RESUME

Basic Information

Name:	Liyun Zhang
Nationality:	China-Xi'an
Institution:	Takemura Lab, Osaka University
	https://www.lab.ime.cmc.osaka-u.ac.jp/
Status:	3rd year PhD student
Address:	Cybermedia Center, 1-32 Machikaneyama, Toyonaka, Osaka, Japan
Phone Number:	+81 08080532280 +86 18092916581
E-mail:	liyun.zhang@lab.ime.cmc.osaka-u.ac.jp liyunzhang9120@gmail.com

Homepage:

https://zhangliyun9120.github.io/

Research Interests

Multi-modal (Vision + Language) Reasoning, Embodied AI, Interactive Robotic Learning

Education

- **Osaka University** 2020.10 - 2024.3
  PhD Candidate, Information Systems Engineering
  Research focus: Mainly engaged in image translation / generation / recognition. multi-modal (vision + language); robotic perception, interactive robotic learning, and embodied AI research.
  Current Projects: using a large language model (LLM) to assist in training a multimodal model, which is carried on controlling (avatar) robotic motions and behaviors in the simulative or real interactive dialogue scene with humans. We learn features from language, audio and video frames, then reason the nonverbal behaviors (facial expression + body gesture) and autonomous actions (manipulate objects + interact with humans) of robots.
- **Xi'an University of Science and Technology** 2012.9 - 2015.7
  Master of Science, Computer Technology Engineering
  Research focus: Uneven illumination image segmentation and object recognition, Linux-based embedded automation robotic system

Employment

- **Visiting Researcher** 2023.2 - 2024.3
  Georgia Institute of Technology https://animesh.garg.tech/
  Description: Multi-modal reasoning and LLMs-based embodied AI
- **Research Associate** 2023.7 - 2024.3
  Osaka University https://www.ist.osaka-u.ac.jp/english/
  Description: Embodied AI and Multi-modal Reasoning

- **Specially Appointed Researcher** 2022.5 - 2023.3
  Sysmex Corporation https://www.sysmex.co.jp/en/index.html
  Description: Identify the area with ointment applied on the forearm (3D partial human body mesh and pose estimation from monocular image)
- **Specially Appointed Researcher** 2021.5 - 2022.3
  Sysmex Corporation https://www.sysmex.co.jp/en/index.html
  Description: Time series missing values imputation using GANs-based bidirectional recurrent model on ICU MIMIC-III datasets
- **Research Associate & Teaching Assistant** 2020.10 - 2021.4
  Osaka University https://www.ist.osaka-u.ac.jp/english/
  Description: Mainly worked on Image translation / generation, SLAM and intelligent robot research & assisting graduate students in experiments.
- **Senior Embedded Software Engineer** 2017.12 - 2018.10
  ZTE Corporation https://www.zte.com.cn/global/index.html
  Description: Research and development of vehicle audio and power software
- **Software R&D Engineer** 2016.11 - 2017.3
  Huawei https://www.huawei.com/us/
  Description: Power management development of smartphone on MTK platform
- **Embedded Software Engineer** 2015.7 - 2017.12
  Huaqin Technology https://en.huaqin.com/
  Description: Smartphone software development, image recognition and robot vision algorithm development

Publications

(*) Peer-reviewed journal articles:

- **Liyun Zhang**, Photchara Ratsamee, Zhaojie Luo, Yuki Uranishi, Manabu Higashida, Haruo Takemura. Panoptic-Level Image-to-Image Translation for Object Recognition and Visual Odometry Enhancement. 2023 IEEE Transactions on Circuits and Systems for Video Technology (TCSVT).
- **Liyun Zhang**, Nanyan Liu, Yuanbin Hou, Xiaojian Liu. Uneven Illumination Image Segmentation Based on Multi-threshold S-F [J]. Opto-Electronic Engineering, 2014, 41(7): 81-87 (OEE).

(*) Peer-reviewed international conference papers:

- **Liyun Zhang**, Photchara Ratsamee, Bowen Wang, Zhaojie Luo, Yuki Uranishi, Manabu Higashida, Haruo Takemura. Panoptic-aware Image-to-Image Translation. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).
- **Liyun Zhang**, Photchara Ratsamee, Yuki Uranishi, Manabu Higashida, Haruo Takemura. Thermal-to-Color Image Translation for Enhancing Visual Odometry of Thermal Vision. 2022 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR).

Awards

- **Special Contribution Award** 2017.3
  Solved the problem of smartphone battery level jump in Huawei
- **Star Staff Award** 2016.10

Acquired "Star Staff" in Huaqin Telecom Technology

- **Technology Innovation Award** 2016.3 & 2015.11
  Huaqin Group Software Department Technology Innovation Second Award 2 times
- **Software copyright** 2015.4
  Steel pipe identification and counting software system
- **Electronic Design Competition Award** 2014.6
  'Automatic orifice positioning system based on embedded Linux' Electronic design competition Third Award
- **Software copyright** 2013.11
  Mine blast hole automatic positioning software system
- **RoboCup Award** 2012.11
  RoboCup China 2012 Middle Size Robot League First Award
- **Excellent Graduation Project (Thesis)** 2012.7
  "Design of Intelligent Bus Stop Announcement System Based on GPS" won the Excellence Award in the Automation Excellent Graduation Project Competition

External funding results

- 2023 Research Abroad Grant (Osaka University Future Fund Globalization Promotion)
- 2023 Osaka University Graduate School of Information Science Search Assistant
- 2022 Sysmex Student Researcher Program Specially Appointed Researcher S
- 2021 Sysmex Student Researcher Program Specially Appointed Researcher S
- 2020 Osaka University Graduate School of Information Science Search Assistant
- 2020 KAKEN — Aerial-Terrestrial-Aquatic Robots for Search and Rescue in an ATA Extreme Environment (Number: 20KK0086)

Skills:

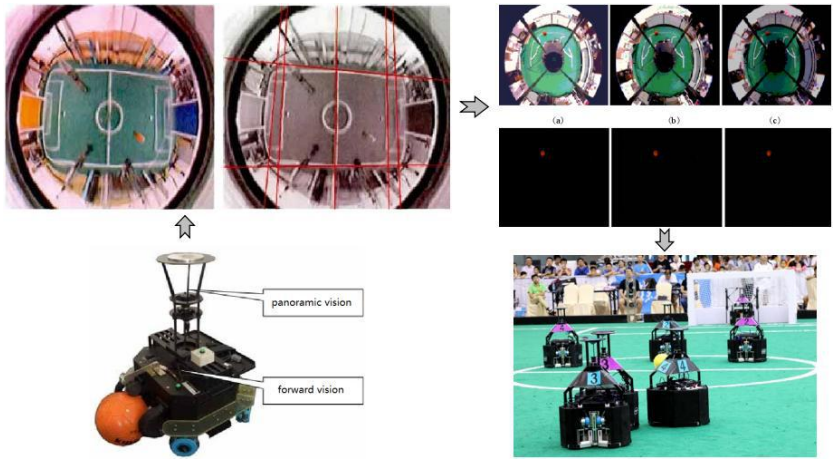
- **Models**
  LLMs, Multimodal model, Reinforcement learning, GANs, Transformer, Diffusion model.
- **Programming**
  Pytorch, Python, ROS, C/C++, Java, Android, QT, Halcon.

Languages:

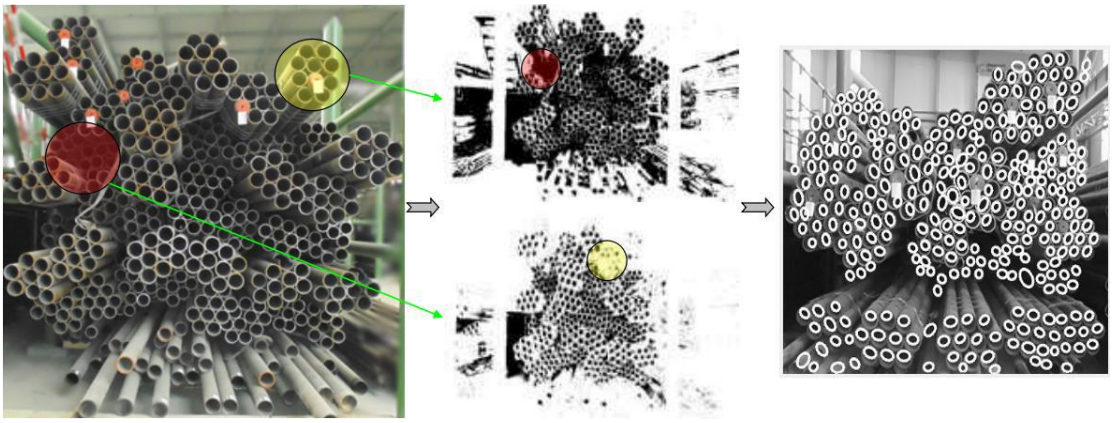
English: TOEIC, CET-6; Japanese: N2

Research Summary

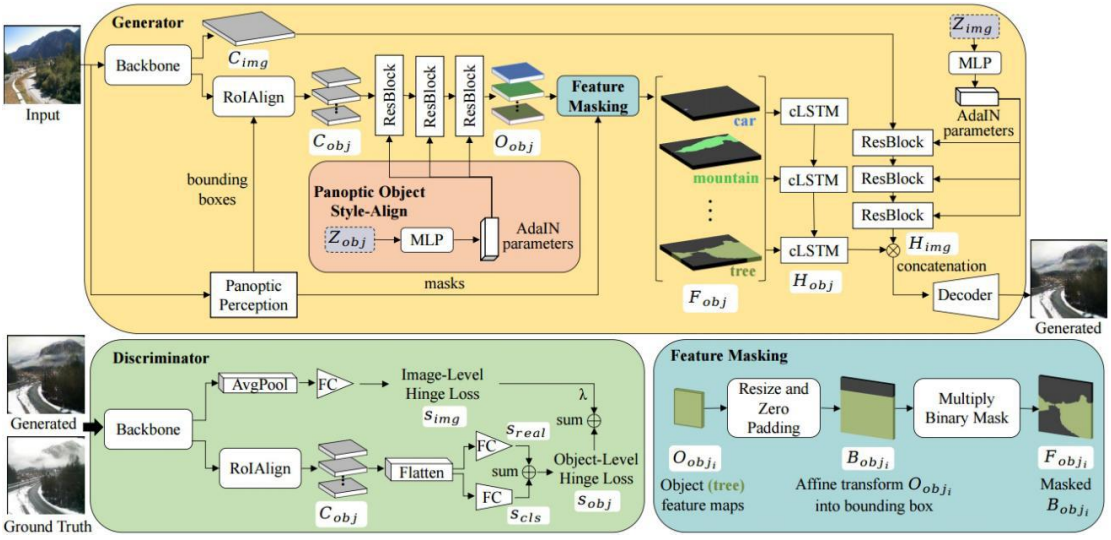
1, Soccer Robot: Recognize and track the football in the image to drive the robot to play autonomously.



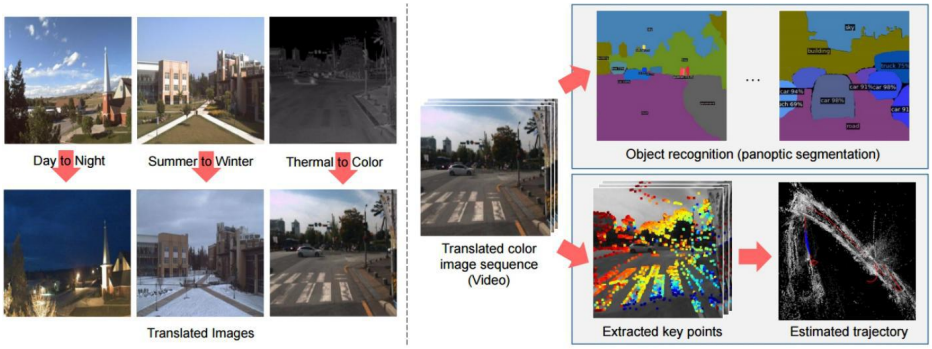
2, Uneven Illumination Image Segmentation Based on Multi-threshold S-F [1]: Segment and recognize image of steel tubes stacked in workshop under uneven illumination, and accurately count the number of steel tubes.



3, Panoptic-aware Image-to-Image Translation [2]: We proposed a panoptic-aware generative adversarial network (PanopticGAN) for image-to-image translation. The panoptic perception (i.e., foreground instances and background semantics of the image scene) is extracted to achieve alignment between object content codes of the input domain and panoptic-level target style codes, then refined by a proposed feature masking module for sharping object boundaries and higher fidelity image generation.

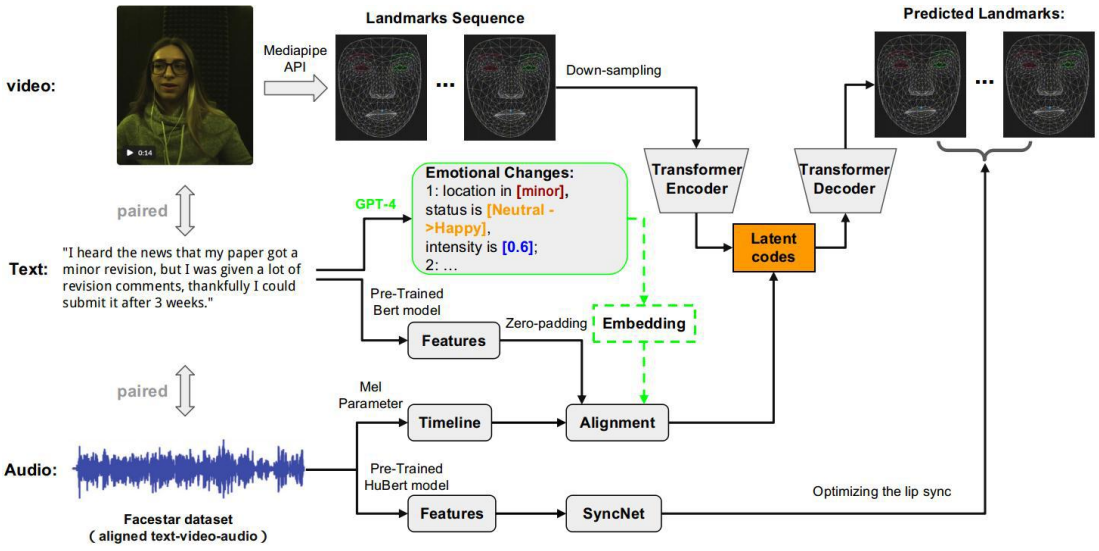


4, Panoptic-Level Image-to-Image Translation for Object Recognition and Visual Odometry Enhancement [3], [4]: We presented a pioneering panoptic-level image translation method. Compared to baselines, our method refines the object semantics on both the foreground and background to avoid semantic information loss. To sharpen the object boundaries, we proposed a feature masking module to extract the object-specific representations along the contours. The proposed method can generate higher-fidelity images/videos to further enhance the performances of the object recognition and visual odometry tasks.



(a) Panoptic-level image-to-image translation (b) Enhancement for object recognition and visual odometry

5, LLMs-based large model of Facial expression control in humanoid robot conversational interaction (preparing for ACMMM 2024) We use large language models (LLMs) to assist in training a multimodal network, LLMs process language to output the position, emotional change state (e.g., Neutral to Happy), and intensity as prior knowledge. The representations of language, audio and video fusion information are learned and transformed into blend-shape weights to achieve control of facial expressions in humanoid robot avatar conversation.



Publications

[1] **Liyun Zhang**, Nanyan Liu, Yuanbin Hou, Xiaojian Liu. Uneven Illumination Image Segmentation Based on Multi-threshold S-F [J]. Opto-Electronic Engineering, 2014, 41(7): 81-87.

[2] **Liyun Zhang**, Photchara Ratsamee, Bowen Wang, Zhaojie Luo, Yuki Uranishi, Manabu Higashida, Haruo Takemura. Panoptic-aware Image-to-Image Translation. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision.

[3] **Liyun Zhang**, Photchara Ratsamee, Zhaojie Luo, Yuki Uranishi, Manabu Higashida, Haruo Takemura. Panoptic-Level Image-to-Image Translation for Object Recognition and Visual Odometry Enhancement. 2023 IEEE Transactions on Circuits and Systems for Video Technology.

[4] **Liyun Zhang**, Photchara Ratsamee, Yuki Uranishi, Manabu Higashida, Haruo Takemura. Thermal-to-Color Image Translation for Enhancing Visual Odometry of Thermal Vision. 2022 IEEE International Symposium on Safety, Security, and Rescue Robotics.

Research Plan

Motivation

The current robots in Fig.1 do not have natural non-verbal communication (facial expressions, body postures, hand gestures and other body language) like humans and autonomous actions when interacting with humans.

Fig.1, Current Problem (from Ishiguro Lab)

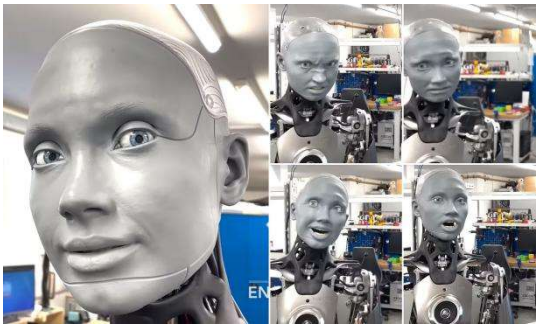


Fig.2, Reference Preview (from Ameca)

Our Goal

We want to learn these high-level abstract behaviors using a Multimodal Large Language Model (MM-LLM) through using the reasoning ability of Multimodal Large Model to train a robotic interactive learning in more complex human daily environment as show in Fig 3. The expected result is to achieve the robot's expression behavior that is closer to the naturalness of human expression such as the Ameca robot in Fig.2, and has the reasoning ability to make autonomous actions appropriately based on the interaction atmosphere.

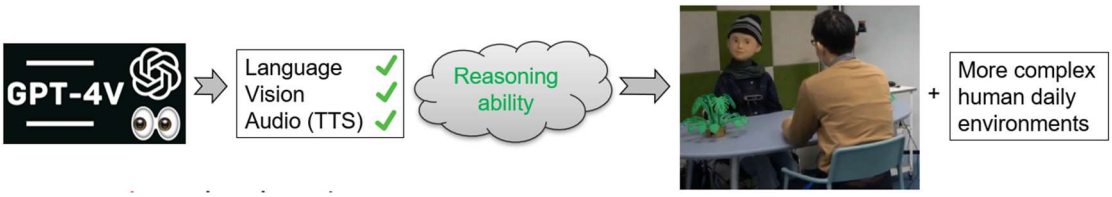


Fig.3, Robotic Interactive Learning

Interactive Robot Learning Between Human and Robot

Our goal is to construct a robot interactive learning in real or virtual environments to learn natural human expressions and behaviors. Take the conversation and interaction between robot and human avatars in a virtual environment as an example, as shown in Fig. 4.

Firstly, we use the pre-trained vision foundation models to extract two types of environmental states from the interaction environment: 1. Non-verbal communication including sequences of facial expression changes, body postures and gestures; 2. Action sequences in the interaction (if present). Secondly, we built a prompt integration module to combine visual domain information and language input, and then input the results to large multi-modal models (LMmMs) such as GPT4-Vision or LMMs that we efficiently fine-tuned using our own datasets for inference. The part so far can be regarded as multimodal understanding.

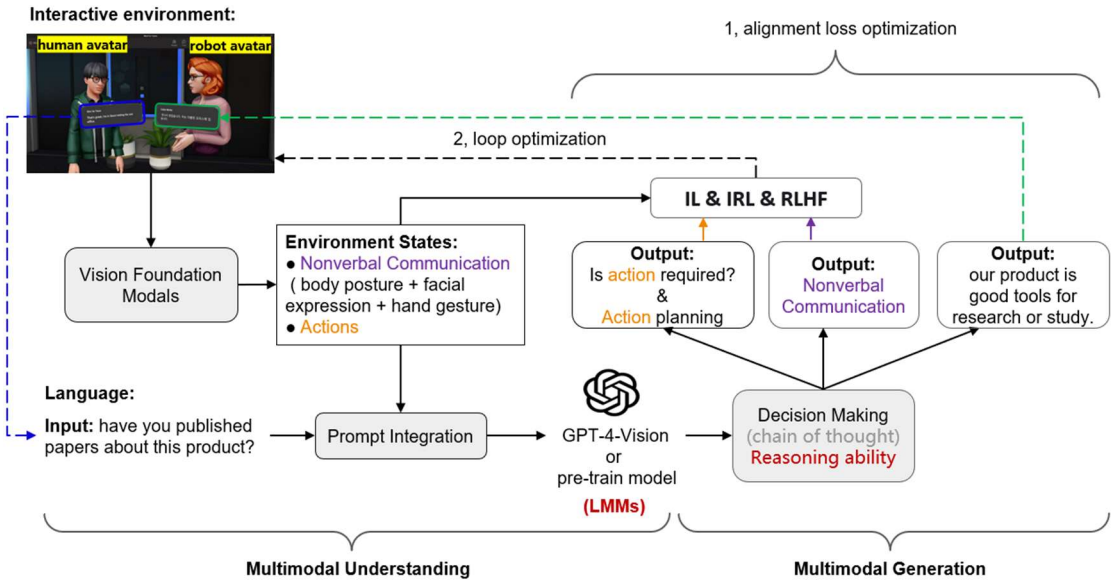


Fig.4, Architectural Concept

The LMMs decision-making makes a chain of thought reasoning to predict three types of outputs: 1. Reasoning about whether an action response is required in the current situation and detailed action planning; 2. Generate non-verbal communication responses corresponding to the input; 3. Generate the language of the reply. Finally, we can use alignment loss to optimize these three outputs. However, we also want to use imitation learning (IL) / inverse reinforcement learning (IRL) / reinforcement learning from human preferences (RLHF) methods to build loop optimization to generate better results for actions and non-verbal communication based on the environment update policies. The latter part can be regarded as multimodal generation.