Yaoting Wang

Phone: +86 156-1553-8152 | Email: yaoting.wang@outlook.com

EDUCATION

University of Edinburgh 2021.09 - 2022.09

MSc Speech and Language Processing

- Distinction MSc dissertation
- Merit-class degree

National University of Limerick, Ireland 2019.07 - 2021.07

BSc Computer Systems

- First-Class Honours with GPA 3.82 / 4.00
- Full Awarded Scholarship (2020)
- Half Awarded scholarship (2019)

Shandong University of Technology 2017.09 - 2021.06

BSc Computer Science Bachelor

- First-class scholarship (2017, 2018)

PROFESSIONAL EXPERIENCE

King Abdullah University of Science and Technology 2024.03 - Present

Visiting Student Research Program Vision-CAIR

- Advised by Prof. Mohamed h. Elhoseiny for vision-language research.

Renmin University of China (RUC) 2023.01 - Present

Research Assistant Intern, GeWu-Lab Beijing

- Assist in research related to multimodal scene understanding: language, vision and audio.

Deepwise 2023.01 - 2023.03

NLP Intern, Deepwise AI Lab Beijing

- Responsible for the research and development of multimodal pneumonia classification using medical text and images on real patient data from the National Institutes for Food and Drug Control.

RESEARCH EXPERIENCE

MiniGPT4-AVSpeech: Towards Robust Audio-Visual Speech Understanding in Overlapped Speech 2024.04 - Present

- We build the first multimodal LLM for robust audio-visual speech understanding in overlapped speech scenarios.
- We build the first dataset Audio-Visual Speech Understanding Bench (AVSU-Bench) with 20K audio-visual speech question-answer pairs for training and accessing the robustness of audio-visual speech models.
- We use the shared-QFormer to allow the lip modality incorporating with speech semantics during the lip-text alignment phase.
- *To be submitted to CVPR 2025.*

AVTrustBench: Assessing Reliability and Robustness in Audio-Visual LLMs 2024.03 - 2024.05

- We want to explore the hallucination of audio-visual LLMs, especially for question-answer task.
- The proposed benchmark (~600K samples) and findings reveal that the majority of existing models fall significantly short of achieving human-like comprehension.
- *Submitted to NeurIPS 2024.*

Ref-AVS: Refer and Segment Objects in Audio-Visual Scenes with Natural Language 2023.12 - 2024.03

- We propose Ref-AVS as a challenging scene understanding task that segments objects of interest with multimodal-cue natural language expressions, and provide the corresponding Ref-AVS benchmark for performance training and validation.
- We design an end-to-end framework for Ref-AVS that efficiently processes the multimodal cues with a crossmodal transformer, serving as a feasible research framework for future development.
- Our work can inspire more methods to build better convenience and accessibility for the visually and hearing impaired population.
- *Accepted by ECCV'24 Main Track.*

Can Textual Semantics Mitigate Sounding Object Segmentation Preference? 2023.09 - 2023.11

- We use multimodal LLM for visual scene understanding and obtain potential sound objects as text cues to enhance audio-visual correlation with language as the bridge, providing further precise guidance for segmentation models.
- We designed task-specific few-shot prompt template with CoT to assist LLM reasoning step-by-step and obtain more accurate text cues.
- We propose a Prompting Mask Queries with Semantics module to seamlessly introduce audio and semantic instructions into visual foundation model like Mask2Former.
- *Accepted by ECCV'24 Main Track.*

Prompting Segmentation with Sound is Generalizable Audio-Visual Source Localizer 2023.04 - 2023.08

- Current methods for the Audio-Visual Segmentation task are based on the encoder-fusion-decoder paradigm and fail to address the challenges posed by limited data and varying data distributions as they do not leverage the prior knowledge of pre-trained models effectively.
- We propose our GAVS model follow the encoder-prompt-decoder paradigm. We introduce the Semantic-aware Audio Prompt to assist the visual foundation model in querying sounding objects from the visual space using audio cues.
- We propose the Correlation Adapter, which minimizes effort in adjusting the visual foundation model to establish cross-modal audio-visual correlation.
- Our method outperforms fusion-based methods significantly in both unseen classes and cross-dataset settings.
- *Accepted by AAAI'24 Main Track and ICCV'23 Workshop.*

Incongruity-Aware Hierarchical Crossmodal Transformer with Dynamic Modality Gating: A Study on Affect Recognition 2022.08 - 2022.12

- We explore how affective information in different modalities can be influenced by each other, specifically highlighting the presence of latent inter-modal incongruity in crossmodal attention.
- To address this issue, we present the Hierarchical Crossmodal Transformer with Dynamic Modality Gating (HCT-DMG), a lightweight model that effectively reduces fusion times by dynamically choosing the primary modality.
- On five benchmarks: MUStARD, UR-Funny, CMU-MOSI, CMU-MOSEI, and IEMOCAP, we not only performing better than baseline methods, but also significantly reducing trainable parameters (<1M).
- In TACL major revision.

Edge Computing 2020.12 - 2021.02

- Assist tutors to expand mobile business selection in edge computing environment with Cuckoo Search algorithm.
- Responsible for obtaining RESTful API information through crawlers.
- Responsible for building a RESTful service intelligent invocation framework using Python.
- *Accepted by the IEEE SCC'21.*

LEADERSHIP EXPERIENCE

Vice Minister, University Science and Technology Innovation Service Center 2018.05 - 2019.05

Secretary, Science and Technology Service Center; Student Union of Computer Science School 2017.10 - 2018.03

Works

1. MiniGPT4-Speech: Towards Robust Audio-Visual Speech Understanding in Overlapped Speech. **Yaoting Wang**, Aditya Katkuri, Shaoxuan Xu, Jian Ding, Jun Chen, Di Hu, Mohamed Elhoseiny. (2024). *To be submitted to CVPR 2025*.
2. AVTrustBench: Assessing Reliability and Robustness in Audio-Visual LLMs. Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, **Yaoting Wang**, Ruohan Gao, Mohamed Elhoseiny, Dinesh Manocha. (2024). *Submitted to NeurIPS 2024*.
3. Can Textual Semantics Mitigate Sounding Object Segmentation Preference?. **Yaoting Wang***, Peiwen Sun*, Yuanchao Li, Honggang Zhang, Di Hu^. (2024). *Accepted by The 18th European Conference on Computer Vision (ECCV 2024). arxiv.*
4. Ref-AVS: Refer and Segment Objects in Audio-Visual Scenes with Natural Language. **Yaoting Wang***, Peiwen Sun*, Dongzhan Zhou, Guangyao Li, Honggang Zhang, Di Hu^. (2024). *Accepted by The 18th European Conference on Computer Vision (ECCV 2024). arxiv.*
5. Stepping Stones: A Progressive Training Strategy for Audio-Visual Semantic Segmentation. Juncheng Ma, Peiwen Sun, **Yaoting Wang**, Di Hu^. (2024). *Accepted by The 18th European Conference on Computer Vision (ECCV 2024).* arxiv.
6. Prompting Segmentation with Sound is Generalizable Audio-visual Source Localizer. **Yaoting Wang***, Weisong Liu*, Guangyao Li, Jian Ding, Di Hu^, Xi Li. (2023). *Accepted by 38th AAAI conference on artificial intelligence (Main track) & ICCV'23 (Workshop).* arxiv.
7. Incongruity-Aware Hierarchical Crossmodal Transformer with Dynamic Modality Gating: A Study on Affect Recognition. **Yaoting Wang***, Yuanchao Li*, Paul Pu Liang, Louis-Philippe Morency, Peter Bell and Catherine Lai^. (2023). *in TACL major revision.* arxiv.
8. Scaling up mobile service selection in edge computing environment with cuckoo optimization algorithm. Ming Zhu, Feilong Yu, Xiukun Yan, Jing Li, **Yaoting Wang**. (2021, September). Accepted by 2021 IEEE International Conference on Services Computing (SCC) (pp. 394-400). IEEE.