SIJIN CHEN

Fudan University | H: +86 187 1792 9716 | csjch3cook@gmail.com
Homepage | GitHub | Google Scholar

About Me

I am an AI researcher working on embodied AI with Dr. Tao Kong at **ByteDance Research**. I received my Master's degree in Artificial Intelligence from **Fudan University** (Sep. 2021 - Jun. 2024), where Prof. Tao Chen is my advisor. I am fortunate to work closely with Dr. Hongyuan Zhu from A*STAR, Singapore, and Dr. Gang Yu, Dr. Xin Chen, and Dr. Chi Zhang from Tencent. Before this, I obtained my Bachelor's degree in Data Science and Big Data Technology also from **Fudan University** (Sep. 2017 - Jun. 2021).

My long-term research goal is to develop vision-language systems that possess the capacity to **comprehend**, **reason**, and **envision** the physical world. Outside my research, I love sports and music.

Research Interests

Multi-modal Learning, Vision and Language, Large Language Models, and Generative AI.

Employment

AI Researcher Jul. 2024 - .
ByteDance Research. Working on embodied AI with Dr. Tao Kong. Beijing, China

Education

Masters in Artificial Intelligence (GPA 3.56/4.00) Sep. 2021 - Jun. 2024
Fudan University. Advised by Prof. Tao Chen. Shanghai, China
Bachelor in Data Science and Big Data Technology Sep. 2017 - Jun. 2021
Fudan University. Shanghai, China

Selected Publications (google scholar)

- MeshXL: Neural Coordinate Field for Generative 3D Foundation Models.
  Sijin Chen, Xin Chen, Anqi Pang, Xianfang Zeng, Yijun Fu, Wei Cheng, Fukun Yin, Yanru Wang, Zhibin Wang, Jingyi Yu, Gang Yu, Bin Fu, Tao Chen.
  [NeurIPS 2024 | project | paper | github]
  [Summary]: Building end-to-end large auto-regressive 3D mesh generation models.
- LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding, Reasoning, and Planning.
  Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, Tao Chen.
  [CVPR 2024 | project | paper | github]
  [Summary]: 3D-LLMs respond to visual and text interactions in complex 3D scenes.
- Vote2Cap-DETR++: Decoupling Localization and Describing for End-to-End 3D Dense Captioning.
  Sijin Chen, Hongyuan Zhu, Mingsheng Li, Xin Chen, Peng Guo, Yinjie Lei, Gang Yu, Taihao Li, Tao Chen.
  [T-PAMI 2024 | paper | github]
  [Summary]: Decoupled feature extraction for localizing and describing objects in 3D scenes.
- End-to-End 3D Dense Captioning with Vote2Cap-DETR.
  Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, Tao Chen.
  [CVPR 2023 | paper | github | youtube]
  [Summary]: Addressing 3D dense captioning as a set prediction problem with parallel decoding.
- 3DET-Mamba: Causal Sequence Modelling for End-to-End 3D Object Detection.
  Mingsheng Li, Jiakang Yuan, **Sijin Chen**, Lin Zhang, Anyu Zhu, Xin Chen, Tao Chen.
  [NeurIPS 2024]
  [Summary]: Exploring state space model's potential as both encoder and decoder for 3D detection.
- M3DBench: Let's Instruct Large Models with Multi-modal 3D Prompts.
  Mingsheng Li, Xin Chen, Chi Zhang, **Sijin Chen**, Hongyuan Zhu, Fukun Yin, Gang Yu, Tao Chen.
  [ECCV 2024 | project | paper | github]
  [Summary]: A large scale dataset querying 3D LLMs with text, 2D, and 3D prompts.

- WI3D: Weakly Incremental 3D Detection via Visual Prompts.
  Mingsheng Li, **Sijin Chen**, Shengji Tang, Hongyuan Zhu, Xin Chen, Fukun Yin, Tao Chen.
  [Under Review | paper]
  [Summary]: Introducing new categories to 3D detectors with 2D foundation models.

Projects

- **Generative 3D Foundation Models.** Jan. 2024 - Jun. 2024
  Put forward **MeshXL**, a family of generative pre-trained transformers for the direct generation of 3D object meshes, accepted to NeurIPS 2024.
- **Language for 3D Scenes.** Aug. 2021 - Mar. 2024
  Proposed **Vote2Cap-DETR**, a set-to-set method for localizing and describing objects in 3D scenes, accepted to CVPR 2023 and won the Scan2Cap challenge at ICCV 2023. Proposed an advanced method, **Vote2Cap-DETR++**, which is accepted to T-PAMI 2024. Presented **LL3DA**, a large language 3D assistant responding to both text and visual interactions with complex 3D scenes, accepted to CVPR 2024. Put forward **M3DBench**, a large-scale multi-modal 3D dataset covering 327k lines of annotations for diverse 3D vision and language tasks, accepted to ECCV 2024.
- **Class-Incremental 3D Detection.** Apr. 2023 - Dec. 2023
  Proposed WI3D, learning to detect new categories from 2D images, under review.
- **Earlier Projects.** Before Sep. 2021
  Self-Supervised Pre-training on 3D Point Clouds. Developed a self-supervised learning algorithm that learns global- and patch-level contrastive representations for 3D point clouds.
  A Smart Advertisement Display System. Developed a human perception system that detects faces, recognizes facial expressions, estimates eye gaze, age, and gender for advertisement recommendation.

Scholarships and Awards

Outstanding Graduate Student Award (rank 1/24).	Apr. 2024
First place winner of the Scan2Cap Challenge at ICCV 2023.	Oct. 2023
National Scholarship (rank 1/46).	Sep. 2023
Second prize of the Scholarship for Outstanding Students.	Sep. 2022
Award for the Scholarship for Outstanding Students.	Sep. 2021
Second prize of the Scholarship for Outstanding Students.	Jun. 2021

Research Intern

Tencent. Jan. 2024 - Jun. 2024
Research Intern, advised by Dr. Xin Chen and Dr. Gang Yu, working on generative 3D foundation models.

Invited Talks

- "**MeshXL**: Neural Coordinate Field for Generative 3D Foundation Models" Jul. 2024
  With a proper ordering protocol, the direct generation of 3D meshes can be modelled into a "next-coordinate generation" problem, and be seeminglessly addressed with modern large language model techniques. A technical report at miHoYo.
- "**Vote2Cap-DETR**: A Set-to-Set Perspective Towards 3D Dense Captioning" Oct. 2023
  By treating 3D Dense Captioning as a translation task from a set of object queries into a set of "box-caption" pairs, we present a set-to-set perspective towards 3D Dense Captioning. A winner presentation for the Scan2Cap challenge at ICCV 2023.
- "End-to-End 3D Dense Captioning with Vote2Cap-DETR" Jun. 2023
  We present an end-to-end transformer model for localizing and describing objects in parallel within diverse 3D environments. A paper presentation at VALSE 2023, Wuxi, China.

Skills

Languages:	Chinese (native), English (proficient), Shanghai dialect
Programming:	Python, R, C, Matlab, SQL
Tools:	PyTorch, Blender, Visual Studio, Spyder, Jupyter Notebook