Paper Reading Report-05

Han Zhang
u7235649

Abstract

*This is my reading report for the paper titled: "***Holistic 3D Scene Understanding from a Single Image with Implicit Representation***", authored by Cheng Zhang et al, and published in CVPR 2021 Computer Vision and Pattern Recognition (cs.CV).*

All ENGN8501 submissions will be subject to ANU's TurnitIn plagiarism check, against both the original paper, internet resources, as well as all other students' submissions. So please make sure that you must write your own reports, and declare the following originality statement:

I, Han Zhang, hereby confirm that I am the sole author of this report and that I have compiled it in my own words.

1. Problem Statement

The research problem of this paper is 3D indoor scene understanding, which uses a color image to reconstruct the layout of the room and each individual object, and estimate its semantics in 3D space. This problem is a computer vision problem which has an important impact on applications such as robotics and virtual reality.

For single image scene reconstruction, recent methods use the relationship between objects to infer 3D bounding boxes from 2D inspections, the extend methods calibrate the inspection objects with CAD models. However, the results are limited by the size of the CAD model database. Some learning-based methods encode the shape as a feature vector and express it in a traditional way, or decompose the shape into a structured representation method of simple shapes. Recently, implicit surface functions have been widely used, some studies combined structured and implicit representations. Graph convolutional networks (GCN) have been widely used and can predict relationships or detect 3D objects from point cloud data.

This paper proposes a comprehensive 3D scene understanding deep learning system based on deep implicit representations, and proposes a scene context network based on image-based implicit shapes, which uses information in local objects to refine the initial 3D pose and scene layout. They also proposed a physical intrusion loss to prevent crossing objects and produce a reasonable object layout.

2. Summarise the paper's main contributions

The author designed a two-stage single image-based holistic 3D scene understanding system using a local structure implicit network. Second, they proposed a new image-based implicit shape embedding network to extract latent shape information. They also proposed a GCN-based scene context network to refine the arrangement of objects. Finally, they proposed a physical intrusion loss to prevent crossing objects and produce a reasonable object layout.

They gave experimental methods and data, and also compared with the state-of-the-art methods. The results showed that their method performed better in terms of object shape, scene layout estimation, and 3D object detection.

3. Method and Experiment

As shown in Figure 1, the system is divided into two stages, the initial estimation stage with Local Implicit Embedding Network (LIEN), and the refinement stage with Scene Graph Convolutional Network (SGCN).

In the first stage, first extract the 2D bounding box from the image, then use the target detection network to restore the target to a 3D bounding box, and use LIEN to directly learn the implicit information from the image and infer 3D shapes. The LIEN includes an image encoder Resnet-18 and a 3-layer MLP to obtain analytic and latent codes. They link the category code with the image features of the encoder, and introduce a prior in LIEN to improve performance.

In the refinement stage, they designed a new SGCN to refine the initial predictions through the context of the scene. They modeled the entire 3D scene as a graph G based on Graph R-CNN, where the nodes represent objects, the scene layout and the relationship between them. They designed different characteristics for different types of nodes.

They also proposed a new physical conflict loss to ensure that the targets will not intersect. They sample points inside each object, put the center point of the Gaussian element as candidate points into the LDIF decoder to filter out points outside the target surface, and queue $\mathbb{S}_i$ to the LDIF of the k-nearest object to verify intersects. The loss function can

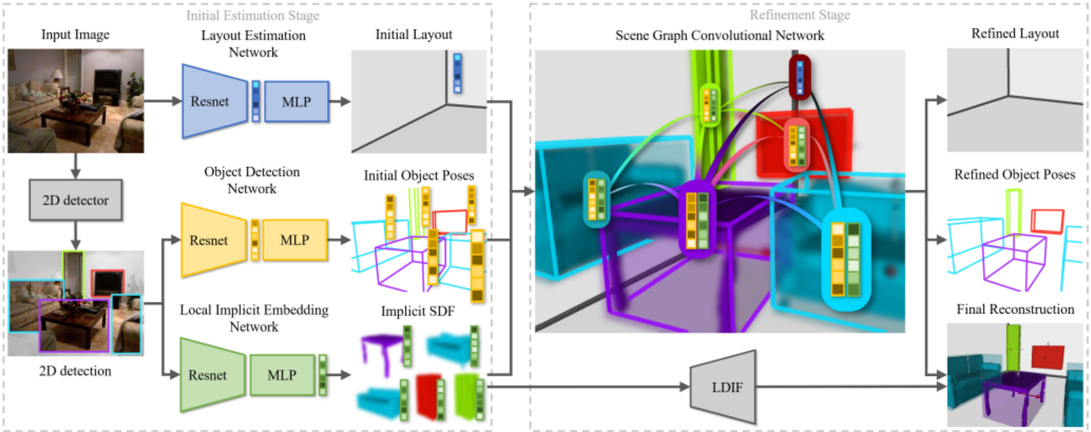

Figure 1: The proposed pipeline.

be expressed as

$$\mathcal{L}_{phy} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathbb{S}_i|}\sum_{x\in\mathbb{S}_i}||relu(0.5 - sig(\alpha LDIF_i(x)))||$$

.

They use Pix3D and SUN RGB-D datasets, first train each module separately then together. They compared their method with AtlasNet, TMN and Total3D in 3D Object Reconstruction, 3D Object Detection, Layout Estimation, Camera Pose Estimation and Holistic Scene Reconstruction, showing that their method performed better. Their ablation study shows GCN, deep implicit feature, physical violation loss and some other factors will influence the performance. Their method also shows good generalization ability on ObjectNet3D dataset.

4. Critical Analysis

4.1. Are the paper's contributions significant?

The contribution is significant. Compared with the existing methods, they gave new network models and loss function, using deep implicit expression to improve the accuracy and fineness of modeling, and showed better generalization performance.

4.2. Are the authors' main claims valid?

The main claims are valid. They gave the mathematical representation and network model of the method, and compared the differences between different methods, and the results showed that their method performed better. They also did ablation study.

4.3. Limitation and weaknesses

The resulting picture shows that there are some details that are still different from the picture when modeling. When estimating the layout of the scene, there are some differences between the details and the grand truth, and it seems to be more inclined to the regular layout. This may be limited by the richness of the data set. Increasing the number and richness of data sets may help.

4.4. Extension and future work

The author can try to use more data sets to train the network to improve the accuracy of 3d modeling and layout estimation in detail. They can also try to synthesize the data set. In the future, this technology may be used in VR, robotics, design and other fields.

4.5. Is the paper stimulating or inspiring ?

This paper is exciting. The new network structure and loss function they proposed greatly improve the accuracy and fineness of 3D modeling and scene layout estimation.

4.6. Conclusion and personal reflection

In conclusion, this paper proposes a method based on learning and deep implicit expression to solve the problem of using a single image for indoor 3D modeling and scene layout estimation end-to-end, and gives a new network structure and loss function to improve Improve the performance of the algorithm. I might try to combine the optical flow method on this basis to assist in modeling. This article tells me that deep implicit expression can express feature measurements more compactly and informatively, and learn from context more effectively.

References

[1] Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, Shuaicheng Liu. *Holis-tic 3D Scene Understanding from a Single Image withImplicit Representation*. CVPR 2021.