Research Experience

My academic journey began with research on medical image segmentation at Sun Yat-sen University (SYSU), under the supervision of Prof. Heye Zhang. As an undergraduate, I focused on nasopharyngeal carcinoma segmentation of CT and MR images. I explored the potential of densely connected convolutional networks to enhance segmentation IoU, culminating in a publication in Sensors [12]. During my master's at HKUST and a research internship at SmartMore, I expanded my focus to general 3D vision tasks, particularly shape-from-images. My primary research involved 3D reconstruction of reflective objects, a task with significant industrial demand and recognized as a long-standing challenge in multi-view 3D Reconstruction. I proposed using polarization information to resolve the ambiguity of reflective surface normals, thereby improving the accuracy of reconstructed geometry. This work resulted in a holistic neural 3D reconstruction pipeline for reflective objects, accepted to ICLR 2024 [24]. Additionally, I investigated the multi-modal generation capabilities of Large Language Models (LLMs) in collaboration with Tencent [25]. Currently, my main interest lies in 3D Vision, and I am working on online feed-forward Gaussian splatting at Microsoft Research Asia (MSRA).

Research Vision

My long-term goal is to develop a versatile feed-forward 3D vision system utilizing large-scale, multi-sensory 3D data gathered from a variety of user scenarios. I believe that the field of 3D Vision with Neural Networks is at a pivotal moment, transitioning from 2D rendering loss to primitive 3D supervision. This shift is driven by the following insights:

Bottleneck of 2D rendering supervision. From NeRFs [15] to Gaussian Splatting [10], the series cam be summarized as learning 3D representations from 2D signals. It circumvents the insufficiency of 3D data, making training with large-scale 2D images feasible. However, performance is often hindered by ambiguities inherent in 2D observations [5, 6, 14, 27]. Additionally, these optimization-based representations tend to overfit specific scenes, reducing their generalizability [1] and limiting their application in data generation [9]. For instance, integrating NeRFs or Gaussian splatting into a standard feed-forward detection network is challenging due to inner optimization-based reconstruction loops. Recent research has explored the use of hyper-networks to directly predict the neural weights of NeRFs [17, 18]. However, training these hyper-networks is challenging, and they are currently handling simple, single objects.

Strengthens of primitive 3D supervision and potential for real-world data. Recent advancements in 3D object generation have showcased impressive high-fidelity details by replacing volume rendering loss with explicit 3D losses [11, 23, 28]. This supervision helps resolve the ambiguities inherent in 2D images. Moreover, feed-forward networks trained on large **synthetic** 3D object datasets like Objaverse & G-Objaverse [3, 16] can generate 3D objects in a single forward pass without optimization [7, 8, 20]. Such philosophy should extend to general scenes. Actually, DUSt3R [22], supervised by point clouds reconstructed from in-the-wild images, has shown significant accuracy improvements and facilitated many downstream tasks like pose-free 3D reconstruction [4]. However, its broader applicability is limited due to **the lack of 3D data from diverse user scenarios** for training. Synthetic 3D objects can be rendered through Blender effortlessly, with accessible ground truth attributes including materials, meshes and cameras. In contrast, constructing a real 3D dataset encounters collection cost and alignment issues. Depending on user scenarios, data from diverse modalities is collected using heterogeneous sensors, including LiDAR in autonomous driving [19], radio waves in digital health [13], and ultrasound in remote sensing [21]. To bring these methods into piratical applications, it is crucial to fuse and align interdisciplinary real-world data, on a scale commensurate with Objaverse, collected from various sensors [19]. The trained backbone model, serving as a strong prior, should be easily adaptable to various downstream tasks.

Open problems. Despite the feasibility discussed, several open problems remain unresolved based on my research at MSRA. Firslty, existing feed-forward (or generalizable, as referred to in some papers [26]) 3D vision models, including pixelNeRF [26], pixelSplat [2], and Large Gaussian Model [20], only support one or two views as inputs. More views require optimization-based global alignment like DUSt3R [22]. Secondly, most of these models are object-centric or foreground-biased, which deteriorates scene perception performance. Additionally, 3D representations in these studies are often assumed to be pixel-aligned, meaning each pixel in an image corresponds to one 3D point or Gaussian. This results in redundant points in general scenes due to overlaps among views and may cause view-inconsistent predictions. Our research on aligning and merging 3D Gaussians across views has shown significant improvements. Ultimately, current methods are trained solely on point clouds or depth maps. I propose that collecting multi-sensory 3D data from multiple devices deployed in diverse user scenarios would enhance performance and generalizability.

References

[1] Aayush Bansal, Xinlei Chen, Bryan Russell, Abhinav Gupta, and Deva Ramanan. Pixelnet: Representation of the pixels, by the pixels, and for the pixels. *arXiv preprint arXiv:1702.06506*, 2017. 1

[2] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19467, 2024. 1

[3] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 1

[4] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*, 2024. 1

[5] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 1

[6] Wenhang Ge, Tao Hu, Haoyu Zhao, Shu Liu, and Ying-Cong Chen. Ref-neus: Ambiguity-reduced neural implicit surface learning for multi-view reconstruction with reflection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4251–4260, 2023. 1

[7] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Liu, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction, 2024. 1

[8] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 1

[9] Nan Huang, Xiaobao Wei, Wenzhao Zheng, Pengju An, Ming Lu, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. S3 gaussian: Self-supervised street gaussians for autonomous driving. *arXiv preprint arXiv:2405.20323*, 2024. 1

[10] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[11] Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024. 1

[12] Yang Li, Guanghui Han, and Xiujian Liu. Dcnet: Densely connected deep convolutional encoder–decoder network for nasopharyngeal carcinoma segmentation. *Sensors*, 21(23):7877, 2021. 1

[13] Yingcheng Liu, Guo Zhang, Christopher G Tarolli, Rumen Hristov, Stella Jensen-Roberts, Emma M Waddell, Taylor L Myers, Meghan E Pawlik, Julia M Soto, Renee M Wilson, et al. Monitoring gait at home with radio waves in parkinson's disease: A marker of severity, progression, and medication response. *Science Translational Medicine*, 14(663):eadc9669, 2022. 1

[14] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7210–7219, 2021. 1

[15] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1

[16] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9914–9925, 2024. 1

[17] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 1

[18] Bipasha Sen, Gaurav Singh, Aditya Agarwal, Rohith Agaram, Madhava Krishna, and Srinath Sridhar. Hyp-nerf: Learning improved nerf priors using a hypernetwork. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[19] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 1

[20] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 1

[21] Lei Wang, Haoran Wan, Ting Zhao, Ke Sun, Shuyu Shi, Haipeng Dai, Guihai Chen, Haodong Liu, and Wei Wang. Scalar: Self-calibrated acoustic ranging for distributed mobile devices. *IEEE Transactions on Mobile Computing*, 23(2):1701–1716, 2023. 1

[22] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 1

[23] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4563–4573, 2023. 1

[24] LI Yang, WU Ruizheng, LI Jiyong, and CHEN Ying-cong. Gnerp: Gaussian-guided neural reconstruction of reflective objects with noisy polarization priors. *arXiv preprint arXiv:2403.11899*, 2024. 1

[25] Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. Seed-story: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683*, 2024. 1

[26] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 1

[27] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 1

[28] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 1