











# Shiyu Hu (胡世宇)




## Research Fellow, Nanyang Technological University (NTU)

 1995.10.29    shiyu.hu@ntu.edu.sg    hushiyu1995    huuuuusy  
 https://huuuuusy.github.io/    http://viig.aitestunion.com/

## Professional Summary

- Excellent Education**  I obtained my bachelor's, master's, and doctoral degree from top universities/research institutions in China, and all theses defense were excellent.
-  I received numerous awards and honors, including the National Scholarship (top 1%) and Beijing Outstanding Graduates (top 5%).
- Solid Foundation**  During my doctoral studies, I have published 14 papers, of which 5 are first-author/corresponding-author publications – including 3 papers in IEEE TPAMI and IJCV, 1 paper in NeurIPS, and a survey in Journal of Images and Graphics. Besides, invited by Springer, I will complete a book in Dec. 2024.
-  The research platform I built and maintained has received over 382k visits from 130+ countries and regions worldwide (until Jun. 2024).
- Wide Communication**  I served as a reviewer for top conferences and journals such as NeurIPS, CVPR, ECCV, AAAI, ACM MM, SCIENCE CHINA Information Sciences, etc., and will conduct a tutorial at the 31th ICIP, 27th ICPR, and 17th ACCV.
-  Since Sep. 2022, I have initiated and organized interdisciplinary seminars based on computer vision (involving 10+ schools and 20+ individuals), covering research areas like computer vision, cognitive science, and human-computer interaction.
-  I assisted and supervised nearly 10 bachelor's, master's, and doctoral students. Besides, I established the Visual Intelligence Interest Group (VIIG) and work with these students to promote research in related directions.




## Work Experience

- 2024.08 - Now  **Research Fellow, School of Physical and Mathematical Sciences (SPMS), Nanyang Technological University (NTU)**
- **Direction:** AI4Science, Computer Vision
  - **PI:** Assoc Prof. Kanghao Cheong (IEEE Senior Member)
- 2018.03 - 2018.11  **Research Assistant, University of Hong Kong (HKU)**
- **Direction:** High Performance Computing, Heterogeneous Computing
  - **PI:** Prof. Choli Wang
- 2016.08 - 2016.09  **Research Intern, Institute of Electronics, Chinese Academy of Sciences (CASIE)**













## Education Background

- 2019.09 - 2024.01  **Ph.D, Institute of Automation, Chinese Academy of Sciences (CASIA)**
- **Major:** Computer Applied Technology
  - **Supervisor:** Prof. Kaiqi Huang (IAPR Fellow, IEEE Senior Member)
  - **Co-supervisor:** Prof. Xin Zhao (IEEE Senior Member)
  - **Thesis title:** *Research of Intelligence Evaluation Techniques for Single Object Tracking*
  - **Thesis committee:** Prof. Jianbin Jiao, Prof. Yuxin Peng, Prof. Yao Zhao (IEEE Fellow, IET Fellow), Prof. Yunhong Wang (IEEE Fellow, IAPR Fellow, CCF Fellow), Prof. Ming Tang
  - **Thesis defense grade:** Excellent

## Education Background (continued)

- 2017.09 - 2019.07  **M.Sc., Department of Computer Science, University of Hong Kong (HKU)**
- **Major:** Computer Science
  - **Supervisor:** Prof. Choli Wang
  - **Thesis title:** *NightRunner: Deep Learning for Autonomous Driving Cars after Dark*
  - **Thesis defense grade:** A+
- 2013.09 - 2017.07  **B.E., Elite Class in School of Information and Electronics, Beijing Institute of Technology (BIT)**
- **Major:** Information Engineering
  - **Thesis title:** *Text Sentiment Analysis Based on Deep Neural Network*
  - **Thesis defense grade:** Excellent
- 2015.07 - 2015.08  **Summer Semester, University of California, Berkeley (UCB)**
- **Major:** New Media
  - **Course grade:** A

## Research Foundation & Interests

- Visual Object Tracking  Research on single object tracking algorithms in general scenes and specific scenarios (e.g., unmanned aerial vehicles).
-  Research on the robustness, generalization, and security of single object tracking algorithms.
- Visual Language Tracking  Research on multi-modal tracking, video understanding, and visual reasoning tasks based on long video sequences.
-  Exploring using Large Language Models (LLMs) and Large Vision Models (LVMs) for long video understanding.
-  Exploring human-computer interaction patterns in long video sequences with visual language tracking as a proxy task.
- Benchmark Construction  Research on the construction strategy of single-modal and multi-modal datasets incorporating human knowledge structure.
-  Research on evaluation mechanisms for robustness, generalization, and safety.
- Intelligent Evaluation  Design of a human-machine universal visual ability evaluation framework.
-  Benchmarking the performance of algorithms based on human abilities in perceptual, cognitive, inferential, etc. Analyzing the bottlenecks of algorithms and human subjects in depth, providing guidance for research on human-like modeling, human-machine collaboration, and human-machine integration.
- AI4Science  **Cognitive Science:** Visual task design, environment construction, and human-machine capability analysis based on human-like modeling principles.
-  **Medical Science:** Research on medical image processing techniques based on artificial intelligence technologies (e.g., cell segmentation and tracking).
-  **Psychology:** Development of gamified assessment systems targeting psychological dimensions such as anxiety, depression, and obsession, along with research on intelligent psychological evaluation technologies. Exploring using LLMs and LVMs for visual comprehension with psychological elements.
-  **Education:** Research on human-computer interaction (HCI) technology for education scenarios, including designing an intelligent education framework from a multidisciplinary perspective, investigating HCI technology, conducting qualitative and quantitative analysis.

## Research Experiences

### Main Research

#### ☰ Idea & Method Design & Experimental Analysis & Paper Writing & Platform Development

2018.03 - 2018.11 📌 **Darknet-Cross: Light-weight Deep Learning Framework for Heterogeneous Computing**

🔧 High-performance Computing & Heterogeneous Computing & Deep Learning Framework

🔧 Darknet-Cross is a lightweight deep learning framework, mainly based on the open-source deep learning algorithm library Darknet and yolov2\_light, and it has been successfully ported to mobile devices through cross-compilation. This framework enables efficient algorithm inference using mobile GPUs.

🔧 Darknet-Cross supports algorithm acceleration processing on various platforms (e.g., Android and Ubuntu) and various GPUs (e.g., Nvidia GTX1070 and Adreno 630).

✓ The work is a part of my master's thesis at the University of Hong Kong (thesis defense grade: A+).

2019.11 - 2022.02 📌 **VideoCube: A Large-scale Multi-dimensional Global Instance Tracking Intelligent Evaluation Platform**

🔧 Visual Object Tracking & Large-scale Benchmark Construction & Intelligent Evaluation Technology

🔧 This work builds upon the concept of human-like modeling and expands the definition of single object tracking (SOT) task. It presents a new task called global instance tracking (GIT), which broadens the range of applications to adversarial scenarios.

🔧 This work proposes a video narrative content decoupling framework based on film theory, and builds a large-scale, multi-dimensional global instance tracking task intelligent evaluation platform called VideoCube, which includes 7.46 million video frames. It is currently the largest SOT benchmark in terms of scale.

🔧 Starting from human-computer confrontation, for the first time, human subjects are introduced into the SOT task in order to measure the visual tracking intelligence.

✓ This work has been published by IEEE TPAMI in Jan. 2023 [A1]. As of Jun. 2024, the platform has received over 382k visits from more than 130 countries and regions worldwide, with over 1,100 downloads and more than 400 algorithm tests.

2023.03 - 2023.09 📌 **MGIT: A Multi-modal Global Instance Tracking Benchmark Based on Hierarchical Semantic Framework**

🔧 Visual Language Tracking & Long Video Understanding and Reasoning & Hierarchical Semantic Information Annotation

🔧 This work extends the GIT task and the VideoCube benchmark by constructing a multi-modal benchmark called MGIT. The MGIT benchmark is designed to capture the complex video narrative relationships and fully encompass the intricate spatio-temporal and causal connections illustrated in long videos.

🔧 This work introduces an innovative multi-granularity semantic information annotation strategy by incorporating the hierarchical structure of human cognition. The strategy aims to provide high-quality semantic information and its effectiveness is validated through experiments.

🔧 This work introduces an evaluation mechanism for the multi-modal SOT task. It conducts a comprehensive experimental analysis on algorithms with various structures, with the goal of thoroughly examining the performance bottlenecks.

✓ This work has been accepted by NeurIPS in Sep. 2023 [A2].

## Research Experiences (continued)

2021.07 - 2023.10

### 📖 **SOTVerse: A User-defined Single Object Tracking Task Space**

🔧 Visual Object Tracking & Dynamic Open Environment Construction & Visual Evaluation Technique

✍️ The proposed task analysis framework, known as the 3E paradigm (where task is defined as the combination of environment, evaluation, and executor), aims to facilitate the efficient utilization of research resources in the evaluation process.

✍️ The research integrates SOT datasets to transform the original static and closed data space into a dynamic open data environment named SOTVerse, comprised of 12.56 million frames. Furthermore, a subspace construction algorithm is introduced to aid researchers in promptly identifying highly challenging sequences and constructing specialized experimental environments aligned with their research objectives.

✍️ A new evaluation system is proposed to adapt to various evaluation needs and task objectives. It conducts a fine-grained analysis of 23 representative algorithms to effectively identify performance bottlenecks in these algorithms.

✓ This work has been accepted by IJCV in Sep. 2023 [A3].

2022.05 - 2023.10

### 📖 **BioDrone: A Bionic Drone-based Single Object Tracking Benchmark for Robust Vision**

🔧 Visual Object Tracking & Drone-based Tracking & Robust Visual Research

✍️ *BioDrone* is the first bionic drone-based SOT benchmark, it features videos captured from a flapping-wing UAV system with a major camera shake due to its aerodynamics. BioDrone highlights the tracking of tiny targets with drastic changes between consecutive frames, providing a new robust vision benchmark for SOT.

✍️ Twenty representative algorithms have been replicated and tested on BioDrone, followed by a detailed analysis of the robustness bottlenecks.

✍️ A baseline algorithm named UAV-KT has been designed. Experimental analysis has been conducted to validate the effectiveness of this method in extracting visual features and maintaining robustness in challenging factors present in UAV scenes.

✓ This work supported the organization of the 3rd High-Speed Low-Power Visual Understanding Challenge as competition data from May to Oct. 2022. The work has been accepted by IJCV in Oct. 2023 [A4].

2022.04 - Now

### 📖 **Intelligent Evaluation Techniques for Visual Object Tracking Based on Visual Turing Test**

🔧 Visual Object Tracking & Evaluation Technique & Visual Turing

✍️ This work proposes the visual Turing test evaluation paradigm by incorporating the concept of the Turing test, enabling a comprehensive assessment of the visual intelligence of algorithms in comparison to human visual capabilities.

✍️ A controlled experimental environment has been developed to facilitate a fair comparison of dynamic visual abilities between humans and machines. This environment incorporates the perceptual and cognitive capabilities that task objects necessitate during the execution of dynamic visual tasks.

✍️ A suitable task object is selected to conduct tests on human-machine dynamic visual abilities, involving 20 representative algorithms and 15 human subjects.


✍️ A universally applicable multi-scale dynamic visual task evaluation framework has been designed. This framework employs center point distance to assess and analyze tasks at three distinct scales, namely frame-level, sequence-level, and group-level.


✓ Two Chinese review papers were published in 2021 and 2023 [A5], [A6]. The experimental content and main conclusions are being finalized [P1], preparing for submission to the Cell Patterns journal [O1]. Besides, invited by Springer, a book will be completed in Dec. 2024 [O2].


## Research Experiences (continued)

### 2024.03 - Now **FIOVA: A Five-in-One Video Annotations Benchmark for Better Human-Machine Comparison**

 Large Vision-Language Models & Evaluation Technique & Visual Turing

 This work proposes the FIOVA benchmark to evaluate the differences between LVLMs and human understanding in video description tasks. FIOVA includes 3,002 long video sequences annotated by five annotators per video, resulting in captions 4 to 15 times longer than existing benchmarks, establishing a robust baseline representing human understanding comprehensively for the first time.


 Conducted an in-depth evaluation of six state-of-the-art LVLMs (VideoLLaMA2, LLaVA-NEXT-Video, Video-LLaVA, VideoChat2, Tarsier, ShareGPT4Video) using the FIOVA benchmark, revealing significant discrepancies between LVLMs and humans, especially in complex videos where human annotators showed substantial disagreement, with LVLMs relying on uniform strategies.


 Highlighted the limitations of using a single annotator as the evaluation groundtruth and proposed new evaluation perspectives to capture differences in semantic understanding, descriptive depth, and consistency, providing guidance for achieving human-level video comprehension in the future.


✓ This work was submitted to ICLR in Oct. 2024 and is under review [R1].

### **A New AI for Education Pipeline to Model More Human-like and Personalised Early Adolescents**

 AI4Education & LLMs & LLM-based Agent

 This work proposes an AI4Education framework, SOE (Scene - Object - Evaluation), to systematically construct LVSA (LLM-based Virtual Student Agents), using a dataset of personalized teacher-student interactions and fine-tuning LLMs with LoRA, conducting multi-dimensional evaluation experiments.

 Developed a theoretical framework for generating LVSA and integrated human subjective evaluation metrics into GPT-4 assessments, demonstrating a strong correlation between human evaluators and GPT-4 in judging LVSA authenticity.


 Validated the capability of LLMs to generate human-like, personalized virtual student agents in educational contexts, laying a foundation for future applications in pre-service teacher training and multi-agent simulation environments.


✓ This work was submitted to ICLR in Oct. 2024 and is under review [R2].

## Independent Developer

### Platform Maintenance & Upgrade

### 2020.07 - 2024.01 **GOT-10k: A Large High-diversity Benchmark and Evaluation Platform for Single Object Tracking**

 Visual Object Tracking & Evaluation Technology & Platform Maintenance

 GOT-10k is constructed to evaluate the generalization ability of trackers on unseen object classes and motion patterns. The platform provides a high-quality video trajectory dataset containing 10,000 video segments, 563 object classes, 87 motion patterns, and 1.5 million tight annotations, where its coverage of object classes is magnitudes wider than other existing tracking benchmarks.

✓ GOT-10k is the supporting platform for a research accepted by IEEE TPAMI. It receives 3.24m page views, 6.7k+ downloads, 17.9k+ trackers from 150+ countries, and gets 18× page views increase after maintenance (statistics by Feb. 2024).

## Research Experiences (continued)

### Collaborative Research

#### ☰ Idea Discussions & Experimental Analysis & Paper Revision

##### 2019.05 - 2019.10 📖 **A Skin Color Detection System without Color Atlas**

🔧 Color Constancy & Skin Color Detection & Illumination Estimation

✍ Under 18 different environmental lighting conditions and with 4 combinations of smart-phone parameters, skin color data was collected from 110 participants. The skin color dataset consists of 7,920 images, with the testing results from CK Company's MPA9 skin color detector serving as the ground truth for user skin colors.

✍ Using an elliptical skin model, the essential skin regions are extracted from the images. The open-source color constancy model, FC<sup>4</sup>, is employed to recover the environmental lighting conditions. Subsequently, the skin color detection results for users are calculated using SVR regression.

✓ The related work has been successfully deployed in Huawei's official mobile application *Mirror* for its AI skin testing function.

##### 2020.11 - 2021.03 📖 **A Project for Cell Tracking Based on Deep Learning Method**

🔧 Medical Image Processing & AI4Science & Cell Segmentation and Tracking

✍ This method follows the tracking by detection paradigm and combines per-frame CNN prediction for cell segmentation with a Siamese network for cell tracking.

✓ This project was submitted to the cell tracking challenge in Mar. 2021, and maintains the second place in the Fluo-C2FL-MS<sup>+</sup> dataset and the third place in the Fluo-C2FL-Huh7 dataset (statistics by Oct. 2023).

##### 2023.08 - 2023.12 📖 **Robust Single-particle Cryo-EM Image Denoising and Restoration Research**

🔧 Medical Image Processing & AI4Science & Diffusion Model

✍ The cryo-electron microscopy at low temperatures can reveal molecular information at almost atomic scale through the reconstruction of 2D micrographs. However, the reconstruction process requires overcoming low signal-to-noise ratio and complex noise structures. This work proposes a diffusion model with a post-processing module to effectively denoise and restore single-particle cryo-EM images.

✍ The effectiveness of the method is validated through experimental results on simulated and real datasets.

✓ The work has been accepted by IEEE ICASSP in Dec. 2023 [A7].

##### 2023.07 - 2024.09 📖 **Robust Visual Language Tracker Based on Human Memory Modeling**

🔧 Visual Language Tracking & Human-like Memory Modeling & Adaptive Prompts

✍ In response to the inadequate robustness of visual language tracking algorithms in long videos, a robust multi-modal tracker called MemVLT is designed from the perspective of human-like memory modeling.

✍ Inspired by the cognitive theory of Complementary Learning System, MemVLT consists of memory storage and interaction modules. It aims to simulate the complex modulation process of human memory between the hippocampus and the neocortex.

✍ The effectiveness of MemVLT is validated through experiments on multiple representative benchmarks for visual language tracking.

✓ This work has been accepted by NeurIPS in Sep. 2024 [A8].



## Research Experiences (continued)

2023.12 -2024.09

### 📖 **Human-like Visual Object Tracking via Visual Search Ability**

📌 Visual Object Tracking & Visual Search Mechanism & Visual Turing Test

✍️ For the difficulty that VOT trackers cannot cope with spatio-temporal discontinuous (STD) scenes, a robust visual target tracking algorithm CPDTrack is designed from the perspective of human-like visual search mechanism modeling. Inspired by the pathways of the human visual system, CPDTrack employs an efficient encoding method for both the local and global visual fields.

✍️ To verify the effectiveness of CPDTrack, a new benchmark STDChallenge is proposed, which consists of challenging sequences with spatio-temporal discontinuities. Additionally, a Visual Turing Test is conducted to measure and quantify the visual search abilities of human subjects.

✍️ The effectiveness of CPDTrack has been verified on the STDChallenge benchmark. Human-computer error consistency analysis in the Visual Turing Test demonstrates that CPDTrack's decision-making is highly similar to that of human subjects.

✓ This work has been accepted by NeurIPS in Sep. 2024 [A9].

2022.10 - Now

### 📖 **Unconstrained Air-writing Technique for Real-World Applications**

📌 Air-writing Technique & Benchmark Construction & Human-machine Interaction

✍️ This study has developed a large-scale and high-quality video dataset named AWCV-100k, which consists of air-writing of Chinese characters. The objective is to establish a more natural and comprehensive experimental environment for human-computer interaction (HCI) research.

✍️ The AWCV-100k dataset comprises 8.8 million video frames, encompassing diverse environmental settings and lighting conditions. It provides comprehensive coverage of 3,755 Chinese characters from the GB2312-80 character set, establishing it as the most extensive and comprehensive air-writing video dataset currently accessible.

✍️ An air-writing character recognition algorithm called VCRec is proposed. This baseline algorithm is capable of extracting fingertip features from sparse visual cues and analyzing them using a spatio-temporal sequence module.


✍️ Representative algorithms and VCRec have been reproduced and tested on the AWCV-100k. Experimental results confirm the robustness and effectiveness of VCRec.


✓ This work has been accepted by IEEE TCSVT in Apr. 2024 [A10]. Subsequent research has built upon the AWCV-100k by incorporating depth information and hand keypoints to create the multi-modal benchmark named MMAW-UCAS2024. Alongside this, a corresponding multi-modal baseline algorithm named MMRec is developed. This research will be submitted to CVPR in Nov. 2024 [O3].


## Research Experiences (continued)

### 2023.01 - Now **Research on Single Object Tracking Task with Similar Object Interference Challenges**

 Visual Object Tracking & Similar Object Interference & Data Mining


 Based on the operational principles and case analyses of mainstream tracking algorithms, this work redefines the challenge of similar object interference from the perspective of algorithms. It focuses on analyzing the cognitive biases of humans and machines when facing similar object interference.


 This study presents a data mining algorithm that enables the automatic extraction of sequences from representative single-object tracking SOT datasets. The extracted sequences consist of instances where similar object interference occurs, and are used to create the TrackingSOI dataset. The extraction process is performed without any manual intervention.


 An algorithm called TransKT is proposed to effectively handle similar object interference. It is capable of distinguishing candidates that have similar appearance information to the target object, thereby achieving robust visual object tracking ability.


✓ A simplified version of this work has been accepted as an oral paper by the CSAI conference in Nov. 2023 [A11]. The full version was submitted to IEEE TCSVT in Jan. 2024 and is currently under review [R3].

### 2023.02 - Now **Intelligent Psychological Assessment System based on Electronic Sandplay**

 Psychological Assessment System & Gamified Assessment & AI4Science

 An intelligent psychological assessment system based on electronic sandbox has been developed to address the limitations of traditional assessments, including participants' lack of inherent motivation and limited insight. It facilitates the evaluation and analysis of various dimensions, such as anxiety, depression, and obsession.


 This system integrates artificial intelligence technology, offering multiple advantages in comparison to traditional questionnaire assessments and gamified assessments. By extracting linked psychological measurement evidence from dynamic process data within games, it improves the authenticity and interpretability of measurements.


 Recruiting participants from educational and public safety settings and conducting validity and reliability tests as well as case analysis, the experimental results confirmed the effectiveness of the electronic sandbox.

✓ The sandbox theme recognition model employed in this work is accepted by PRCV in Sep. 2023 [A12]. As an interdisciplinary and systematic work spanning psychology, game design, and artificial intelligence, the development trajectory of its technical route and framework were submitted to top psychology journals in China [A13], [R4]. The subsequent research proposes VS-LLM, a visual-semantic depression assessment based on LLM for drawing projection test. The work has been accepted by PRCV in June 2024 [A14].

### 2024.01 - Now **Diverse Text Generation for Visual Language Tracking Based on LLM**

 Visual Language Tracking & Large Language Model & Evaluation Technique

 Design a framework named DTLLM-VLT based on LLM to automatically generates extensive and multi-granularity text to enhance environmental diversity.

 Deploy DTLLM-VLT in representative benchmarks and conduct a detailed evaluation of the VLT method. Experimental results demonstrates its potential to enhance the comprehension of vision datasets.

✓ This work has been accepted as an oral paper and won the best paper honorable mention by CVPR 3rd VDU workshop in Apr. 2024 [A15]. The subsequent research is dedicated to the development of a novel visual language tracking benchmark named DTVLT. This benchmark is constructed on the foundation of DTLLM-VLT and includes comprehensive evaluation and analysis. DTVLT was submitted to ICLR in Oct. 2024 and is under review [R5].



## Research Experiences (continued)

### 📖 Research on the Dilemma and Countermeasures of Human-Computer Interaction in Intelligent Education

📌 Intelligent Education Technology & Human-Computer Interaction & AI4Science

✍️ Incorporating insights and methodologies from education, cognitive psychology, and computer science, this project establishes a theoretical framework for understanding the evolution of HCI within the intelligent education.

✍️ Drawing upon the established theoretical framework, this project conducts a comprehensive analysis of the evolution of HCI in educational settings, transitioning from collaboration to integration. Furthermore, it delves into the key issues arising from this transformative process within the realm of intelligent education.

✍️ Building upon the core issues unearthed, this project investigates strategies for leveraging theoretical guidance and technical enhancements to enhance the efficacy of HCI in intelligent education, ultimately striving towards effective human-computer integration.

✓ The project is funded by the 2023 Intelligent Education PhD Research Fund, supported by the Institute of AI Education Shanghai and East China Normal University, and is currently in progress [O4].

### 2024.05 - Now 📖 Leveraging Aligned Target-Context Cues for Robust Vision-Language Tracking

📌 Visual Language Tracking & Multi-modal Learning & Contextual Clues

✍️ Aiming at the limitations of visual language trackers in tracking long sequences, a novel tracker, ATCTrack, is proposed. This method obtains multi-modal cues aligned with the target state through target-context feature modeling to achieve a robust tracking effect.

✍️ For the visual modality, an efficient temporal visual target-context modeling method is proposed to provide timely visual clues for the tracker; for the textual modality, a target word recognition method based on text content is proposed to ensure focus on target-related words while reducing the interference of auxiliary context words.

✍️ Compared with existing text processing methods based on visual-text similarity, ATCTrack simplifies the task while improving accuracy and achieving SOTA performance on representative benchmarks.

✓ This work was submitted to AAAI in Aug. 2024 and is under review [R6].

### 2024.07 - Now 📖 Enhancing Vision-Language Tracking by Effectively Converting Textual Cues into Visual Cues

📌 Visual Language Tracking & Multi-modal Learning & Grounding Model

✍️ Aiming at the imbalance between video and text data in visual language tracking, a plug-and-play method is proposed to convert text clues into visual clues, thereby improving the tracking effect.

✍️ This method leverages the strong text-image alignment capability of the foundation model to convert textual clues into interpretable visual heatmaps. It fuses these heatmaps with the search image via a heatmap guidance module to effectively guide tracking.

✍️ This method simplifies the processing of text prompts by the VLT system and is compatible with existing trackers. Experimental results on multiple mainstream benchmarks demonstrate its effectiveness in enhancing VLT performance and achieving SOTA performance on representative benchmarks.

✓ This work was submitted to ICASSP in Sep. 2024 and is under review [R7].

## Research Experiences (continued)

### ■ A Benchmark for Visual Language Tracking based on Multi-modal Interaction

✚ Visual Language Tracking & Multi-modal Interaction & Evaluation Technology

✍ Aiming at the lack of a multi-round interaction mechanism in current visual language tracking, a new benchmark, VLT-MI, is proposed, introducing the multi-round interaction mechanism for the first time.



✍ A new interaction paradigm is designed to provide more accurate text and bounding boxes after multiple tracking failures, improving the tracker's robustness and expanding the downstream applications of VLT tasks.

✍ By comparing with the traditional VLT benchmark, the tracker's accuracy and robustness under the interactive paradigm proposed in this work are verified, providing the field with a more sophisticated multimodal tracker evaluation method and new ideas for future dataset expansion.

✓ This work was submitted to NeurIPS 1st Workshop on Visual-Language Model in Sep. 2024 and is under review [R8].

## Research Publications

### Acceptance

- 1 **S. Hu**, X. Zhao, L. Huang, and K. Huang, "Global instance tracking: Locating target more like humans," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 45, no. 1, pp. 576–592, 2023. [DOI: 10.1109/TPAMI.2022.3153312](#).
- 2 **S. Hu**, D. Zhang, M. Wu, X. Feng, X. Li, X. Zhao, and K. Huang, "A multi-modal global instance tracking benchmark (mgit): Better locating target in complex spatio-temporal and causal relationship," in *The 37th Conference on Neural Information Processing Systems (NeurIPS, Poster)*, vol. 36, 2023, pp. 25 007–25 030.
- 3 **S. Hu**, X. Zhao, and K. Huang, "Sotverse: A user-defined task space of single object tracking," *International Journal of Computer Vision (IJCV)*, vol. 132, pp. 872–930, 2024. [DOI: 10.1007/s11263-023-01908-5](#).
- 4 X. Zhao , **S. Hu** , Y. Wang, J. Zhang, Y. Hu, R. Liu, H. Ling, Y. Li, R. Li, K. Liu, and J. Li, "Biodrone: A bionic drone-based single object tracking benchmark for robust vision," *International Journal of Computer Vision (IJCV)*, vol. 132, pp. 1659–1684, 2024. [DOI: 10.1007/s11263-023-01937-0](#).
- 5 **S. Hu**, X. Zhao, and K. Huang, "Visual intelligence evaluation techniques for single object tracking: A survey (单目标跟踪中的视觉智能评估技术综述)," *Journal of Images and Graphics* (《中国图象图形学报》), 2023.
- 6 K. Huang, X. Zhao, Q. Li, and **S. Hu**, "Visual turing: The next development of computer vision in the view of human-computer gaming (视觉图灵：从人机对抗看计算机视觉下一步发展)," *Journal of Graphics* (《图学学报》), vol. 42, no. 3, p. 339, 2021. [DOI: 10.11996/JG.j.2095-302X.2021030339](#).
- 7 J. Zhang, T. Zhao, **S. Hu**, and X. Zhao, "Robust single-particle cryo-em image denoising and restoration," in *The 49th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP, Poster)*, 2024.
- 8 X. Feng, X. Li, **S. Hu**, D. Zhang, M. Wu, J. Zhang, X. Chen, and K. Huang, "Memvlt: Visual-language tracking with adaptive memory-based prompts," *The 38th Conference on Neural Information Processing Systems (NeurIPS, Poster)*, 2024.
- 9 D. Zhang, **S. Hu**, X. Feng, X. Li, M. Wu, J. Zhang, and K. Huang, "Beyond accuracy: Tracking more like human via visual search," *The 38th Conference on Neural Information Processing Systems (NeurIPS, Poster)*, 2024.
- 10 M. Wu, K. Huang, Y. Cai, **S. Hu**, Y. Zhao, and W. Wang, "Finger in camera speaks everything: Unconstrained air-writing for real-world," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2024.

- 11 Y. Wang, **S. Hu**, and X. Zhao, "Rethinking similar object interference in single object tracking," in *The 7th International Conference on Computer Science and Artificial Intelligence (CSAI, Oral)*, 2023, pp. 251–258.
- 12 X. Feng, **S. Hu**, X. Chen, and K. Huang, "A hierarchical theme recognition model for sandplay therapy," in *The 6th Chinese Conference on Pattern Recognition and Computer Vision (PRCV, Poster)*, 2023, pp. 241–252.  
DOI: 10.1007/978-981-99-8462-6\_20.
- 13 K. Huang, Y. Kang, C. Yan, **S. Hu**, L. Wang, T. Tao, and W. Gao, "A review of intelligent psychological assessment based on interactive environment (基于交互环境的智能化心理测评)," *Chinese Mental Health Journal* (《中国心理卫生杂志》), 2024.
- 14 M. Wu, Y. Kang, X. Li, **S. Hu**, X. Chen, Y. Kang, W. Wang, and K. Huang, "Vs-llm: Visual-semantic depression assessment based on llm for drawing projection test," *The 7th Chinese Conference on Pattern Recognition and Computer Vision (PRCV, Poster)*, 2024.
- 15 X. Li, X. Feng, **S. Hu**, M. Wu, D. Zhang, J. Zhang, and K. Huang, "Dtllm-vlt: Diverse text generation for visual language tracking based on llm," *3rd Workshop on Vision Datasets Understanding and DataCV Challenge in The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024 (CVPRW, Oral, Best Paper Honorable Mention)*, 2024.
- 16 Y. Zhang, C. Liu, W. Chen, X. Xu, F. Wang, H. Li, **S. Hu**, and X. Zhao, "Revisiting instance search: A new benchmark using cycle self-training," *Neurocomputing (Neu)*, vol. 501, pp. 270–284, 2022. DOI: 10.1016/j.neucom.2022.06.027.

## Preprint

- 1 **S. Hu**, X. Zhao, Y. Wang, Y. Shan, and K. Huang, *Nearing or surpassing: Overall evaluation of human-machine dynamic vision ability*, 2023. URL: [https://openreview.net/forum?id=LGbzyW\\_pnsc](https://openreview.net/forum?id=LGbzyW_pnsc).

## Under Review

- 1 **S. Hu\***, X. Li\*, X. Li, J. Zhang, Y. Wang, X. Zhao, and K. Cheong, "Can lvlms describe videos like humans? a five-in-one video annotations benchmark for better human-machine comparison," *The 13th International Conference on Learning Representations (ICLR, Under Review)*, 2024.
- 2 Y. Ma\*, **S. Hu\***, X. Li, Y. Wang, S. Liu, and K. Cheong, "Students rather than experts: A new ai for education pipeline to model more human-like and personalised early adolescences," *The 13th International Conference on Learning Representations (ICLR, Under Review)*, 2024.
- 3 Y. Wang, **S. Hu**, D. Zhang, M. Wu, T. Yao, Y. Wang, L. Chen, and X. Zhao, "Target or distractor? rethinking similar object interference in single object tracking," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT, Under Review)*, 2024.
- 4 Y. Ren, X. Feng, **S. Hu**, Y. Kang, C. Yan, Y. Zeng, L. Wang, and K. Huang, "Intelligent psychological assessment with sandplay based on evidence-centered design theory (基于证据中心设计理论的智能心理沙盘测评系统)," *Acta Psychologica Sinica* (《心理学报》), *Under Review*, 2024.
- 5 X. Li, **S. Hu**, X. Feng, D. Zhang, M. Wu, J. Zhang, and K. Huang, "Dtvlt: A multi-modal diverse text benchmark for visual language tracking based on llm," *The 13th International Conference on Learning Representations (ICLR, Under Review)*, 2024.
- 6 X. Feng, **S. Hu**, X. Li, D. Zhang, M. Wu, J. Zhang, X. Chen, and K. Huang, "Atctrack: Leveraging aligned target-context cues for robust vision-language tracking," *The 39th Annual AAAI Conference on Artificial Intelligence (AAAI, Under Review)*, 2024.
- 7 X. Feng, D. Zhang, **S. Hu**, X. Li, M. Wu, J. Zhang, X. Chen, and K. Huang, "Enhancing vision-language tracking by effectively converting textual cues into visual cues," *The 50th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP, Under Review)*, 2024.

- 8 X. Li, **S. Hu**, X. Feng, D. Zhang, M. Wu, J. Zhang, and K. Huang, "Visual language tracking with multi-modal interaction: A robust benchmark," *1st Workshop on Video-Language Models in The 39th Conference on Neural Information Processing Systems (NeurIPS-W, Under Review)*, 2024.

## Ongoing Research

- 1 **S. Hu**, J. Zhu, Y. Wang, X. Zhao, and K. Huang, "Vt<sup>3</sup>: A visual tracking turing test of human-machine dynamic vision ability," *Cell Patterns (In Preparation)*, 2024.
- 2 X. Zhao, **S. Hu**, and X. Yin, *Visual Object Tracking - An Evaluation Perspective*. Springer, 2024.
- 3 M. Wu, X. Li, **S. Hu**, Y. Cai, K. Huang, and W. Wang, "Unconstrained multimodal air-writing benchmark: Writing by moving your fingers in 3d," *The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2025 (CVPR, In Preparation)*, 2024.
- 4 Y. Ma, Z. Yang, Y. Kang, X. Xue, and **S. Hu**, "From collaboration to integration: Research on the dilemma and countermeasures of human-computer interaction in intelligent education (从协同走向融合: 智能教育中人机交互的困境及对策研究)," *Intelligent Education PhD Research Fund, supported by the Institute of AI Education Shanghai and East China Normal University (In Progress)*, 2024.

## Paper Summary

- Journal  **TPAMI**: IEEE Transactions on Pattern Analysis and Machine Intelligence (CCF-A Journal, Top-1 journal in computer vision, IF=20.8). Acceptance×1 (first author×1).
-  **IJCV**: International Journal of Computer Vision (CCF-A Journal, Top-2 journal in computer vision, IF=11.6). Acceptance×2 (first author×1, corresponding-author×1).
-  **TCSVT**: IEEE Transactions on Circuits and Systems for Video Technology (CCF-B Journal, IF=8.3). Acceptance×1, under review×1.
-  **JIG**: Journal of Images and Graphics (《中国图象图形学报》, CCF-B Chinese Journal). Acceptance×1 (first author×1).
-  **JOG**: Journal of Graphics (《图学学报》, CCF-C Chinese Journal). Acceptance×1.
-  **Neu**: Neurocomputing (CCF-C Journal, IF=5.5). Acceptance×1.
-  **CMHJ**: Chinese Mental Health Journal (《中国心理卫生杂志》, CSSCI Journal, Top Psychological Journal in China). Acceptance×1.
-  **APS**: Acta Psychologica Sinica (《心理学报》, CSSCI Journal, Top-1 Psychological Journal in China). Under review×1.
- Conference  **NeurIPS**: Conference on Neural Information Processing Systems (CCF-A Conference). Acceptance×3 (first author×1).
-  **ICLR**: International Conference on Learning Representations (CAAI-A Conference). Under review×3 (first author×2).
-  **AAAI**: Annual AAAI Conference on Artificial Intelligence (CCF-A Conference). Under review×1.
-  **NeurIPS-W**: Workshop in Conference on Neural Information Processing Systems (CCF-A Conference Workshop). Under review×1.
-  **CVPRW**: Workshop in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CCF-A Conference Workshop). Acceptance×1 (oral & best paper honorable mention×1).
-  **ICASSP**: IEEE International Conference on Acoustics, Speech, and Signal Processing (CCF-B Conference). Acceptance×1, under review×1.
-  **PRCV**: Chinese Conference on Pattern Recognition and Computer Vision (CCF-C Conference). Acceptance×2.

## Paper Summary (continued)

- CSAI: International Conference on Computer Science and Artificial Intelligence (EI Conference).  
*Acceptance*  $\times 1$  (*oral*  $\times 1$ ).

## Skills

- Languages
  - Mandarin Chinese (native speaker) and English.
- Coding
  - Python, Java, Matlab, C,  $\LaTeX$ .
- Development
  - Android, Flask, SQLite.
- Linux
  - Shell, OS virtualization.
- Misc.
  - Academic research, leadership, presentation.

## Awards and Honors

- 2024
  - Best Paper Honorable Mention**, the 3rd Workshop on Vision Datasets Understanding and DataCV Challenge in CVPR 2024.
  - Beijing Outstanding Graduates**, Beijing Municipal Education Commission (Top 5%).
- 2023
  - China National Scholarship**, Ministry of Education of the People's Republic of China (Top 1%).
  - First Prize of Climbing Scholarship**, Institute of Automation, Chinese Academy of Sciences.
- 2022
  - Merit Student**, University of Chinese Academy of Sciences.
- 2017
  - Academic Scholarship**, Beijing Institute of Technology.
  - Excellent Innovative Student**, Beijing Institute of Technology.
- 2016
  - College Scholarship**, Chinese Academy of Sciences.
  - Academic Scholarship**, Beijing Institute of Technology.
  - Excellent League Member on Youth Day Competition**, Beijing Institute of Technology.
- 2015
  - National First Prize**, Contemporary Undergraduate Mathematical Contest in Modeling (Top 1%).
  - Academic Scholarship**, Beijing Institute of Technology.
  - First Prize of Mathematics Modeling Competition**, Beijing Institute of Technology.
  - Outstanding Individual on Summer Social Practice**, Beijing Institute of Technology.
  - Second Prize on Summer Social Practice**, Beijing Institute of Technology (Team Leader).
  - Outstanding Student Cadre**, Beijing Institute of Technology.
  - Outstanding League Cadre on Youth Day Competition**, Beijing Institute of Technology.
  - Outstanding Youth League Branch**, Beijing Institute of Technology (Team Leader).
  - Top 10 Activities on Youth Day Competition**, Beijing Institute of Technology (Team Leader).
- 2014
  - Academic Scholarship**, Beijing Institute of Technology.
  - Outstanding Student**, Beijing Institute of Technology.
- 2013
  - Academic Scholarship**, Beijing Institute of Technology.

## Academic Activities and Services

- Tutorial
  - 31th IEEE International Conference on Image Processing (ICIP)**
    - Title:** An Evaluation Perspective in Visual Object Tracking: from Task Design to Benchmark Construction and Algorithm Analysis
    - Date & Location:** 27-30 October, 2024, Abu Dhabi, United Arab Emirates
    - Duration:** Half-day (Three Hours)



## Academic Activities and Services (continued)

- **27th International Conference on Pattern Recognition (ICPR)**
  - **Title:** Visual Turing Test in Visual Object Tracking: A New Vision Intelligence Evaluation Technique based on Human-Machine Comparison
  - **Date & Location:** 01-05 December, 2024, Kolkata, India
  - **Duration:** Half-day (Three Hours)
- **17th Asian Conference on Computer Vision (ACCV)**
  - **Title:** From Machine-Machine Comparison to Human-Machine Comparison: Adapting Visual Turing Test in Visual Object Tracking
  - **Date & Location:** 08-12 December, 2024, Hanoi, Vietnam
  - **Duration:** Half-day (Three Hours)

Associate Editor ■ **Journal:** Innovation and Emerging Technologies

Reviewer ■ **Conference:** NeurIPS, ICLR, CVPR, ECCV, AAAI, ACMMM, AISTATS, etc.

- **Journal:** SCIENCE CHINA Information Sciences, IEEE Access, Journal of Computational Science, Journal of Electronic Imaging, Digital Signal Processing, etc.

## Assisted Student Supervision

- Ph.D. Student ■ **Meiqi Wu**, 2022.08-Now, University of Chinese Academy of Sciences (Computer Vision & Human-computer Interaction)
- **Xiaokun Feng**, 2023.04-Now, Institute of Automation, Chinese Academy of Sciences (Visual Object Tracking & Visual Language Tracking)
- **Yiping Ma**, 2023.08-Now, East China Normal University (Intelligent Education Technique & Human-computer Interaction)
- **Dailing Zhang**, 2023.08-Now, Institute of Automation, Chinese Academy of Sciences (Visual Object Tracking & Visual Turing Test & AI Agent)
- **Yipei Wang**, 2024.08-Now, Southeast University (Multimodal Large Lanugage Model & Visual Object Tracking)
- **Xuchen Li**, 2024.08-Now, Institute of Automation, Chinese Academy of Sciences (Visual Language Tracking & Multimodal Large Lanugage Model & AI4Science)
- M.S. Student ■ **Yiping Ma**, 2022.05-2023.07, Nanjing Normal University (Intelligent Education Technique & Speech Emotion Recognition)
- **Yipei Wang**, 2022.08-2024.07, Southeast University (Visual Object Tracking & LLM for Recommendation System)
- B.E. Student ■ **Junyou Zhu**, 2022.09-2023.08, University of Chinese Academy of Sciences (Visual Turing Test)
- **Lihang Hu**, 2022.09-2023.08, University of Chinese Academy of Sciences (Visual Object Tracking)
- **Dailing Zhang**, 2022.09-2023.08, Southeast University (Visual Object Tracking)
- **Xuchen Li**, 2023.04-2024.07, Beijing University of Posts and Telecommunications (Visual Object Tracking & Visual Language Tracking)

## References

Professors Kaiqi Huang and Xin Zhao served as my Ph.D. supervisor and co-supervisor, respectively, with whom I collaborated on research in computer vision. Additionally, Prof. Choli Wang oversaw my M.Sc. studies at HKU, and I had the privilege of working with him on high-performance computing projects. Currently, I am lucky to work with Prof. Kanghao Cheong at NTU.



**Prof. Kaiqi Huang**

Professor, IAPR Fellow, IEEE Senior Member, 10,000 Talents Program - Leading Talents

Director of Center for Research on Intelligent Systems and Engineering (CRISE)

National Laboratory of Pattern Recognition (NLPR)

Institute of Automation, Chinese Academy of Sciences (CASIA)

95 Zhongguancun East Road, Beijing, China

✉ kqhuang@nlpr.ia.ac.cn

**Prof. Choli Wang**

Honorary Professor

Department of Computer Science, University of Hong Kong (HKU)

Pokfulam, Hong Kong, China

✉ clwang@cs.hku.hk

✉ choliwang@gmail.com

**Prof. Xin Zhao**

Professor, IEEE Senior Member, Beijing Science Fund for Distinguished Young Scholars

School of Computer and Communication Engineering, University of Science and Technology Beijing (USTB)

30 Xueyuan Road, Beijing, China

✉ xinzhao@ustb.edu.cn

✉ xzhaopersonal@foxmail.com

**Prof. Kanghao Cheong**

Associate Professor, IEEE Senior Member

Assistant Dean, School of Physical & Mathematical Sciences

Assistant Dean, College of Science

Nanyang Technological University (NTU)

50 Nanyang Avenue, Singapore

✉ kanghao.cheong@ntu.edu.sg