Paper Reading Report-05

Shreya Chawla
u7195872

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 (University of Electronic Science and Technology of China) et al, and published in IEEE CVPR 2021.*

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

1. Problem Statement

3D indoor scene understanding from a single RGB image is a computer vision task. It aims to reconstruct 3D room layout, estimate object poses as oriented bounding boxes and object meshes thus achieving semantic understanding and reconstruction of the scene. This is an ill-posed problem making it difficult and complex. As cluttered scene has heavy occlusions it may cause 3D reconstruction pipelines to fail. Other competing solutions include [2].

Previous works like Total3DUnderstanding propose to solve the problem in "box-in-the-box" fashion representing the scene as object meshes with orientated bounding boxes placed in a cuboid room layout. It decomposes the task in a bottom-up way into layout estimation, camera pose estimation, 2D detection, object pose estimation and object reconstruction. However, the existing works still struggle on understanding and utilizing the scene context like humans to better arrange placement of objects.

It is not well considered how to avoid the intersection between objects in the final scene reconstruction. Due to the use of mesh representation by the single object reconstruction method, the output shape is often low quality. Similar to previous work, this paper deep implicit representation to overcome this.

Previous works rely solely on Graph Convolution Network (GCN) for scene understanding. However, the problems of object shape quality and shape intersection remain unsolved. Also, GCN requires rich features like shape priors and context information. To solve these problems, this paper uses implicit representations.

Implicit representations like Signed Distance Function can naturally represent whether a point is inside an object and provide a gradient to push it away from the object, which can be exploited to construct a Physical Violation Loss (PVL) to penalize intersection. This serves as the motivation for the methodology proposed. Moreover, the latest Local Deep Implicit Functions (LDIF) provides a compact while precise method to decode object shapes from a latent vector by incorporating implicit and structured representation, which can further improve the results of single object reconstruction, and to provide more structured shape priors for GCN.

2. Summary of the paper's main contributions

The paper [1] presents a new pipeline to solve the problem. A novel physical loss to avoid incorrect context between objects is introduced by the authors. They propose to incorporate GCN with implicit representations.

3. Method and Experiment

The pipeline consists of two stages - estimation stage and refinement stage. In the initial estimation stage, the image is passed through Layout Estimation Network (LEN) to obtain the initial 3D cuboid room layout. The LEN composes of ResNet and MLP. With a 2D object detector, the objects are extracted. From each bounding box or each image crop they estimate the initial 3D object pose as 3D bounding boxes from Object Detection Network (ODN) and a new Local Implicit Embedding Network (LIEN) embeds an implicit shape code for corresponding object. LIEN help effectively capture contextual information among object.

In the refinement stage, they model layout and objects as a graph to understand the scene context and refine the initial estimation with a GCN. The initial estimations, features, and shape codes from the previous stage are concatenated and embedded into node representations to be further updated with message passing. The updated representations are later decoded into residuals to refine the initial estimations. This GCN, called Scene Graph Convolutional Network (SGCN), is novel to this paper. From different sources, different features are designed for different types of nodes as the input features are the key to an effective GCN. These nodes include - layout node, object node and relationship node. The refined poses are incorporated with the object shapes decoded from shape codes with LDIF and machine code algorithm, to get the final reconstruction of the whole scene. During training, upon all the losses used in Total3DUnderstanding including direct supervision on network outputs, they propose PVL based on the insight that objects should not intersect, to penalize the points inside both adjacent objects.

They also evaluate their technique on other aspects including supporting relation, geometry accuracy, and room layout. Their qualitative and quantitate results show the advantages of pipeline proposed compared to other leading methods. Their method outperforms SOTA by 5.2% in layout estimation while on par with SOTA on camera pose estimation for NYU-37 dataset. Their method gave more accurate bounding box estimation and with less intersection in qualitative comparisons. They significantly bested SOTA over all semantic categories by improving AP by 18.83%.

4. Critical Analysis

4.1. Significance of the paper's contributions

This paper is incremental as shown in several evaluation results. The model for holistic scene understanding leverages deep implicit representation. It not only reconstructs accurate 3D object geometry, but also shows ability to learn better scene context using GCN compared to the prior methods. The novel PVL is shown to deliver accurate scene and object layout. Experiments show that the model outperforms in various tasks in holistic scene understanding.

4.2. Validity of the authors' main claims

Extensive comparative evaluation results suggests the method's superiority. Quantitative comparisons on different metrices with existing leading techniques are made. For object reconstruction comparison they report the mean Chamfer distance to be 6.72 which is the lowest. For 3D object detection comparison, their mAP outperforms all the SOTA for all the classes with mean being 45.21 which is greater than Total3D by almost double. Comparisons on layout IoU, detection mAP, supporting error, average collision volume, corner error, and pixel error also validate the authors' claims.

Their qualitative experiments show better accuracy and sharp edges in predicted scene compared to other SOTA. They compare the input image on the basis of oblique view, camera view and scene reconstructed view with ground truth, Total3D method and their results. The method displays good generalization ability and robustness on other datasets (ObjectNet3D). They verified the effectiveness of each component - SGCN, PVL and Deep Implicit Feature.

4.3. Limitation and weaknesses

In some extreme cases, heavy occlusion might cause their pipeline to fail as some objects might be completely hidden behind other object. Heavy occlusion could also cause failure cases as then the pose of the object might be incorrectly predicted. Extremely cluttered scenes are another case where the model predictions might be very different from the ground truth.

4.4. Extension and future work

One possible future work would to use different architecture to find relation between objects. Another one could be to apply the presented techniques for different application tasks. This methodology could be used to rebuild architectural ruins for reconstruction. It can also be used for indoor interior designing and in property e-sales and marketing by generating a reconstructed scene allowing users to better visualize. This technique could be used for forensic scene reconstruction to 3D visualize the crime scene.

4.5. Is the paper stimulating or inspiring ?

The novel loss function and SGCN are well explained and experimented on. Several qualitative and quantitative comparisons along with sound reasoning for are provided. This provides a better understanding of the results and makes the paper interesting. The impact of deep implicit representation in the paper's results could inspire future works.

4.6. Conclusion and personal reflection

A two stage single image holistic 3D scene understanding system based on deep implicit representation is presented in this paper. A local First, draw a short conclusion about this paper. A local implicit shape embedding network based on LDIF is introduced which extracts latent shape information from one image and leads to superior geometry accuracy. GCN based scene context network refines the object arrangement which well exploits the latent and implicit features. The newly proposed physical violation loss with implicit representation, effectively prevents the object intersection.

If I were tasked to solve this research problem, I would use depth as a parameter to get better correlation between objects in the RGB image.

References

[1] C. Zhang, Z. Cui, Y. Zhang, B. Zeng, M. Pollefeys and S. Liu. *Holistic 3D Scene Understanding from a Single Image with Implicit Representation*. In Proc. IEEE/CVF Conf. on CVPR, 2021. 1

[2] W. Yin, J. Zhang, O. Wang, S. Niklaus, L. Mai, S. Chen and C. Shen. *Learning to Recover 3D Scene Shape from a Single Image* In Proc. IEEE/CVF Conf. on Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 204-213. 1