

Zhang Handuo

+65 83410772 · zhanghanduo@gmail.com · Personal Site: handuo.top · github.com/zhanghanduo
Nationality: Chinese (SG PR) · Date of Birth: 28/05/1988 · [LinkedIn-Handuo-64785367](https://www.linkedin.com/in/handuo-64785367)

RESEARCH INTEREST

I am an AI passionate about advancing NLP, computer vision, robot navigation, and multi-modal learning, with a focus on transforming complex research into practical, scalable solutions. My interests lie in GenAI system building, LLM post-training, as well as enhancing robot capability. With expertise in C++, Python, and frameworks like TRL and DeepSpeed, I excel at hands-on development, optimizing models for edge devices, and building intelligent workflows.

KEY COMPETENCIES

Python, C/C++, SQL, ROS	PyTorch, TRL, DeepSpeed,	AutoGen, LangChain, LangGraph
Computer Vision Expertise	TorchTitan, Unsloth, LLaMa-Factory	Dify, RagFlow, MCP Server
Team Collaboration & Leadership	Model Optimization & Deployment	Micro-Service Dev Experience

PROFESSIONAL EXPERIENCE

Shanda Group AI Research Institute (Singapore)
AI Scientist

Sep 2023 - Present

Lead a team of 5 in developing GenAI solutions for intelligent query parsing, intent recognition, and automated task flows, delivering AI-driven “smart hands and feet” to empower multiple industries.

Project 1 – G-Mate (2024-07 to Present)

Overview: An online self-media product promotion assistant designed for Key Opinion Leaders (KOLs) and B-end users, offering features like hot-spot tracking, multi-modal content creation, strategy suggestion & scheduling, and automated promotion workflows. Product Page: <https://web.gm8.ai/>.

- **Event Map:**
 - Designed an incremental Knowledge Graph connecting events to over 20k entities (e.g., brands, influencers, products), updated in real time using auto-tagging and community hierarchy management.
 - Improved entity linkage accuracy by 20% through AI-driven disambiguation, ensuring reliable connections between trending events and relevant stakeholders.
 - Enabled sub-second query responses for event insights, supporting 500+ daily queries from users with 99% uptime.
- **Content Engine:**
 - Spearheaded the development of a GenAI pipeline supporting content creation in 3 modalities (text, image, video), tailored to KOL personas and memories, aided by Kling API.
 - Boosted engagement metrics (likes, shares) by 25% compared to human benchmarks, validated via A/B testing across 10,000+ posts, using both GPT-4o and self-hosted fine-tuned Qwen-32B model.
 - Slashed content creation time from 2 hours to 15 minutes per piece, improving team productivity by 800% and enabling rapid campaign scaling.
- **Tools Model:**
 - Fine-tuned LLama3.3 70B model with LLaMa-Factory and 16xA100 to handle 100+ parallel function calls, optimizing it for Twitter/Tiktok API calls, and KOL-specific workflows like scheduling and analytics retrieval: ([Huggingface watt-tool-70B](#) and [Huggingface watt-tool-8B](#)).
 - Enhanced multi-turn conversation coherence by 25%.
 - Secured **first place on the [BFCL Berkley Function-Calling Leaderboard](#)**, surpassing industry standards by 15% in accuracy compared to other open-source models (from 2024 Dec till present).

Project 2 – W-Hiring (2024-03 to 2024-06)

Overview: An AI-driven HR solution integrating cutting-edge AI techniques with corporate recruitment needs, streamlining talent acquisition for enterprises.

- **Workflow Automation:**
 - Automated the end-to-end recruitment pipeline (doc parsing done by others) with promptflow framework, reducing manual effort by 70% and shrinking time-to-hire from 30 days to 15 days across 50+ job postings.
 - Developed an AI-powered interview question generator, tailoring questions to job roles and improving candidate assessment precision by 20%, as rated by hiring managers.
- **RAG System:**
 - Created a recruiting knowledge base with 6,000+ documents and a talent knowledge graph linking 500,000 candidates by skills, experience, qualifications and preferences.
 - Integrated Text2SQL for fuzzy semantic search, enabling HR teams to query candidates in natural language with 89% precision (e.g., “Find Python developers graduated from QS top 50 with 3+ years”).
 - Engineered a hybrid search system merging keyword matching, semantic reranking and generative models, enhancing candidate-job match accuracy by 15% (e.g., 90% match success vs. 75% baseline).
 - Enabled real-time matching at under 5 seconds per candidate, supporting high-volume recruitment for 100+ applicants monthly.

Project 3 – Watt Personal Assistant (2023-09 to 2024-02)

Overview: An AI-powered, Zapier-like product enabling multi-agent collaboration to automate complex tasks through conversational interfaces, enhancing user productivity.

- **Intent & Tool API Selection:**
 - Transitioned from a RAG system for API set retrieval to supervised fine-tuning for direct API function call, boosting intent recognition accuracy from 89% to 96% across 50,000+ user queries.
 - Reduced average token usage per query by 70% (from 1200 to 360 tokens), cutting operational costs by 60% and improving response times to under 2 second.
 - Decreased user-reported tool selection errors by 20%.
- **Function Calling:**
 - Implemented a function-call workflow supporting 200+ unique functions (<https://lupan.watt.chat/>), with parameter extraction and iterative refinement for tasks like email drafting and calendar scheduling.
 - Improved parameter accuracy from 80% to 95%, reducing user corrections by 60% in a sample of 1,000 multi-step tasks.
 - Raised task completion rates from 70% to 90%, enabling seamless automation for complex workflows like project management.
- **User Profiling:**
 - Developed a user memory bank based on usage data, capturing preferences (e.g., persona, task frequency and case scenarios) and behavior patterns.
 - Reduced average task completion time by 30% (from 10 minutes to 7 minutes) with proactive suggestions tailored to individual profiles.

Nanyang Technological University (Singapore)

April 2023 - Present

Part-time Lecturer

Lecturer of NTU post-graduate course EE7207 Neural Networks and Deep Learning which covers cutting-edge techniques like attention mechanism, transformer architecture, graph neural networks and modern large language model frameworks.

My part of the lecture includes GNN with its variants, the advanced training & inference techniques of CNN, and transformer, and the applications of modern networks (including BatchNorm, LayerNorm, GroupNorm, Fully Convolutional Networks, EfficientNet, MobileNet, ViT, etc) and use cases (object detection, semantic/instance segmentation and segmentAnything).

Mind Pointeye Pte. Ltd. (Singapore)
AI Scientist

April 2021 - Aug 2023

I work as the tech lead of 7 AI and data engineers on the development and deployment of video analysis algorithms and IoT projects (<https://mindpointeye.com/>).

Object Detection & Optimization:

- Built an object detection pipeline with quantization-aware mixed-precision training.
- Optimized training with a curriculum learning approach (https://github.com/zhanghanduo/yolox_pl), cutting training time by 20% while maintaining precision (mAP on COCO dataset).
- Applied post-training quantization in C++ for edge devices (e.g., Rockchip RK3588 NPU and SE50221), reducing model size by 4x and accelerating inference by 3x with minimal accuracy loss.

Time Series Forecasting:

- Developed models for stock return forecasting and IoT sensor life prediction.
- Created a hybrid model that combines transformer/LSTM for temporal feature extraction with a GNN for relation modeling, for multi-stock relative ranking forecasting problems, improving accuracy by 12% over traditional ARIMA methods and 10% over DeepAR on historical stock data.

Video Algorithms & IoT Analysis:

- Designed a feature extraction pipeline for the "Multi-embedding query for person MOT-ReID System" using deep convolutional networks, improving re-identification accuracy by 10% on the Market-1501 dataset.
- Implemented a spatio-temporal graph convolutional network for human action recognition and intrusion detection, achieving 95% detection accuracy in real-time video streams.

Nanyang Technological University Robotics I Lab
Project Officer (2015) and PhD Candidate (2016-2021)

Sep 2015 - Jun 2021

I work as the team leader of 3 PhD students on the collaboration project with Singapore Technology Kinetics. The project aims to develop a high speed stereo vision system and apply it onto unmanned ground vehicles with multi-sensor fusion using deep neural networks. Project page: [Stereo Perception](#).

- Developed a real-time stereo matching algorithm using semi-global matching (SGM) in C++, achieving 18 fps on edge device Nvidia Jetson TK2 for obstacle avoidance and prior input for self-localization.
- Implemented multi-object detection & tracking system for objects up to 50 meters, achieving mean errors of 1.58m (distance), 1.25° (bearing), and 0.45m (size) within 30 meters. I integrated DarkNet YOLO-3 (https://github.com/zhanghanduo/yolo3_pytorch) with TorchScript to achieve 10 fps on Jetson TK2.
- Built a visual SLAM pipeline with ORB features and bundle-adjustment (optional with Lidar, IMU and GPS), reducing translation error to 0.043% and rotation error to 0.41° in dynamic and heavy traffic scenarios.
- Integrated deep learning object detection with a Kalman filter for multi-object tracking (MASS), achieving a True Positive Rate (TPR) of 0.947 and MOTA of 0.915 on the [KITTI MOT Benchmark](#) ranking 4th in 2019. Now it is positioned at 71 due to various more advanced DL based MOT models.
- Integrated real-time mapping with bird's-eye view representation.

EDUCATION & CERTIFICATIONS

Ph.D in Robot Vision & Control

Nanyang Technological University
2016 - 2022, Cum. GPA: 4.67/5.00

M.SC in Pattern Recognition & Intelligent System

Northeastern University (China)
2011 - 2013, GPA: 3.94/4.00

Bachelor in Automation Control

Northeastern University (China)
2007 - 2011

EXTRACURRICULAR ACTIVITIES

- Teaching Assistant for 'Data Structure' and 'Signal Processing' courses, explaining complex concepts and supporting undergraduates.
- 2013 National Graduate Scholarship
- 2011: 1st Prize, 8th National Graduate Mathematical Contest in Modeling, China
- 2010: Meritorious Winner (First Prize), American Mathematical Contest in Modeling

PUBLICATIONS

PhD Thesis: Visual metric and semantic localization for UGV, 2021.

(1) GMC: Grid Based Motion Clustering in Dynamic Environment

Handuo Zhang, K Hasith, Han Wang, Intelligent System Conference (IntelliSys), 2019.

(2) LaCNet: Real-time End-to-End Arbitrary-shaped Lane and Curb Detection with Instance Segmentation Network

Hui Zhou, Han Wang, **Handuo Zhang**, K Hasith, ICARCV 2020.

(3) Multiple Object Tracking With Attention to Appearance, Structure, Motion and Size

K Hasith, Han Wang, **Handuo Zhang**, IEEE Access, 2019.

(4) Real Time Multiple Object Tracking using Deep Features and Localization Information

K Hasith, **Handuo Zhang**, Han Wang, ICCA, 2019.

(5) A consistent and long-term mapping approach for navigation

Handuo Zhang, K Hasith, Han Wang, International Journal of Research in Advent Technology (IJRAT), 2019.

(6) Heading Reference-Assisted Pose Estimation for Ground Vehicles

Han Wang, Rui Jiang, **Handuo Zhang**, SS Ge, IEEE Transactions on Automation Science and Engineering (T-ASE), 2018.

(7) A hybrid feature parametrization for improving stereo-SLAM consistency

Handuo Zhang, K Hasith, Han Wang, International Conference on Control and Automation (ICCA), 2017.

(8) Ultra-wideband aided fast localization and mapping system

Chen Wang, **Handuo Zhang**, TM Nguyen, L Xie, International Conference on Intelligent Robots and Systems (IROS), 2017.

(9) Stereo vision based negative obstacle detection

K Hasith, **Handuo Zhang**, Han Wang, ICCA, 2017.

(10) Object co-segmentation via weakly supervised data fusion

Shiping Wang, **Handuo Zhang**, Han Wang, Computer Vision and Image Understanding (CVIU), 2017.