ZHIHAO ZHANG

+1 949-981-1208 ⋄ zzh_jackfram@outlook.com

EDUCATION

Renmin University, China *2016.9 - 2020.7*
B.S. in Computer Science Overall GPA: 3.74/4.0 (5%)

- Academic achievement: Annually Academic Achievement Scholarship 2016-2018, Dean's Scholarship of RUC 2018, National Undergraduate Training Programs Scholarship for Innovation and Entrepreneurship

University of Edinburgh, UK *2018.9 - 2019.6*
Full year visiting student, major in computer science Overall GPA: 4.0/4.0
Carnegie Mellon University, USA *2020.9 - Now*
Master of Science in Robotics (MSR) Overall GPA: 4.17/4.0

Carnegie Mellon University, USA *2022.9 - 2027 (expected)*
Ph.D. in Computer Science

RESEARCH EXPERIENCES

Carnegie Mellon University, Catalyst *Pittsburgh, U.S*
Research Assistant, advised by Prof. Zhihao Jia *2021.3-now*

- Machine Learning System

Carnegie Mellon University, Intelligent Control Lab *Pittsburgh, U.S*
Research Assistant, advised by Prof. Changliu Liu *2020.9-2021.3*

- Deep learning theory related topics, eg. Neural Tangent Kernel, Rademacher Complexity, Norm Based NN Capacity Measurement.

University of California Berkeley, Mechanical Systems Control Lab *Berkeley, U.S*
Research Intern, advised by Prof. Masayoshi Tomizuka *2019.10-2020.3*

- "Social-WaGDAT: Interaction-aware Trajectory Prediction via Wasserstein Graph Double-Attention Network", an interactive trajectory prediction method using GNN framework

Carnegie Mellon University, Intelligent Control Lab *Pittsburgh, U.S*
Research Intern, advised by Prof. Changliu Liu *2019.6-2019.10*

- AutoEnv, an integrated platform for autonomous driving related tasks. Components include preprocessing, algorithm implementation(TRPO, PS-GAIL, RLS), simulation and evaluation. Now published as an open source code base v1.0 on GitHub. Link `https://github.com/JackFram/Autoenv`

RUC Multimedia and Intelligence Lab *Beijing, China*
Research Assistant, advised by Prof. Qin Jin *2018.6-2019.3*

- Visual-dialog challenge 2018, design an encoder-decoder framework incorporate attention mechanism to achieve multiple round of Q&A. Encoder is consisted of a ResNet50 for image feature extraction and LSTM for question encoding, decoder is a LSTM for answering questions.

PUBLICATIONS

- **Communication Bounds for the Distributed Expert Problem**, under review
  with Zhihao Jia, Qi Pang, David Woodruff, Wenting Zheng (alphabetic order)
- **SpecInfer: Accelerating Generative LLM Serving with Speculative Inference and Token Tree Verification**, preprint
  Xupeng Miao*, Gabriele Oliaro*, **Zhihao Zhang***, Xinhao Cheng, Zeyu Wang, Rae Ying Yee Wong, Zhuoming Chen, Daiyaan Arfeen, Reyna Abhyankar, Zhihao Jia
- **GradSign: Model Performance Inference with Theoretical Insights**, The Tenth International

Conference on Learning Representations (ICLR 2022)
Zhihao Zhang, Zhihao Jia

- **Social-WaGDAT: Interaction-aware Trajectory Prediction via Wasserstein Graph Double-Attention Network**, IEEE Transactions on Intelligent Transportation Systems (TITS)
  Jiachen Li, Hengbo Ma, **Zhihao Zhang**, Masayoshi Tomizuka

AWARDS

- Meta Research Award, 2022
- Google Faculty Research Award, 2022