Wenbo Hu

My interest in computer vision began during my sophomore year. When I participated in a school project on wildfire detection in California, I was introduced to Convolutional Neural Network (CNN) by my team leader and the study material is CS231N by Fei-Fei Li. Therefore, Stanford became the only dream school that I will think of when mentioning computer vision. Almost immediately, I was astonished by the power of CNN in distinguishing images with high accuracy through a trivial layer-by-layer architecture. Later on, I discovered this technique could solve many practical and challenging problems encountered in many aspects of our lives, including facial recognition and yield prediction. CNN's wide application encouraged me to dive deeper into its mechanisms and underlying features and eventually kicked off my deep exploration into computer vision.

As I began my exploration, I realized that mathematical foundations were crucial to understanding computer vision, allowing me to freely implement algorithms and not stumble over research papers. I took several math classes including optimization, probability, and signal processing, which enabled me to become familiar with the concepts of MCMC, variational inference, and convex/non-convex optimization. These classes prepared me for editing neural network architectures and applying scientific research in industry. After two quarters of applying math to implement machine learning models, I undertook an internship at Synthesis Electronic Technology in the computer vision group. My job was to accelerate lightweight object detection models and apply them to industrial scenarios. Although familiar with segmentation and detection models, applying and improving the new yolov5 series was still challenging for me. To overcome this hurdle, I researched various resources to fully understand the architecture. When the Swin-Transformer was published that summer, I was inspired to edit the yolov5 architecture by applying transformers. My workmates and I started began our research into transformer networks using images as direct inputs. After reading and exploring papers for a month, we applied diverse techniques and added a transformer to our yolov5 model resulting in a 5% mAP 50 improvement in our working dataset.

Nevertheless, working on industrial problems doesn't only mean applying and improving SOTA models. In industry, what is favorable is not only a delicate model, but a sophisticated solution that best meets business requirements and ethics. During my internship, our business problem involved monitoring workers in factories and reporting violations. After discussing with my workmates, I realized that designing a model with the highest possible accuracy is not adequate and ethical. The intention was to remind workers so others can stay alert. Therefore, the model I built should minimize false positives as much as possible since we never want to disturb a diligent worker. This experience gave me a valuable lesson on how to understand a business problem. Besides implementing specific details and finalizing networks, I need to acknowledge what both my manager and the customers want. To make sure we meet business ethics, the network should not be used for pushing workers. I experimented and tested a robust bound for human "slacking off" time and kept our model confidential from others editing the accuracy threshold. Discussing and cooperating with customer needs and meeting data ethics was a great lesson for me—as opposed to just implementing algorithms and hypertuning a fancy model.

Getting familiar with computer vision applications, I fully devoted myself to research at the start of my next school year. I began following Professor Hao Su who introduced me to the exciting

world of computer vision. Working in Su's lab, I was inspired by front-end computer vision research ideas and ended up exploring reinforcement learning (RL) and generalizable manipulation skills, which are one of the cornerstones of embodied AI. In our research group, we worked on building a large-scale robotic. We built the benchmark based on a simulated part-based interactive environment (Sapien) which was published at CVPR 2020. I worked on the rigid body environment tasks and implemented them in a unified OpenAI Gym interface with fully simulated dynamic interaction, supporting multiple observation modes and multiple controllers. These tasks solidify my RL skills by interacting with agents, environments, and action space and applying dense reward functions. I also handled interface wrapper functions and guaranteed that my code was flexibly written and could be adapted to changing demands. Since we were building a benchmark and therefore all competitors would read our code, I learned how to clearly document the code and split it into separate functions, so it was clear as to what each piece of code did. We also kept our code reproducible so that others could run it themselves. Through these high standards and rigorous implementing procedures, I gained a valuable experience in this large project. I also discovered research exploration was not only about exploring trending fields but also exploring uncharted fields. This benchmark can push cutting-edge research work by addressing the critical problems encountered. From its application, we also discovered that minimizing the difference between robots in simulation environments and real performance is an interdisciplinary gap that lacks exploration. Hence, I independently began research in this field with a Ph.D. Jiayuan. We first experimented with actual robot motion planning after hand-eye calibration but failed to get the expected result. I then rendered 3D images with axes to clearly dissect my algorithm. Fixing the bugs in an unfamiliar field makes me realize how important visualization is to computer vision since each object has its frames and axes. Now, we combined our ideas from real robots to modify the training policies in the simulation environment and we will wait for our results.

Computer vision is a field that updates fast and needs both breadth and depth of field knowledge. Whenever I'm in my comfort zone, I force myself to meet more challenging problems to expand my tool set. From CNN networks, handling 3D vision data such as point clouds and meshes, and now exploring robotics/RL algorithms for "active" AI, I always seek to advance my skills. Although all domains of computer vision shape my career goals, I have a deeper interest in exploring the 3D vision world relating to objection detection, MVS, and NerF, as well as solving autonomous driving problems by implementing SOTA algorithms. I'm eager to focus on algorithm development on real-time embedded platforms, to prototype and test autonomous vehicles, and to work with others to complete functional safety analyses and productization efforts.

Stanford University has the best computer science program in the world. Here, I can learn a robust set of skills encompassing current and emerging state-of-the-art AI and computer vision topics. Stanford students are the core members that revolutionize every generation of technology and make Silicon Valley the way it is today. I will choose the specialization of Artificial intelligence, specifically computer vision, and work on its interactions with robotics. My research mentor Hao Su is the author of Fei-Fei Li's world-impacting ImageNet benchmark, and he received his Ph.D. under LJ Guibas at Stanford. Since all his teaching was conducted in the Stanford way, I strongly believe I will express my talents at Stanford. I believe the study opportunities at Stanford will become a critical component of my professional development, and hence your favorable consideration of my application will be sincerely appreciated.