Jianbiao Mei (梅剑标)

Research Area: 3D Perception & Embodied AI jianbiaomei@zju.edu.cn (Email)

EDUCATION

- **Zhejiang University** *College of Control Science and Engineering* 2021.09 – PhD Candidate
  - Research Interest: World models, LLM agents, 3D perception, Video segmentation
  - Advanced Perception on Robotics and Intelligence Learning (APRIL) Lab
  - Supervisor: Yong Liu
- **Zhejiang University** *Chu Kochen Honors College* 2017.09 – 2021.06 Bachelor of Engineering in Automation
  - GPA: 4.5/5.0 Major: Automation (Control)
  - Advanced Perception on Robotics and Intelligence Learning (APRIL) Lab

RESEARCH

1. World models

- **DreamForge: Motion-Aware Autoregressive Video Generation for Multiview Driving Scenes** 2024.08-2024.11
  - We introduce perspective guidance and develop object-wise position encoding to enhance street and foreground generation. This innovative object-wise position encoding improves foreground modeling and inherently provides local 3D correlation, leading to better object generation.
  - We propose motion-aware temporal attention to incorporate motion cues and understand the appearance changes of the video. Besides, by utilizing motion frames and an autoregressive generation paradigm, we achieve long video generation with a model trained on short sequences.
  - We integrate the proposed DreamForge with a realistic simulation platform (DRIVEARENA) to enhance coherent driving scene generation and offer more reliable open-loop and closed-loop evaluation for vision-based driving agents.
- **DriveArena: A Closed-loop Generative Simulation Platform for Autonomous Driving** 2024.06-2024.09
  - We propose the first high-fidelity closed-loop simulator for autonomous driving, DRIVEARENA, which can provide realistic surround images and integrate seamlessly with existing vision-based driving agents.
  - DRIVEARENA supports simulation using road networks from any city worldwide, enabling the creation of diverse driving scenario images with varying styles.
  - The Driving Agent, Traffic Manager, and World Dreamer communicate via network interfaces, enabling a highly flexible and modular framework. It allows each component to be replaced with different methods without requiring specific implementations.
- **Vision-Centric 4D Occupancy Forecasting and Planning via World Models** *AAAI'25 Oral* 2024.04-2024.8
  - We propose Drive-OccWorld, a vision-centric world model designed for forecasting 4D occupancy and flow, and we explore the integration of the future forecasting capabilities of world models with end-to-end planning.
  - We design a simple yet efficient semantic- and motion-conditional normalization module for semantic enhancement and motion compensation, which improves forecasting and planning performance.
  - We provide a unified conditioning interface that integrates flexible action conditions into future generations, enhancing the controllability of Drive-OccWorld and facilitating a broader range of downstream applications.

2. LLM agents

- **O^2-Searcher: A Searching-based Agent Model for Open-Domain Open-Ended Question Answering** 2025.3-2025.05
  - We introduce a novel RL-based search agent O^2-Searcher, which dynamically acquires and flexibly utilizes external knowledge via an efficient, local search environment. This design enables an effective decoupling of LLM's internal knowledge from its sophisticated reasoning processes.
  - We propose a unified training mechanism, allowing the agent to efficiently handle both open-ended and closed-ended question types. Through meticulously designed reward functions, O^2-Searcher learns to identify problem types and adaptively adjust its answer generation strategies.

- We construct O^2-QA, a high-quality open-domain QA benchmark specifically designed to evaluate LLMs' performance on complex open-ended questions. It comprises 300 manually curated open-ended questions from diverse domains, along with $\sim 30k$ associated cached web pages.
- Extensive experiments show O^2-Searcher, which uses a 3B-base LLM, significantly outperforms other SOTA LLM agents on O^2-Searcher. It also achieves SOTA on multiple closed-ended QA benchmarks against comparably-sized models, with performance matching 7B models.

- **LeapAD: A Dual-Process Approach to Autonomous Driving** *NeurIPS'24* 2023.11-2024.05
  - Through performing SFT on collected datasets (Rank2Tell, DriveLM, and CARLA), we empower VLM models such as QwenVL to perceive the scenarios of autonomous driving and reason the key objects and potential risks.
  - We propose a dual-process decision-making module inspired by human cognition theory. Without human involvement, our approach enables the fast, empirical Heuristic Process to inherit the capabilities of the slow, rational Analytic Process in a self-supervised manner.
  - We have explored combining in-context learning, external memory, and reflection modules to achieve continuous learning in closed-loop autonomous driving.

3. 3D perception

- **SGN: Camera-based Semantic Scene Completion with Sparse Guidance Network** *TIP'24* 2023.5-2023.08
  - We propose an end-to-end camera-based SSC framework called SGN, propagating semantics from the semantic- and occupancy-aware seed voxels to the whole scene based on geometry prior and occupancy information.
  - We adopt the dense-sparse-dense design and propose hybrid guidance and effective voxel aggregation to enhance intra-categories feature separation and expedite the convergence of the semantic diffusion.
  - Extensive experiments on the SemanticKITTI and SSCBench-KITTI-360 benchmarks demonstrate the effectiveness of our SGN, which is more lightweight and achieves the new state-of-the-art (with only 12.5 M parameters and 7.16 G memory for training).
- **CenterLPS: Segment Instances by Centers for LiDAR Panoptic Segmentation** *MM'23* 2022.12-2023.05
  - We propose a new detection-free and clustering-free framework, dubbed as CenterLPS, with the paradigm of center-based instance encoding and decoding for LiDAR panoptic segmentation.
  - We develop a sparse center proposal network based on the pseudo heatmap to predict instance centers and feature embedding, which can well capture object information of instances.
  - A center-aware transformer is designed to collect context between different center feature embeddings and around centers. The center-based queries facilitate the learning of the transformer.
  - We introduce dynamic convolution with position/shape priors to decode instance masks. A mask fusion module is devised to unify the semantic and instance predictions.

4. Video segmentation

- **Delving Deeper Into Mask Utilization in Video Object Segmentation** *TIP'22* 2021.09-2022.03
  - We provide a unified testbed for eight different VOS encoders to investigate an effective mask fusion strategy. This is the first work to compare a wide range of VOS models from the perspective of mask utilization under the same experimental conditions.
  - We explore the mask benefits on the matcher and propose an insightful mask-enhanced matcher to eliminate the background distraction and enhance the target feature in the matching process.
  - We propose a new network, dubbed MaskVOS, which sufficiently makes use of the reference masks in both the encoder and matcher. The effectiveness of our model is demonstrated on three benchmark datasets, highlighting the importance of effectively using the mask in VOS.
- **Fast Real-time Video Object Segmentation with a Tangled Memory Network** *ACM TIST'23* 2020.12-2021.06
  - We propose a fast real-time Tangled Memory Network (TMN) to excavate the rich target information like contour and edge features contained in the mask, which is not fully exploited by existing methods. Several variants with different model sizes are implemented to adapt to different application platforms easily.
  - We design a Target State Estimator (TSE) to predict an IoU score, providing reliable mask prediction feedback during online inference. Moreover, based on this score, a simple memory bank organization mechanism is devised to maintain the memory bank efficiently.

- We conduct comprehensive experiments on the public benchmarks DAVIS and YouTube-VOS, demonstrating that our method obtains competitive results while running at high speed (66 FPS on the DAVIS16-val set).

INTERNSHIP

- **Shanghai Artificial Intelligence Laboratory** 2023.12 – Research Intern

1. LLM agents

- Developing **O^2-Searcher** and **O^2-QA benchmark**, an advanced search agent designed to autonomously search the internet, delivering precise answers for deterministic question answering tasks and generating detailed key findings for open-ended queries, powered by an innovative R1-inspired training paradigm.
- Core developer of **MiGo** for **DeepResearch**. Developing a research agent as an intelligent assistant for in-depth academic paper comprehension and open-ended report generation.
- Developing **LeapAD** and exploring the application of LLMs in autonomous driving. We aim to develop a closed-loop AD system capable of reasoning about unseen scenarios and utilizing knowledge in a human cognition manner, along with a closed-loop learning process involving continuous interaction and exploration combined with rational analysis.

2. Driving simulation

- Focusing on diffusion-based autonomous driving scene generation. We developed the first high-fidelity closed-loop simulator for autonomous driving, **DRIVEARENA**. This simulator provides realistic surround images and integrates seamlessly with existing vision-based driving agents.
- Developing **DreamForge**, long video generation for driving simulation, enhancing scene modeling through perspective guidance and object-wise position encoding, capturing motion dynamics with motion-aware temporal attention, and integrating the DRIVEARENA simulation platform for coherent scene generation and reliable evaluation.

AWARDS

- Outstanding Graduate of Zhejiang University Academic Scholarship of Zhejiang University
- ECCV W-CODA Corner Case Scene Generation 2nd Place Award

PUBLICATIONS (selected)

[1] **Jianbiao Mei***, Yukai Ma*, Xuemeng Yang, et al., Continuously Learning, Adapting, and Improving: A Dual-Process Approach to Autonomous Driving, Advances in Neural Information Processing Systems (**NeurIPS**), 2024.

[2] **Jianbiao Mei**, Yu Yang, Mengmeng Wang, et al., Camera-based Semantic Scene Completion with Sparse Guidance Network, IEEE Transactions on Image Processing (**TIP**), 2024.

[3] **Jianbao Mei**, Yu Yang, Mengmeng Wang, et al., LiDAR Video Object Segmentation with Dynamic Kernel Refinement, Pattern Recognition Letters (**PRL**), 2024.

[4] **Jianbao Mei***, Yu Yang*, Mengmeng Wang, et al., CenterLPS: Segment Instances by Centers for LiDAR Panoptic Segmentation, In Proceedings of the 31st ACM International Conference on Multimedia (**ACM MM**), 2023.

[5] **Jianbiao Mei***, Yu Yang*, Mengmeng Wang, et al., PANet: LiDAR Panoptic Segmentation with Sparse Instance Proposal and Aggregation, IEEE/RSJ International Conference on Intelligent Robots and Systems (**IROS**), 2023.

[6] **Jianbiao Mei**, Yu Yang, et al., SSC-RS: Elevate LiDAR Semantic Scene Completion with Representation Separation and BEV Fusion, IEEE/RSJ International Conference on Intelligent Robots and Systems (**IROS**), 2023.

[7] **Jianbiao Mei***, Mengmeng Wang*, Yu Yang, et al., Fast Real-time Video Object Segmentation with a Tangled Memory Network, ACM Transactions on Intelligence Systems and Technology (**ACM TIST**), 2023.

[8] Yu Yang*, **Jianbiao Mei***, Liang Liu, et al., DQFormer: Towards Unified LiDAR Panoptic Segmentation with Decoupled Queries, IEEE Transactions on Geoscience and Remote Sensing (**TGRS**) 2025.

[9] Yu Yang*, **Jianbiao Mei***, Yukai Ma, et al. Driving in the Occupancy World: Vision-Centric 4D Occupancy Forecasting and Planning via World Models for Autonomous Driving. AAAI Conference on Artificial Intelligence (**AAAI Oral**), 2025.

[10] Yukai Ma*, **Jianbiao Mei***, Xuemeng Yang, et al., LiCROcc: Teach Radar for Accurate Semantic Occupancy Prediction using LiDAR and Camera, IEEE Robotics and Automation Letters (**RAL**), 2024.

[11] Mengmeng Wang*, **Jianbiao Mei***, Lina Liu, et al., Delving Deeper Into Mask Utilization in Video Object Segmentation, IEEE Transactions on Image Processing (**TIP**), 2022.