YUE WANG RESEARCH STATEMENT

In recent years, several trends have contributed to the renaissance of robotics hardware. However, 3D visual perception still fails to meet the demands of robotics. There are at least four factors leading to this gap. First, current computer vision algorithms have been largely targeted to 2D images. These algorithms are extremely good at classifying or recognizing objects in an image, but they fail to reason about the physical 3D environment. For example, computer vision algorithms typically struggle to estimate the 3D locations and orientations of the objects from 2D inputs. Second, several 3D computer vision algorithms focus on studying synthetic data and do not generalize to real-world scenarios; when deployed to real applications, many of them fail due to the domain gap. Third, objects can be invisible in a single modality or a single view. Current robotics perception modules take only RGB or RGBD images as input, neglecting other modalities. Finally, the rapid progress in computer vision is largely due to ImageNet and other image datasets. These images are mostly sourced from the internet and do not reflect the real-world image distribution. This discrepancy creates a serious generalization issue when models are adapted to real applications.

I aim to build reliable 3D deep learning algorithms and to enable human-level perception for robotics. Specifically, I identify three key insights to close the gap: (1) Perception algorithms must operate in 3D environments, leading to an end-to-end pipeline that includes minimal human intervention. (2) Multi-modality representations are the key to equip robots with the ability to see through occlusion and clutter and reason about the physics. (3) 3D data is hard to acquire from both the internet and the real-world, so we have to design algorithms that learn 3D representations with minimal supervision. To tackle these mutually-reinforced challenges, I have been conducting research by baking these insights into each project. Below, I elaborate efforts I have taken and/or plan to take in each direction.

1. Moving from 2D to 3D: Geometric Deep Learning. Most computer vision algorithms focus on recognizing objects on 2D images and often fail due to occlusion and view changes. These algorithms are inherently constrained by the information loss when modeling 3D physical world with 2D representations. To that end, we must design algorithms that consume 3D data, e.g., point clouds or meshes. Moreover, machine learning algorithms designed for 2D cannot generalize to point clouds or meshes even with significant changes. This representation difference calls for designing novel deep learning operators to consume 3D data.

In my first project as a PhD student, I introduced a new deep learning framework for point clouds, dubbed Dynamic Graph CNN (DGCNN) [1]. DGCNN constructs a k-NN graph based on the embeddings of point clouds; each point is connected to its neighbors within a certain range. Graph convolutional networks are employed to learn features from the point clouds; then, features among each local neighborhood are aggregated by symmetric functions. The same operations are stacked sequentially to build a hierarchical representation of input point clouds. DGCNN achieves state-of-the-art performance in several point cloud recognition benchmarks. In addition, DGCNN learns to label semantic parts without explicit supervision, shedding lights into 3D self-supervised learning. DGCNN is one of the most popular papers in ACM TOG (6th entry of downloads among all TOG papers as of 11/28/2021). Furthermore, DGCNN motivates a branch of learning-based methods on high energy physics and dominates particle tagging [2].

In addition to shape level recognition tasks on point clouds, I also investigate large scale scene understanding on point clouds. In our works [3, 4] on 3D object detection, we design models to directly operate on real-world point clouds captured by LiDAR sensors. Our methods are immediately applicable to autonomous driving. Furthermore, many low-cost autonomous systems rely on multi-view cameras to detect obstacles. This challenging problem boils down to how to predict 3D bounding boxes from 2D observations. In contrast to existing methods which operate in the 2D space, our method (DETR3D [5]) operates in the 3D space and query features from 2D observations; it relates 3D bounding boxes with their 2D projections in the multi-view images via camera matrices. Our method predicts a sparse set of bounding boxes and do not require any post-processing steps such as non-maximum suppression. I am currently collaborating with Toyota Research Institute and Li Auto to transfer our algorithms into production.

Beyond high-level recognition tasks, geometric deep learning should be able to support low-level visual tasks. Point cloud registration is a key problem for computer vision applied to robotics, medical imaging, and other applications. This problem involves finding a rigid transformation from one point cloud into another so that they align. Iterative Closest Point (ICP) [6] and variants [7, 8, 9, 10] have been widely used for registration. However, due to the non-convexity of its objective function, ICP is prone to local minima, reducing its generalizability and/or practicality. To tackle the non-convexity of rigid transformation objective, we proposed a simple, flexible, and general framework titled Deep Closet Point (DCP) [11] (and its follow-up PRNet [12]), which learns priors from data. DCP uses DGCNN and Transformers [13] to encode two point clouds, from which pointwise features are obtained. Then, a point-to-point matching can be approximated by dot products of point features. Finally, with the neural network empowered matching function, a singular value decomposition (SVD) extracts the rigid transformation. DCP outperforms existing optimization and/or learning-based methods by a significant margin. The work was presented at the International Conference on Computer Vision (ICCV 2019).

2. Learning with Redundancies: Multi-modality 3D Models. Our experience of the world is multi-modal – we see objects, hear sounds, feel textures, smell odors, and taste flavors; similarly the embodied intelligence should interpret such multi-modal signals together. Most robotics perception algorithms only take RGB or RGBD images, often suffering from object occlusion and view changes. To provide more redundancies, I have been studying multi-modality 3D deep learning. In [4], I introduced an object detection model that consumes both point clouds and virtual range images. Our insight is that certain objects are hard to detect in one modality but can be easily recognized in another. By fusing information from multiple sources, the model can make sense of the environment from a holistic perspective. This model achieves state-of-the-art performance on the largest available Waymo Open Dataset [14] as of publication.

I am continuing to investigate how to use multi-modality information more effectively. In an extension work [15], we use knowledge distillation to flow information between modalities. In particular, we design a single-modality student model to mimic the behavior of a multi-modality teacher model. As a consequence, the student model implicitly learns a generative model of the multi-modal world even without having complete signals. This student model achieves comparable performance with significantly lower overhead and is more suitable for production. Moreover, in [3] we also distill information from observations in the future to the current moment to enable future forecasting across modalities.

3. Learning without Annotations: 3D Self-supervised Learning. Often, the amount of unlabeled data for training a deep learning model is far larger than the amount of labeled data. Specifically, in the 3D case, data labeling is very costly. Hence, using large amounts of 3D unlabeled data is a central problem in both academia and industry. I have been investigating 3D self-supervised learning from two perspectives: (1) general representations can transfer to tasks where data is limited, so we can learn general representations from largely available 3D data with a proxy training objective; (2) for specific tasks where annotations are extremely challenging to acquire, we can look for their associative tasks and train the model with implicit supervision.

In recent ongoing work, I am developing learning representations by maximizing the mutual information between point clouds and images, which are captured at the same time. These two modalities are encoded with neural networks; the embeddings of the two modalities are attracted if they belong to the same scene and otherwise repelled. By pre-training on this instance discrimination task, these encoders are discovering semantics jointly without using any annotations. Once the pre-training is finished, the encoding networks (with potentially additional sub-networks) can be fine-tuned on labeled data with a task-specific training objective.

In addition to learning using a large amount of unlabeled data, I investigated how to meta-learn general models on a extremely small amount of data. The machine learning community has seen increasing interest in meta-learning. Most (if not all) approaches involve fast adaptation algorithms and ignore the role of representations. In our recent ECCV paper RFS [16], we demonstrated that a simple pre-trained embedding model can outperform complicated meta-learning algorithms. This finding sheds lights into meta-learning research – models and algorithms are equally important.

Still, for a specific task that is distant from the pre-training objective, a general representation is not sufficient to solve the problem. For example, predicting dense correspondence between shapes is independent of attracting a point cloud embedding and an image embedding. Therefore, we must design a point cloud network that can learn without using explicit correspondence supervision. In our recent work [12], I propose using registration as a proxy task to learn keypoint detection and correspondence. We use a pipeline similar to DCP [11] to encode two point clouds; set correspondence in the latent space; use the correspondence to predict a transformation. As a consequence, the model is driven to predict correspondences without trained on the annotated data.

Research Impact

My research on point cloud processing was recognized by the Nvidia Graduate Fellowship (five recipients per year worldwide) and was named the first place recipient of the William A. Martin Master's Thesis Award for 2021. DGCNN [1] was (arguably) the most popular paper in ACM TOG 2019 and has received over 1,800 citations since its appearance in 2019. DGCNN has motivated a wide range of research and publications across deep learning, computer vision, and high energy physics [2]. DCP [11], which was the first paper to combine deep learning with ICP, has been widely adopted in point cloud registration. This work provides a completely new perspective to address rigid alignment problems. It also inspired other works that incorporated optimization procedures into deep learning architectures.

My recent projects on 3D object detection are immediately applicable to industry. Pillar-OD [4] has been deployed in Waymo's production platform. Object DGCNN [3] is under technology transfer via a collaboration with Toyota Research Institute. DETR3D [5], at the same time, enables low-cost autonomous driving for electronic vehicles including Tesla and Li Auto.

In addition, RFS [16] leads a novel direction to study "learning to learn" from a representation perspective. It provides the community with a competitive baseline model. Also, it motivates a rethinking of few-shot learning benchmarks and the associated role of meta-learning algorithms.

Future Plans

My research aims to enable embodied agents to understand the 3D world. My general approach towards that goal is to build reproducible research from perception and control to 3D content generation on real platforms – algorithms have to be demonstrated, reproduced, and deployed on customer-available robots such as self-driving cars, indoor robots, and/or AR/VR devices. In addition to current projects, I plan to work on the directions as stated below.

End-to-end Autonomous Driving. Autonomous driving is an integrated showcase of robotic perception. However, no research platform for autonomous driving is publicly available; this scenario significantly slows down the progress in autonomous driving research. In the first three years of my career, I will assemble a team to build such a research platform. This platform will include a physical simulator and a real car, which enable both simulation-to-real transfer learning and demonstrations in real scenarios. I will focus on multi-modal perception and 3D self-supervised learning to enable human level perception for autonomous driving with minimal supervisions. In addition to perception, I will investigate learning-based planning and control, which are much unexplored. My goal is to enable autonomous driving from perception to control with machine learning.

Furthermore, I strive to open-source the whole project to the community, as open-sourcing is crucially important to computer science development. Autonomous driving is a real-world problem, and we should take inputs from industry. To that end, I plan to pursue joint research with industrial research labs to facilitate collaborations between academia and industry. In addition, I will study ethical issues to develop safe, reliable, and equitable autonomous systems.

Indoor Robotics. Beyond outdoor robots such as self-driving cars, I intend to work on indoor scene understanding and robots. My goal is to design a configurable robot that consists of affordable chips and sensors; we will release the configuration and make our works on this robot reproducible. I will enable fast prototyping of research ideas on this robot, and my research focus will include three topics: learning-based localization and mapping (SLAM), 3D meta-learning under different indoor setups, and 3D scene understanding. In addition to conducting research on this robot, I plan to extend it to be a teaching platform for robotics. I expect to collaborate with other robotics professors and co-design a robotics perception class.

3D Content Generation. Thanks to recent developments in neural representations [17] and generative adversarial networks [18], content creation has significantly advanced. However, current methods still fail the need of robotic applications. To enable end-to-end 3D visual computing, I would like to explore neural rendering and 3D content generation. Combined with the indoor robot perception system, this can serve as a real-time content generator to provide realistic VR/AR experiences. Furthermore, I plan to setup a light stage to capture human dynamics and object appearance under various geometries and lights, which can be used as training data for neural rendering . I will release the data constantly and invite researchers to study generative modeling for 3D content creation.

References

[1] Yue Wang, Yongbin Sun, Sanjay E. Sarma Ziwei Liu, Michael M. Bronstein, and Justin M. Solomon. Dynamic Graph CNN for Learning on Point Clouds. *ACM Transactions on Graphics (TOG)*, 2019.

[2] Huilin Qu and Loukas Gouskos. Particlenet: Jet tagging via particle clouds. *CoRR*, abs/1902.08570, 2019.

[3] Yue Wang and Justin M. Solomon. Object dgcnn: 3d object detection using dynamic graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[4] Yue Wang, Alireza Fathi, Abhijit Kundu, David Ross, Caroline Pantofaru, Thomas Funkhouser, and Justin Solomon. Pillar-based object detection for autonomous driving. In *The European Conference on Computer Vision (ECCV)*, 2020.

[5] Yue Wang, Vitor Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, , and Justin M. Solomon. DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries. In *The Conference on Robot Learning (CoRL)*, 2021.

[6] Paul J. Besl and Neil D. McKay. A Method for Registration of 3-D Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 14(2):239–256, February 1992.

[7] Szymon Rusinkiewicz and Marc Levoy. Efficient Variants of the ICP Algorithm. In *Third International Conference on 3D Digital Imaging and Modeling (3DIM)*, June 2001.

[8] Aleksandr Segal, Dirk Hähnel, and Sebastian Thrun. Generalized-ICP. In Jeff Trinkle, Yoky Matsuoka, and José A. Castellanos, editors, *Robotics: Science and Systems*. The MIT Press, 2009.

[9] Sofien Bouaziz, Andrea Tagliasacchi, and Mark Pauly. Sparse Iterative Closest Point. In *Proceedings of the Eleventh Eurographics/ACMSIGGRAPH Symposium on Geometry Processing*, 2013.

[10] François Pomerleau, Francis Colas, and Roland Siegwart. A Review of Point Cloud Registration Algorithms for Mobile Robotics. *Foundations and Trends in Robotics*, 4(1):1–104, May 2015.

[11] Yue Wang and Justin M. Solomon. Deep closest point: Learning representations for point cloud registration. In *The International Conference on Computer Vision (ICCV)*, 2019.

[12] Yue Wang and Justin M. Solomon. Prnet: Self-supervised learning for partial-to-partial registration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.

[14] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tak-wing Tsui, James C. Y. Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jon Shlens, Zhi-Feng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] Yue Wang, Alireza Fathi, Jiajun Wu, Thomas A. Funkhouser, and Justin M. Solomon. Multi-frame to single-frame: Knowledge distillation for 3d object detection. In *The Workshop on Perception for Autonomous Driving at the European Conference on Computer Vision*, 2020.

[16] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: A good embedding is all you need? In *The European Conference on Computer Vision (ECCV)*, 2020.

[17] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *The European Conference on Computer Vision (ECCV)*, 2020.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2014.