Research Statement

Yueh-Cheng Liu
ycliu0610@gmail.com

1 Introduction

Nowadays, robots have been widely deployed into factories and warehouses. However, there is still a considerable gap in bringing autonomous robots into our daily lives, such as autonomous vehicles or home robots. One of the main reasons is that the ability to perceive and fully understand the noisy, unstructured, and dynamic environment is still lacking.

Therefore, my primary research objective is to *enable machines to reconstruct and understand the 3D world* for the benefit of autonomous robots. I believe that this perceptual intelligence includes **semantic understanding** and **geometric understanding** of the environment. For example, a home service robot is required to model the 3D indoor structure accurately to navigate through the environment. To execute an action command successfully, it must understand the object or location the human refers to.

Thanks to the development of deep learning, scene understanding has recently experienced major improvement by training large neural networks to perform specific tasks. We can also train neural networks to estimate geometric or structural information, such as depth estimation and indoor layout estimation. However, handling the noisy and unseen environment remains challenging when the system is deployed into the real world. Besides, the high demand for large-scale labeled data to train the system is a critical issue. Therefore, in my vision, building a robust and data-efficient 3D scene understanding algorithm that can easily scale as well as generalize to the new environment is crucial.

2 3D Semantic and Geometric Understanding

Semantic 3D scene understanding and data-efficient training. We observed that existing neural networks for 3D semantic segmentation were poor at inferring texture information, such as detecting framed pictures hung on walls in 3D. Therefore, we study the fusion of features from different modalities (e.g., color, geometry) and global scene context to improve the the performance of 3D semantic segmentation. At the end, we proposed a novel network architecture based on PointNet++ [1] with the 2D-3D fusion layer that combined the advantage of 2D and 3D networks. The method was the second place on the ScanNet benchmark in 2018 and was published at 3DV 2019 [2].

On the other hand, I notice that the scarcity of 3D labeled data is a major obstacle to building a learning-based 3D scene understanding system. Hence, I attempted to mitigate this issue in two ways – pretraining and active learning. In my first work [3], since 2D labeled (image) data are more abundant than 3D, I adopted knowledge distillation and contrastive learning to learn latent 3D representation from 2D neural networks as an objective for 3D pretraining. The work demonstrated that 3D pretraining by this "learning from 2D" approach significantly improves downstream tasks like 3D object detection and 3D semantic segmentation. In the other work [4], my partner and I designed an active learning framework for 3D semantic segmentation to reduce labeling efforts. During the training process, the algorithm selects point cloud regions with large uncertainty and diversity for label acquisition. As a result, the method with only 15% of data can achieve performance similar to fully-supervised training using all the data, and it is accepted to ICCV 2021.

Learning-based robotic grasp detection. 3D geometric understanding is a key for robots to perform actions, , such as grasping and manipulation, under the challenge of dense and noisy environments, especially with unseen objects. In our work published in CoRL 2020 [5], we studied the grasp detection task on 3D point clouds, where we aim to directly estimate 6-DoF grasp poses to pick up objects on the plate. I believe that it is possible for machines to recognize some parts of objects with certain geometric patterns that are easy to grasp using 3D deep learning, even for novel objects. As a result, we proposed a novel coarse-to-fine 6-DoF grasp representation that 3D neural networks can directly regress. Therefore, the method can predict multiple diverse and robust grasp poses in one single inference, which is 20 times faster than previous methods requiring multi-stage inference.

Monocular floor plan reconstruction. Floor plans are high-level abstractions for indoor maps that are useful in many robotic tasks, such as indoor navigation. In contrast to most of the previous works, which rely on reconstructed 3D point clouds as inputs, our goal was to sequentially estimate the floor plan of the indoor scene using monocular 360-images. Leveraging camera pose estimation from monocular visual SLAM and 360-layout estimation approaches, we proposed a novel algorithm reconstructing the floor plant through aggregating and aligning multiple 360-layout geometries. In addition, we showed that we could handle the challenges of identifying room instances using only geometric information (e.g., camera poses and 360-layout estimation). This work [6] was submitted to the IEEE Robotics and Automation Letters.

3 Future Work

In my past research experience, I studied semantic scene understanding and geometric scene understanding for various 3D computer vision tasks, considering the two topics independently. In the next step, I would like to explore combining both semantic and geometric understanding for better robotic perception systems. On the other hand, inspired by my attempts for data-efficient learning, I would like to study enabling the generalization ability of the vision system of autonomous robots in new environments with minimal human supervision. Moreover, since accurately modeling the 3D environment is crucial, I am also interested in indoor 3D reconstruction and addressing the current challenges in this field, such as efficiency and scalability for large-scale environments. Lastly, since my ultimate goal is to improve perception and the understanding of the 3D world for robots, I would like to investigate the robotic application tasks, such as visual navigation for embodied robots, to understand the real needs and limitation of a robotic perceptual system thoroughly.

Integration of geometric and semantic information in SLAM and 3D reconstruction systems. Most current 3D semantic scene understanding approaches rely on reconstructed 3D scenes as inputs. However, it is crucial to run at a per-frame rate for real-world applications. Besides, incorporating the semantic information may improve existing SLAM and 3D reconstruction algorithms by building semantic-meaningful maps. For example, it could help long-term localization tasks since semantic landmarks are robust to weather and lighting. Also, the system could become more interpretable for humans compared to traditional ones based on key features. Furthermore, we may also determine how likely the landmark will stay in the same place as an uncertainty prior based on semantic information (e.g., object classes). For example, in a dynamic environment, objects on the table could be easily moved around, but large furniture or building structure is more likely to stay in the same place for a long time.

Several works have shown promising results in this direction: SceneCode [7] builds dense and consistent semantic maps in a visual SLAM system by optimizing learned semantic and geometric latent codes; Fusion++ [8] demonstrates using object instances as the memory-efficient map representation for SLAM; Panoptic 3D scene reconstruction [9] shows that jointly estimating the semantic instances and geometric information is favorable against considering each task independently.

Self-supervised learning and learning by interaction. Most computer vision algorithms were based on passive perception (i.e., learned from offline datasets). In contrast, I would like to explore learning visual representation or scene understanding through actively interacting with the physical world. I believe that learning from interactions, such as manipulation or navigation, can enhance the generalization ability of autonomous robots for unseen objects or new environments with minimal human supervision. For example, [10, 11] learn visual representations through manipulation and show that the representation could further assist in performing various downstream actions; Chaplot et al. [12] demonstrate self-supervised learning for scene understanding by navigating embodied agents in the indoor environment. Furthermore, through interaction, a robot may actively perceive the world and improve the robustness. For example, if a machine fails to recognize an object, a robotic arm equipped with cameras can move around and attempt to capture different views of the object.

Fast generalization and hierarchical implicit 3D representation for SLAM and 3D reconstruction. Implicit neural representation has been widely used for novel view-synthesis and 3D reconstruction tasks [13, 14]. Recently, Sucar et al. [15] has shown great success in building a visual SLAM system using implicit 3D representation as maps. However, implicit representations or NeRF-like algorithms require lengthy training for every new scene since the MLP needs to be trained from scratch. In addition, as

shown in Peng et al. [16], implicit neural representations are limited to small objects or scenes and can hardly handle large-scale scenes (e.g., a building with multiple rooms). Therefore, I believe that a potential direction is to construct hierarchical (implicit or hybrid) representations for 3D scene reconstruction or SLAM, which can maintain the coarse structure and local details simultaneously (e.g., [17]). This direction is also related to the problem of scalable and memory-efficient map representations for visual SLAM. On the other hand, the method should be able to generalize to new scenes efficiently.

References

[1] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *arXiv preprint arXiv:1706.02413*, 2017.

[2] H.-Y. Chiang, Y.-L. Lin, Y.-C. Liu, and W. H. Hsu, "A unified point-based framework for 3d segmentation," in *2019 International Conference on 3D Vision (3DV)*. IEEE, 2019, pp. 155–163.

[3] Y.-C. Liu, Y.-K. Huang, H.-Y. Chiang, H.-T. Su, Z.-Y. Liu, C.-T. Chen, C.-Y. Tseng, and W. H. Hsu, "Learning from 2d: Pixel-to-point knowledge transfer for 3d pretraining," *arXiv preprint arXiv:2104.04687*, 2021.

[4] T.-H. Wu, Y.-C. Liu, Y.-K. Huang, H.-Y. Lee, H.-T. Su, P.-C. Huang, and W. H. Hsu, "Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 510–15 519.

[5] K.-Y. Jeng, Y.-C. Liu, Z. Y. Liu, J.-W. Wang, Y.-L. Chang, H.-T. Su, and W. H. Hsu, "Gdn: A coarse-to-fine (c2f) representation for end-to-end 6-dof grasp detection," *arXiv preprint arXiv:2010.10695*, 2020.

[6] B. Solarte, Y.-C. Liu, C.-H. Wu, Y.-H. Tsai, and M. Sun, "360-dfpe: Leveraging monocular 360-layouts for direct floor plan estimation," *arXiv preprint arXiv:2112.06180*, 2021.

[7] S. Zhi, M. Bloesch, S. Leutenegger, and A. J. Davison, "Scenecode: Monocular dense semantic reconstruction using learned encoded scene representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 776–11 785.

[8] J. McCormac, R. Clark, M. Bloesch, A. Davison, and S. Leutenegger, "Fusion++: Volumetric object-level slam," in *2018 international conference on 3D vision (3DV)*. IEEE, 2018, pp. 32–41.

[9] M. Dahnert, J. Hou, M. Nießner, and A. Dai, "Panoptic 3d scene reconstruction from a single rgb image," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[10] E. Jang, C. Devin, V. Vanhoucke, and S. Levine, "Grasp2vec: Learning object representations from self-supervised grasping," *arXiv preprint arXiv:1811.06964*, 2018.

[11] Z. Xu, Z. He, J. Wu, and S. Song, "Learning 3d dynamic scene representations for robot manipulation," *arXiv preprint arXiv:2011.01968*, 2020.

[12] D. S. Chaplot, M. Dalal, S. Gupta, J. Malik, and R. R. Salakhutdinov, "Seal: Self-supervised embodied active learning using exploration and 3d consistency," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[13] D. Azinović, R. Martin-Brualla, D. B. Goldman, M. Nießner, and J. Thies, "Neural rgb-d surface reconstruction," *arXiv preprint arXiv:2104.04532*, 2021.

[14] A. Bozic, P. Palafox, J. Thies, A. Dai, and M. Nießner, "Transformerfusion: Monocular rgb scene reconstruction using transformers," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[15] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison, "imap: Implicit mapping and positioning in real-time," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6229–6238.

[16] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger, "Convolutional occupancy networks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 523–540.

[17] C. Jiang, A. Sud, A. Makadia, J. Huang, M. Nießner, T. Funkhouser *et al.*, "Local implicit grid representations for 3d scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6001–6010.