

Mr. Wu Size

S-Lab for Advanced Intelligence
Nanyang Technological University
50 Nanyang Avenue, Singapore 639798

Cellphone: +65 8048 1678
Email: size001@e.ntu.edu.sg
Homepage: <https://wusize.github.io/>

EXPERIENCES

Nanyang Technological University (NTU)

Ph.D. student in College of Computing and Data Science

Singapore

Jan 2022—Dec 2025

SenseTime, Beijing

Research Intern in Computer Vision

Beijing, China

Oct 2020—Dec 2021

University of Science and Technology of China (USTC)

Bachelor of Electronic and Information Engineering (AI)

Hefei, China

Aug 2017—Jun 2021

GPA: 3.8/4.3 (Ranking Top 10%)

RESEARCH INTERESTS

Multimodal Perception & Generation (Multimodal LLMs); Scene Understanding (Object Detection & Image Segmentation); Multi-view Human Detection & Tracking

PAPERS

- [1] **Size Wu**, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. Preprint, 2025.
- [2] **Size Wu**, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-lmm: Grounding frozen large multimodal models. In *CVPR*, 2025.
- [3] Shilin Xu, Xiangtai Li, **Size Wu**, Wenwei Zhang, Yining Li, Guangliang Cheng, Yunhai Tong, Kai Chen, and Chen Change Loy. Dst-det: Simple dynamic self-training for open-vocabulary object detection. In *TCSVT*, 2024.
- [4] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, **Size Wu**, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *CVPR*, 2024.
- [5] **Size Wu**, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipsef: Vision transformer distills itself for open-vocabulary dense prediction. In *ICLR*, 2024.
- [6] **Size Wu**, Wenwei Zhang, Lumin Xu, Sheng Jin, Wentao Liu, and Chen Change Loy. Clim: Contrastive language-image mosaic for region representation. In *AAAI*, 2024.
- [7] **Size Wu**, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *CVPR*, 2023.
- [8] **Size Wu**, Sheng Jin, Wentao Liu, Lei Bai, Chen Qian, Dong Liu, and Wanli Ouyang. Graph-based 3d multi-person pose estimation using multi-view images. In *ICCV*, 2021.

RESEARCH PROJECTS

Multimodal Perception & Generation

NTU, Jan 2024—Preprint

- **Unified Multimodal Understanding and Generation.** Propose Harmon, a unified framework for multimodal understanding and generation with a shared visual representation. Under review [1] **Code**.
- **Grounding Large Multimodal Models:** Propose to ground frozen multimodal LLMs by exploiting location priors in LLM’s attention weights. The large multimodal models are endowed with grounding ability without losing conversation capability. **CVPR 2025** [2] **Code**.

Open-World Scene Understanding

NTU, Jan 2022—Dec 2023

- **Enhancing Region Representation for Open-Vocabulary Object Detection and Image Segmentation:** Propose resource-efficient and effective methods to improve region representation regarding vision-language alignment. An approach that generates pseudo region-text pairs is accepted by **AAAI 2024** [6] (**Code**), and a self-distillation approach is accepted by **ICLR 2024** [5] (**Code**) as spotlight (top 5%).
- **Aligning Bag of Regions for Open-vocabulary Object detection:** Propose to distil knowledge from pre-trained vision-language models (VLMs) on a bag of regions for open-vocabulary object detection (OVD). The method effectively exploits the VLMs’ ability to represent co-existing and contextually related object concepts. It achieves state-of-the-art performance on multiple OVD benchmarks. **CVPR 2023**. [7] **Code**.

Multiview Detection & Tracking

SenseTime, Beijing, Oct 2020—Dec 2021

- **Multi-view 3D Human Pose Estimation:** Propose three task-specific graph neural networks (GNNs) for multi-view human pose estimation. The GNNs efficiently match human centres, predict human locations and refine human pose estimations. The proposed method achieves state-of-the-art performance on CMU Panoptic and Shelf datasets with significantly lower computation complexity. **ICCV 2021.** [8][Code](#).
- **Multi-view Multi-person Tracking:** Aggregating geometric and appearance features for the Multi-camera Multiple People Tracking problem. The method is applied to industrial projects with 3 patents applied. It also achieves 3rd place in a challenge in ICCV 2021 Multi-camera Multiple People Tracking Workshop.

OPEN-SOURCE PROJECT

MMPose

Oct 2020—May 2023

I am one of the contributors to MMPose, an open-source project on GitHub for 2D and 3D human pose estimation. I am the core developer of multi-view 3D pose estimation in MMPose.

CORE QUALIFICATIONS

Programming: Python, C/C++, MATLAB

English: TOEFL 106/120, **GRE 337/340+4.0/6.0**

AWARDS AND HONORS

- AISG PhD Fellowship. 2023.
- Future Star Award. SenseTime. 2021.
- Outstanding Final Year Project (**Top5%**). USTC. 2021.
- Honored Class for **Artificial Intelligence**. USTC. 2019—2021.
- Open-Vocabulary Object Detection Contest: **First Place Award**. CSIG. 2023.
- Security AI Challenger: **Ninth Place Award (out of 36489 teams)**. Tianchi, Alibaba Cloud. 2021.
- AI Innovation and Application Competition: **Second Place Award**. CAICT. 2021.
- Multi-camera Multiple People Tracking Challenge: **Third Place** in Top-down View Track. ICCV 2021.
- National Encouragement Scholarship. USTC. 2018, 2019.
- Scholarship For Outstanding Students. USTC. 2018—2020.
- Talent Program in Information Science and Technology. USTC. 2017—2021.