Statement of Purpose

I am currently a master's student in Applied Data Science at the University of Southern California (USC), advised by Professor Laurent Itti and Professor Ram Nevatia. During my master's study, I have co-authored six papers, including one CVPR'21 paper, and submitted two first-author papers to CVPR'22. I want to pursue a Computer Science Ph.D. focusing on understanding the important human-centric properties beyond accuracy, e.g., humanity (Humanoid Neural Network), interpretability (Explainable Artificial Intelligence), and stability (Adversarial Machine Learning). I think they are crucial to understanding ML models and making them trustworthy as well as reliable.

My motivation to pursue my Ph.D. comes from my past research experience. My research aims to understand the logic of neural networks (NN), enables humans to provide guidance to the NN after identifying the root cause of the error, and gives insights into how humans perceive this world. Continuous exploration in understanding the logic of NNs excites me as I gain deeper knowledge in bridging the gap between NNs and Human intelligence for efficient knowledge exchange. Diving into theories and mathematical details behind algorithms gives me great opportunities to make fundamental improvements to machine learning models' abilities. All these valuable experiences and learnings inspired me to pursue an academic career, and continue pushing the boundary of NNs' capabilities.

To better understand how to explain the reasoning logic behind NN's predictions, I worked on Interpreting with Structural Visual Concepts at iLab, advised by Professor Laurent Itti and Doctor Ziyan Wu. This project aims to interpret models' logic by answering WHY and WHY NOT questions. For example, if the NN's prediction is "ambulance", we want to answer the questions, "Why does the model think this is an ambulance?" and "Why not a fire engine?". Most of the current methods focus on answering "Why". They try to discover the low-level correlation between input pixels and the final decision. In our work, we use structural concept graphs (SCGs) to provide human-intuitive high-level explanations. "Concept" represents the important primitive visual features for a class. For example, the "wheel" and the "police logo" are important concepts for "police vans". After extracting class-specific concepts, we would construct SCGs based on the spatial relationships between them. Extensive experiments showed that our framework could answer "why" and "why not" with faithful, logical, concept-level explanations, which can help us improve the original model's performance. I was really proud to share these findings in our paper at CVPR'21.

After understanding the logic behind model predictions, I was eager to explore decision stability: why adversarial examples can fool NNs. Thus, I joined the USC Iris Computer Vision Lab, collaborating with Professor Ram Nevatia to defend NNs against patch attacks. Compared with traditional adversarial attacks, which generate perturbations on arbitrary pixels, patch attacks manipulate a specified region of the input image to

mislead the target NN. Considering that a block of adjacent pixels can cause the misprediction of NNs, an intuitive solution is to detect these pixels and protect NNs from perceiving them. Since the attack patches tend to be highly textured and distinctive in appearance compared to natural images, I utilized an image semantic segmentor to detect such patches and to mask the adversarial pixels with the average pixel values. This enables a task-agnostic defense method compatible with multiple downstream models and can generalize to patches of any size and shape. A research paper based on this work is now under review for CVPR'22, of which I am the first author.

Working on adversarial learning made me realize the fragility of NNs, and I began to consider the differences between humans and NNs. Human decisions are much more robust, and the human vision system (HVS) is the gold standard for complex vision tasks, e.g., zero-shot learning, continual learning, and novel view imagination. To investigate the contributions of three important features (shape, texture, and color) to the HVS, I designed a humanoid vision engine (HVE) consisting of three feature extractors and a representation learning model to mimic the way of human brain processes these features respectively. Through this framework, we can compute the contribution of each feature during predictions with a gradient-based method. After getting the representations of shape, texture, and color with our feature extractors, we can also use those feature representations to simulate human abilities in imagination and open-world zero-shot learning. These results are the first step towards better understanding the contributions of object features to classification, zero-shot learning, and imagination tasks. I also submitted this work to CVPR'22 as the first author.

Throughout my one and a half years of immersion in computer vision, I have always been determined to pursue an academic career. There are three directions that I would like to endeavor in my future academic life: 1) *Humanoid Neural Network* that guides models to learn from humans and give insights into how humans perceive this world; 2) *Explainable Artificial Intelligence* that creates interfaces to help NNs and humans communicate, interact, and exchange knowledge; 3) *Adversarial Learning* that defends against adversarial (patches) attacks via original image restoration. I believe all these questions are fundamental for bridging the gap between NNs and human intelligence, and their solutions will play significant roles in implementing a human-centric AI. With all these topics in mind, I feel the need and urgency to apply for the Ph.D. program.

At XXX, there are many professors whose research groups appeal to me. Particularly, I find the work of Professor **AAA**, **BBB**, **CCC** and **DDD** exciting and relevant to my research interests. It would be my honor to be able to join XXX as a Ph.D. student. Overall, I believe XXX's abundant resources and collaborative environment can provide the best guidance to my academic career, and I would also love to give back to the XXX community with my positive spirit and ceaseless hard work.