As humans, our natural ability to comprehend the surroundings allows us to navigate, interact, and make decisions effortlessly. In contrast, understanding 3D environments remains challenging for AI systems, particularly in dynamic and complex real-world scenarios. These challenges of 3D scene understanding inspire my passion for enabling machines to effectively navigate, manipulate objects, and interact with humans. Achieving this requires models to have a fundamental knowledge, having an ability to detect and recognize objects within dense 3D scenes. Currently, I am focusing on leveraging vision-language models to achieve a comprehensive understanding of 3D scenes through multi-view RGB images, point clouds, and implicit representations. My research aims to enhance machines' ability to understand and interact with the physical world using natural language. These advancements will pave the way for applications in areas such as autonomous driving, robotics, and augmented reality. To achieve this, I aspire to become an independent researcher who can collaborate effectively with peers and leading experts to drive impactful advancements in AI and computer vision. Pursuing a Ph.D. in Computer Science at **University Name** is a pivotal step in my journey as it provides an ideal environment to deepen my expertise and further enhance my research capabilities.

Background: Graduating with the Best Thesis Award from VNUHCM - University of Information Technology, I built a robust foundation in mathematics, machine learning, and computer vision, accompanied by four published research papers. Furthermore, I had two years as an AI Resident at VinAI Research - recognized as one of the world's top 20 companies in AI research in 2022 - I collaborated closely with leading researchers, expanding my expertise in developing AI algorithms for advanced 3D perception and computer vision applications. Through rigorous coursework and hands-on experience, I have strengthened my knowledge and skills in essential tools to conduct impactful research in computer vision.

Research Experience: During my undergraduate studies, I was captivated by the rapid evolution of object detection algorithms. This early exposure inspired me to first-author my first research paper [2] as a sophomore, paving the way for three more published works [9, 3, 8] centered on (oriented) object detection. These works were a formative experience that shaped my academic attitude and sparked my passion for conducting high-quality academic research.

During my junior year as an undergraduate at VNUHCM-UIT, I joined VinAI Research to further develop my expertise in language-driven understanding. My passion for 3D vision was sparked when I participated in the **1st Workshop on Open-Vocabulary 3D Scene Understanding** [1] at ICCV 2023, where I proposed an idea of integrating foundation model into 3D scenes and implemented an end-to-end pipeline, from data preprocessing to model evaluation, leading the team to ultimately achieve the second place. This experience fueled my interest in 3D scene understanding, as traditional methods face limitations in scaling to real-world scenarios with vast category diversity, often relying on extensive, costly annotations. This journey has led me to explore Open-Vocabulary 3D Instance Segmentation, pushing the boundaries of scalable 3D perception.

Existing 3D scene understanding methods often struggle to identify small and rare objects, such as light switches, door knobs, and window locks — due to limitations in training data and the dense nature of 3D scene captures. To address these challenges, I led the design and the development of **Open3DIS** [6], which integrates 2D information with supervised 3D features, effectively aggregating 2D masks across frames. By leveraging strong semantic cues, Open3DIS achieves state-of-the-art performance across four open-vocabulary benchmarks, with 1.5x enhancement in AP score compared to prior methods. The paper, in which I was the **first author**, was accepted to CVPR 2024. This work was so remarkable that my team later achieved the first place in the **2nd Workshop on Open-Vocabulary 3D Scene Understanding** at CVPR 2024. Open3DIS has garnered substantial recognition within the 3D perception community as it provides a fine-grained level of 3D instance segmentation, significantly advancing object recognition capabilities in robots navigation and autonomous systems.

Furthermore, I recognized a significant challenge in Open-Vocabulary 3D scene understanding: the need for predefined class names during testing, which requires human intervention and hinders the autonomy of scene perception in AI agents. While one approach is to establish extensive vocabularies, this often leads to inaccuracies and hampers performance due to the overwhelming number of categories. To address these issues, we introduced a new task, **Open-Ended 3D Point Cloud Instance Segmentation** [7], leveraging Multimodal-LLM to aggregate 3D visual tokens onto a 3D point cloud, generating unconstrained class names for 3D objects, readily for real-world language interaction with humans. This results in a paper in which I am the **first author**. The paper is still being under-reviewed at a top-tier conference.

My most recent work tackled several limitations of previous approaches, including Open3DIS and Open-Ended 3D Point Cloud Instance Segmentation. Earlier methods segmented each frame of a video independently, then combined them using unsupervised merging heuristics, leading to an over-segmented and inconsistent representation of an object. In contrast, **Any3DIS** [5] offers robust 3D instance segmentation across all objects in a 3D scene by leveraging the proposed 3D-Aware Segment-AnythingV2 model and Mask Optimization module. Our method achieves state-of-the-art performance on two challenging indoor datasets, significantly surpassing the current state-of-the-art Open3DIS's performance by 30% on AP score and 10 times faster. I am also the **first author** of this work, and it is currently under review at a top-tier conference. These accomplishments have solidified my passion for advancing 3D understanding and inspired me to pursue graduate studies to further contribute to this field.

Alongside my research in 3D computer vision, building on my aspiration to contribute to my home country, I had a privilege to collaborate with Vietnam's Ministry of Information and Communications in the development **VinMap** [4] project with goals to advance public knowledge, bolster cybersecurity for the Vietnamese government, and pioneer future research in map analysis. This project received strong support from the Vietnamese public, capturing the interest of researchers nationwide and marking my very first contribution to the country's development.

Future Research: My research focuses on attaining a fine-grained understanding of every object in a 3D environment, a crucial step toward enabling machines and robots to perceive the world as humans do. I believe that AI development parallels human development; over time, AI agents equipped with a robust foundational understanding at the most detailed level of their surroundings will excel in navigating and interpreting complex, dense 3D environments. However, significant challenges remain due to the limited availability of high-quality 3D training data and the constraints of current 3D representations. I aim to leverage self-supervised learning techniques or vision-language models to address data limitations, driven by the goal of developing AI agents that can **extract fine-grained semantic information from any 3D scene representation** or **explore new, more expressive, and meaningful representations** to enhance object recognition and interaction capabilities in diverse and arbitrary 3D environments. Furthermore, I aspire to advance our understanding of how robots physically interact with complex 3D environments by effectively **fusing information from multiple modalities**, enabling reliable, and robust 3D reasoning.

I would be thrilled to pursue a Ph.D. at **University Name**, where several faculty members' research aligns closely with my long-term goals. Prof. **X**'s research on developing robust strategies for 3D scene understanding in open-world scenarios closely aligns with my interest in enabling machines to perceive the world as humans do. Prof. **Y**'s work on 2D perception, especially video understanding resonates with my aspirations to study fine-grained interactions between humans and objects in realistic contexts. Prof. **Z**'s investigations into robotic movement and interaction with arbitrary environments strongly complement my goal of enabling machines to learn in ways akin to human behavior. The faculty and research centers at **University Name** create an exceptional environment for advancing my research aspirations. I am confident that the Doctoral Program in Computer Science at **University Name** will provide the ideal community for me to develop my expertise, cultivate my skills and achieve success as an independent researcher in academic and industry.

References

[1] F. Engelmann, A. Takmaz, J. Schult, E. Fedele, J. Wald, S. Peng, X. Wang, O. Litany, S. Tang, F. Tombari, et al. Opensun3d: 1st workshop challenge on open-vocabulary 3d scene understanding. *arXiv preprint arXiv:2402.15321*, 2024.

[2] P. Nguyen, L. Ngo, T. Truong, T. T. Nguyen, N. D. Vo, and K. Nguyen. Page object detection with yolof. In *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 205–210. IEEE, 2021.

[3] P. Nguyen, T. Truong, N. D. Vo, and K. Nguyen. Rethinking classification of oriented object detection in aerial images. *International Journal of Advanced Computer Science and Applications*, 13(9), 2022.

[4] P. Nguyen, A. Do, and M. Hoai. Detecting omissions in geographic maps through computer vision. In *2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6. IEEE, 2024.

[5] P. Nguyen, M. Luu, A. Tran, C. Pham, and K. Nguyen. Any3dis: Class-agnostic 3d instance segmentation by 2d mask tracking. *arXiv preprint arXiv:2411.16183*, 2024.

[6] P. Nguyen, T. D. Ngo, E. Kalogerakis, C. Gan, A. Tran, C. Pham, and K. Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4018–4028, 2024.

[7] P. D. Nguyen, M. Luu, A. Tran, C. Pham, and K. Nguyen. Open-ended 3d point cloud instance segmentation. *arXiv preprint arXiv:2408.11747*, 2024.

[8] N. D. Vo, P. Nguyen, T. Truong, L. Ngo, K. Huynh, D. Dinh, and K. Nguyen. Analysis of fog density on oriented object detection in aerial images. In *2022 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–6. IEEE, 2022.

[9] N. D. Vo, P. Nguyen, T. Truong, H. C. Nguyen, and K. Nguyen. Foggy-dota: An adverse weather dataset for object detection in aerial images. In *2022 9th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 269–274. IEEE, 2022.