In Computer Science, 2015.