Haoyi Zhu

hyizhu1108@gmail.com | https://haoyizhu.site/ | Haoyi Zhu | Google Scholar

EDUCATION

University of Science and Technology of China (USTC), China *Sep. 2023 – present*
Ph.D. in Computer Science
Shanghai Jiao Tong University (SJTU), China *Sep. 2019 – Jun. 2023*
B.E. in Artificial Intelligence Honor Class

PUBLICATIONS

Total Citations: 1300

Journal Papers

PonderV2: Pave the Way for 3D Foundation Model with A Universal Pre-training Paradigm [paper] [code]
Haoyi Zhu*, Honghui Yang*, Xiaoyang Wu*, Di Huang*, Sha Zhang, Xianglong He, Tong He, Hengshuang Zhao, Chunhua Shen, Yu Qiao, Wanli Ouyang
*IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), 2025*

AlphaTracker: A Multi-Animal Tracking and Behavioral Analysis Tool [paper] [code]
Zexin Chen, Ruihan Zhang, Hao-Shu Fang, Yu E Zhang, Aneesh Bal, Haowen Zhou, Rachel R Rock, Nancy Padilla-Coreano, Laurel R Keyes, **Haoyi Zhu**, Yong-Lu Li, Takaki Komiyama, Kay M Tye, Cewu Lu
Frontiers in Behavioral Neuroscience, 2023

AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time [paper] [code]
Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu*, **Haoyi Zhu***, Yuliang Xiu, Yong-Lu Li, Cewu Lu
[>8K GitHub Stars] *IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), 2022*

Conference Papers (Selected)

Aether: Geometric-Aware Unified World Modeling [paper] [project]
Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, Tong He
arXiv preprint, 2025

SPA: 3D Spatial-Awareness Enables Effective Embodied Representation [paper] [project]
Haoyi Zhu, Honghui Yang, Yating Wang, Jiange Yang, Limin Wang, Tong He
*International Conference on Learning Representations (**ICLR**), 2025*

Tra-MoE: Learning Trajectory Prediction Model from Multiple Domains for Adaptive Policy Conditioning [paper]
Jiange Yang, **Haoyi Zhu**, Yating Wang, Gangshan Wu, Tong He, Limin Wang
*IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2025*

Point Cloud Matters: Rethinking the Impact of Different Observation Spaces on Robot Learning [paper] [project]
Haoyi Zhu, Yating Wang, Di Huang, Weicai Ye, Wanli Ouyang, Tong He
Advances in Neural Information Processing Systems (NeurIPS), 2024

RH20T: A Comprehensive Robotic Dataset for Learning Diverse Skills in One-Shot [paper] [project]
Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, **Haoyi Zhu**, Cewu Lu
IEEE International Conference on Robotics and Automation (ICRA), 2024

UniPAD: A Universal Pre-Training Paradigm for Autonomous Driving [paper] [code]
Honghui Yang, Sha Zhang, Di Huang, Xiaoyang Wu, **Haoyi Zhu**, Tong He, Shixiang Tang, Hengshuang Zhao, Qibo Qiu, Binbin Lin, Xiaofei He, Wanli Ouyang
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024

X-NeRF: Explicit Neural Radiance Field for Multi-Scene 360° Insufficient RGB-D Views [paper] [code]
Haoyi Zhu, Hao-Shu Fang, Cewu Lu
IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023

MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge [paper] [project]

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, **Haoyi Zhu**, Andrew Tang, De-An Huang, Yuke Zhu, Anima Anandkumar
[Outstanding Paper Award] *Advances in Neural Information Processing Systems (NeurIPS), 2022*

For the full publication list, please visit my Google Scholar page.

RESEARCH EXPERIENCE

General 3D & Embodied Representation Team, Embodied AI Center, Shanghai AI Lab
Advisors: Prof. Tong He, Prof. Wanli Ouyang, and Prof. Xiaogang Wang

World Model & Embodied Foundation Model *Nov. 2024 – Present*

- Exploring scalable embodied AI. Mainly two aspects: World Model and learning from human demonstrations.
- Proposed **Aether**, a geometry-aware unified world model. **Aether** features 3 core capabilities: (1) 4D dynamic reconstruction, (2) action-conditioned video prediction, and (3) goal-conditioned visual planning. Building upon video generation models, our framework demonstrates unprecedented synthetic-to-real generalization despite *never* observing real-world data during training.
- Mentored several junior students on robot learning topics, including action tokenizers, cross-embodiment policies, point cloud policies, *etc.*

General 3D Embodied Representation Learning *Mar. 2023 – Oct. 2024*

- Conducted extensive research on general 3D representation for embodied AI, leading to the publication of three first-author papers.
- Proposed a universal 3D pre-training framework, ***PonderV2***, which leverages differentiable neural rendering to learn point cloud representations. Achieved state-of-the-art (SOTA) performance across 11 indoor and outdoor benchmarks.
- Performed an in-depth investigation of observation spaces in robot learning, culminating in the discovery of the critical importance of ***Point Cloud Matters***.
- Developed ***SPA***, an innovative representation learning framework that highlights the significance of 3D spatial awareness in embodied AI. Delivered the most comprehensive evaluation of embodied representation learning to date, surpassing over 10 SOTA methods.

Jim Fan's Team (currently GEAR Lab), NVIDIA AI
Advisors: Dr. Jim Fan & Prof. Anima Anandkumar

Building Open-Ended Embodied Agents with Internet-Scale Knowledge *Feb. 2022 – Jul. 2022*

- Contributed to the ***MineDojo*** project, a new framework designed to build generally capable agents with internet-scale knowledge in Minecraft. Won the **Outstanding Paper Award** at NeurIPS 2022.
- Co-developed the YouTube database, which comprises over 730,000 narrated Minecraft videos, totaling approximately 300,000 hours of content and 2.2 billion English transcripts.
- Constructed the initial version of MineCLIP, a video CLIP foundation model that acts as a learned reward function, enabling agents to solve diverse open-ended tasks specified in natural language.
- Authored APIs for three databases (YouTube, Minecraft Wiki, and Reddit), along with the corresponding documentation and part of the official project website.

Machine Vision and Intelligence Group (MVIG), Shanghai Jiao Tong University
Advisors: Dr. Hao-Shu Fang & Prof. Cewu Lu

General Robot Manipulation *Aug. 2021 - Jul. 2023*

- Contributed to ***RH20T***, the largest robotic manipulation dataset to date, featuring over 140 skills and exceeding 40 TB in size.
- ***RH20T*** includes more than 110,000 contact-rich real-world robotic manipulation sequences encompassing diverse skills, robots, viewpoints, objects, and backgrounds. It provides multimodal data, including visual, tactile, audio, and proprioception information, paired with human demonstrations for each task.

- Investigated 3D hand keypoint labeling through multi-view images using 2D hand pose estimation combined with mesh optimization.
- Proposed ***X-NeRF***, an explicit neural radiance field representation for novel view synthesis under challenging multi-scene 360° settings with insufficient RGB-D views.

Fast and Accurate Multi-Person Whole-Body Pose Estimation *Aug. 2020 - Aug. 2021*

- Contributed to the development of ***AlphaPose***, a fast and accurate system for multi-person whole-body pose estimation and tracking.
- Achieved **over 8K stars** on GitHub, ranking in the **top 0.01%** of all GitHub repositories.
- Participated in the annotation of the Halpe full-body human keypoint and human-object interaction detection (HOI-Det) datasets.
- Extended the parametric pose NMS algorithm to whole-body scenarios and conducted extensive experiments.
- Recognized as an "Excellent Open-Source Project in China, 2020," being the only university project among the 10 winners.

INVITED TALKS

- 2025/03/27, Graduate Academic Forum, USTC, “Towards Spatial Intelligence: From 3D Vision to Embodied AI"
- 2024/12/27, ZhiXingXing Embodied AI Frontier Lecture, Shanghai, “Towards Spatial Intelligence: From 3D Vision to Embodied AI" [link]