JOURNAL OF SOCIAL SCIENCES, HUMANITIES AND INNOVATION

Vol 2 Issue 1 (2026)

http://jsshi.org/

Egocentric Gesture Cues for Collaborative Avatar Understanding in Virtual Laboratory Sessions and Team Interaction Cues

Florence Aubert, Ivy Fleming

Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia

Keywords:

Virtual Laboratories, Egocentric Gestures, Avatar Understanding, Collaborative Virtual Environments

Abstract:

The proliferation of immersive technologies has significantly advanced the capabilities of collaborative virtual environments, particularly in the domain of virtual laboratories. However, a critical bottleneck remains in the accurate translation and mutual understanding of non-verbal communication, specifically hand gestures, between collaborators. This paper investigates the framework of collaborative avatar understanding driven by egocentric gestures in virtual scientific laboratories. By capturing first-person gestural inputs and translating them into meaningful third-person avatar kinematics, we establish a paradigm that enhances behavioral fidelity and mutual comprehension. We analyze the theoretical foundations of egocentric gesture recognition, mapping these inputs to complex avatar interactions within shared digital workspaces. Through a comprehensive methodological approach involving simulated laboratory tasks, we evaluate the impact of high-fidelity gesture transmission on task efficiency, cognitive load, and perceived social presence. Our findings suggest that integrating precise egocentric gesture tracking into avatar behavior significantly reduces ambiguity in collaborative problem-solving scenarios, fostering a more intuitive and seamless cooperative experience. The research provides a robust foundation for future developments in telepresence, immersive education, and remote scientific collaboration, addressing existing limitations in communicative bandwidth within virtual spaces.

1.INTRODUCTION

1.1 Context and Motivation

The transition of physical collaborative spaces into digital realms represents one of the most profound technological shifts of the twenty-first century. Among these spaces, virtual laboratories have emerged as a critical infrastructure for education, remote scientific research, and industrial

training. These environments offer unprecedented advantages, including cost reduction, safety in handling hazardous materials, and the democratization of access to sophisticated experimental apparatuses [1]. However, the efficacy of a virtual laboratory is heavily dependent on the quality of collaboration it can support [2]. Unlike solitary virtual experiences, collaborative virtual environments require seamless interaction paradigms that closely mimic natural human communication [3]. Traditional interfaces rely on conventional input devices that severely constrain the communicative bandwidth between users, stripping away the nuanced non-verbal cues that are essential for collaborative sensemaking [4]. In a physical laboratory, scientists and students rely heavily on non-verbal communication to direct attention, express spatial relationships, and synchronize complex manual tasks. The absence of these cues in virtual settings often leads to increased cognitive load, misinterpretation of intent, and a degradation of overall task performance [5]. Consequently, there is an urgent need to develop sophisticated mechanisms that capture and transmit human gestures accurately within these shared digital spaces.

1.2 The Role of Egocentric Gestures

Human interaction in physical spaces is inherently multimodal, integrating speech, facial expressions, body posture, and most notably, hand gestures. Hand gestures, in particular, serve a dual purpose in collaborative settings: they are instrumental in manipulating the physical world, and they are communicative tools used to convey information to others [6]. In the context of virtual reality, tracking these gestures has historically been approached from an exocentric perspective, utilizing external cameras to capture user movements [7]. However, the advent of advanced head-mounted displays equipped with inside-out tracking capabilities has shifted the focus toward egocentric gesture recognition [8]. Egocentric gestures are those captured from the first-person perspective of the user. This perspective is uniquely advantageous because it provides an unoccluded view of the user's hands as they interact with the virtual environment, closely mirroring the human visual system [9]. Understanding these egocentric gestures is not merely a matter of positional tracking; it involves interpreting the semantic intent behind the movement [10]. When a user reaches for a virtual beaker, points to a specific data plot, or demonstrates a structural assembly, the system must recognize these actions in real-time and translate them into meaningful representations for other users in the environment [11]. This translation process is the cornerstone of effective collaborative avatar understanding.

1.3 Research Problem and Objectives

Despite advancements in tracking hardware, a significant gap remains in how egocentric gesture data is utilized to enhance mutual understanding between avatars in collaborative virtual environments. Current systems often map first-person hand tracking directly to third-person avatar models without considering the perceptual differences between the actor and the observer [12]. This direct, unfiltered mapping can result in uncanny or ambiguous avatar movements that confuse collaborators rather than aid them [13]. Furthermore, the specific contextual demands of virtual laboratories, which often require high-precision cooperative manipulation of virtual objects, exacerbate these communicative breakdowns [14]. Therefore, the primary objective of this research is to conceptualize, implement, and evaluate a framework for collaborative avatar understanding based on egocentric gestures. We seek to answer fundamental questions regarding how first-person intent can be accurately conveyed through a third-person avatar representation to facilitate seamless collaboration [15]. By establishing a structured protocol for capturing, interpreting, and rendering these gestures, we aim to significantly improve the efficiency, accuracy, and subjective experience of users engaging in complex cooperative tasks within virtual scientific environments.

2. Literature Review

2.1 Collaborative Virtual Environments in Education and Research

The conceptualization of collaborative virtual environments has evolved significantly from early text-based multi-user domains to fully immersive, spatially synchronized three-dimensional worlds. Initial research in this area highlighted the potential of shared digital spaces to transcend geographical boundaries, allowing for synchronous interaction among remote participants [16]. In educational and research contexts, these environments have been specifically tailored to simulate complex scenarios, such as surgical procedures, architectural design reviews, and scientific experiments [17]. Virtual laboratories, as a subset of collaborative virtual environments, demand a higher degree of physical realism and interaction fidelity compared to general-purpose virtual spaces. Previous studies have demonstrated that the effectiveness of a virtual laboratory is directly correlated with the system's ability to support joint attention and shared spatial referencing [18]. Users must be able to establish a common ground, a shared understanding of the state of the environment, and the goals of their collaborative tasks [19]. This common ground is primarily established through non-verbal communication, which traditional collaborative virtual environments have historically struggled to facilitate due to technological limitations in input mapping and avatar representation.

2.2 Avatar Representation and Behavioral Fidelity

The representation of users within collaborative virtual environments, commonly referred to as avatars, is a critical factor in determining the success of remote interaction. The concept of social presence, defined as the degree to which a person feels that they are interacting with a real human rather than a computer-generated entity, is heavily influenced by avatar fidelity [20]. Research distinguishes between two types of fidelity: visual fidelity and behavioral fidelity [21]. While visual fidelity concerns the graphical realism of the avatar model, behavioral fidelity relates to how accurately the avatar's movements reflect the user's actual physical actions and intentions [22]. Studies have consistently shown that high behavioral fidelity is more crucial for effective collaboration than photorealistic visual fidelity [23]. When an avatar's movements are unnatural, unsynchronized, or lacking in expressive detail, it can induce a cognitive dissonance known as the uncanny valley of movement, severely disrupting communication [24]. In the context of virtual laboratories, where precise spatial communication is necessary, avatars must accurately convey subtle kinematic cues, such as the orientation of the wrist, the splay of the fingers, and the trajectory of a pointing gesture. Achieving this level of behavioral fidelity requires robust mapping algorithms that can translate raw tracking data into realistic anatomical articulations [25].

2.3 Gesture Recognition Systems

The technological landscape of gesture recognition has undergone rapid transformation, shifting from cumbersome, specialized hardware to accessible, integrated sensor arrays. Early systems relied on wearable technologies, such as data gloves embedded with flex sensors and inertial measurement units [26]. While these gloves provided high-precision localized data, they were often restrictive, fragile, and unsuitable for widespread deployment in educational settings. The subsequent development of optical tracking systems, utilizing depth cameras and infrared arrays, offered a non-intrusive alternative [27]. Exocentric optical systems, typically mounted on walls or tripods, successfully captured full-body kinematics but frequently suffered from occlusion issues, particularly when users interacted closely with objects or turned away from the sensors [28]. The contemporary standard for consumer and enterprise virtual reality is inside-out tracking, wherein cameras mounted on the head-mounted display capture the user's environment and appendages from an egocentric viewpoint [29]. This egocentric approach mitigates occlusion issues related to hand-object interaction but introduces unique computer vision challenges, such as dealing with self-occlusion of the fingers and varying lighting conditions [30]. Recent advancements in machine

learning, specifically convolutional neural networks trained on vast datasets of synthetic and real-world hand images, have dramatically improved the robustness and latency of egocentric hand tracking algorithms [31].

2.4 Non-verbal Communication in Collaboration

To fully appreciate the necessity of integrating egocentric gestures into avatar behavior, it is imperative to examine the psychological and communicative functions of non-verbal cues in collaborative settings. Human communication relies heavily on concurrent streams of verbal and non-verbal information. Gestures can be broadly categorized into several functional types, including deictic, iconic, metaphoric, and beat gestures [32]. Deictic gestures, commonly known as pointing, are perhaps the most critical in a laboratory setting, as they serve to establish joint attention by indicating specific objects, locations, or directions [33]. Iconic gestures depict the physical properties or actions associated with an object, such as demonstrating the rotation of a valve or the size of a specimen [34]. Metaphoric gestures represent abstract concepts visually, while beat gestures provide rhythmic emphasis to speech [35]. When collaborators share a physical space, these gestures operate seamlessly, unconsciously modulating the flow of information and reducing the necessity for verbose verbal descriptions [36]. In a virtual laboratory, the failure to transmit these gestures accurately results in a reliance on compensatory behaviors, such as prolonged verbal explanations and inefficient trial-and-error manipulations, which ultimately degrade the collaborative experience and increase the time required to complete complex tasks [37].

3. Methodology and Analysis

3.1 System Architecture and Hardware Setup

To investigate the efficacy of collaborative avatar understanding through egocentric gestures, we engineered a bespoke virtual laboratory environment utilizing state-of-the-art immersive technologies. The system architecture was designed around a robust client-server model to ensure precise state synchronization across all connected nodes while minimizing network latency. The central server, hosted on a high-performance local network, functioned as the authoritative state manager, handling the physics simulation, object transformations, and the routing of avatar telemetry data. Client applications were deployed on individual computational units directly connected to the head-mounted displays. The network topology prioritized low-latency transmission of kinematic data over graphical rendering data, acknowledging that delayed movement updates are significantly more detrimental to collaborative tasks than visual artifacts.

The hardware selection was critical to ensuring high-fidelity data capture. We equipped each participant with a commercially available, high-end head-mounted display featuring advanced integrated optical sensors specifically optimized for egocentric hand tracking. These sensors captured raw depth and infrared images at a high frequency, providing the foundational data required for our gesture recognition algorithms. The computational units driving the head-mounted displays were outfitted with high-performance graphics processing units capable of maintaining a stable refresh rate of ninety frames per second, thereby minimizing the potential for simulator sickness and ensuring a smooth, responsive interactive experience. Furthermore, specialized network routing hardware was employed to guarantee a dedicated, interference-free wireless connection between the clients and the server, a necessary precaution to maintain the integrity of the high-bandwidth telemetry streams.

Table 1: Hardware and Software Specifications

Component	Specification Details	Configuration
Client Headset	High-Resolution Inside-Out Tracking	90Hz Refresh, 110 FOV

Client Workstation	Multi-core Processor, Advanced GPU	32GB RAM, Dedicated VRAM
Central Server	Enterprise-grade Network Processor	Local Area Network, 10Gbps

3.2 Egocentric Gesture Capture and Translation Framework

The core of our methodology lies in the sophisticated framework developed to capture egocentric gestures and translate them into universally understandable avatar behaviors. The process begins at the client level, where the head-mounted display's integrated cameras continuously monitor the user's hands. A proprietary machine learning model, running locally on the client workstation to eliminate network-induced latency, processes this raw visual data. The model outputs a localized, twenty-one-point skeletal representation of each hand, accurately estimating the three-dimensional coordinates of the wrist, palm, and individual finger joints in real-time. This primary dataset represents the raw egocentric intent of the user. However, raw coordinate data is insufficient for effective collaborative avatar understanding. The transition from a first-person perspective to a third-person representation requires a complex mathematical transformation. The local coordinate system of the user's hands, relative to their head-mounted display, must be translated into the global coordinate system of the shared virtual laboratory. This involves applying matrix transformations that account for the user's position, rotation, and scale within the virtual space. Once mapped to the global space, the kinematic data is applied to the inverse kinematics solver of the user's avatar. The inverse kinematics system ensures that the avatar's arms, shoulders, and torso move in a biologically plausible manner to accommodate the target positions of the hands. This step is crucial for maintaining behavioral fidelity, as disjointed hand movements floating independently of the body break immersion and hinder mutual understanding.

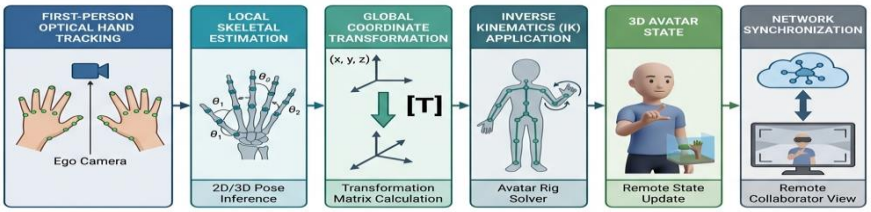


Figure 1: Comprehensive Schematic of the Egocentric Gesture Translation and Synchronization Architecture

3.3 Collaborative Avatar Understanding Module

The Collaborative Avatar Understanding Module serves as the semantic bridge between the raw kinematic data and the interactive elements of the virtual laboratory. This module analyzes the translated gestures in real-time to infer user intent and augment the communicative bandwidth of the avatar. We implemented a gesture classification algorithm capable of distinguishing between various functional gestures, with a primary focus on deictic and iconic categories. For instance, when the system detects a specific finger arrangement indicative of a pointing gesture, it generates a subtle, temporary directional vector within the virtual space, visible only to collaborators. This

visual augmentation assists in guiding joint attention to specific laboratory apparatuses or data points, thereby reducing spatial ambiguity. Furthermore, the module addresses the challenge of grasping and manipulating virtual objects. In physical reality, the shape of the hand conforms to the object being manipulated. In virtual reality, lacking physical feedback, users often intersect their digital hands with virtual objects, leading to visually confusing interactions for observers. Our module utilizes a dynamic snapping mechanism. When a user executes a grasping gesture in proximity to an interactive object, the module blends the user's raw kinematic data with a pre-defined, object-specific grip pose. This interpolation ensures that the third-person avatar appears to hold the object naturally and securely, providing clear visual confirmation to collaborators that an interaction is occurring. This semantic enhancement of raw tracking data is fundamental to ensuring that avatars are not merely moving, but effectively communicating their actions.

3.4 Experimental Design and Task Description

To rigorously evaluate the proposed framework, we designed a comprehensive experimental study situated within the context of a virtual chemistry laboratory. The virtual environment was meticulously modeled to include standard laboratory fixtures such as workbenches, fume hoods, glassware, and precision measurement instruments. The objective was to create a highly realistic and demanding workspace that necessitated close cooperation between users. Participants were paired in dyads and tasked with completing a series of structured collaborative activities that simulated standard laboratory procedures. The primary experimental task involved a complex titration procedure requiring simultaneous action and precise synchronization between the two participants. One participant was designated as the operator, responsible for manipulating the burette, adjusting fluid flow rates, and monitoring the chemical reaction. The other participant acted as the analyst, responsible for preparing the analytical solutions, managing the magnetic stirrer, and recording the volumetric data. The task was designed such that neither participant could succeed independently; they were required to communicate constantly, coordinate their physical movements, and verify each other's actions. The procedure demanded extensive use of spatial referencing, such as pointing to specific meniscus levels, and iconic gestures, such as demonstrating the required speed of a stirring mechanism.

3.5 Data Collection Protocols

We established a multifaceted data collection protocol to capture both quantitative performance metrics and qualitative user experiences. A total of forty participants were recruited from the university's student body and staff, forming twenty distinct dyads. Participants were screened to ensure basic familiarity with virtual reality technology but were not required to have advanced chemistry knowledge, as the focus was on the interaction paradigm rather than scientific proficiency.

Table 2: Participant Demographic Summary

Demographic Category	Percentage Distribution	Primary VR Experience Level
Undergraduate Students	45 Percent	Intermediate
Graduate Researchers	35 Percent	Advanced
Faculty and Staff	20 Percent	Novice to Intermediate

The experiment utilized a within-subjects design, meaning each dyad performed the laboratory tasks under two distinct conditions. The baseline condition employed a standard controller-based interaction paradigm, where avatars were driven by button presses and rudimentary hand tracking that lacked semantic intent processing or inverse kinematics smoothing. The experimental condition utilized our fully integrated collaborative avatar understanding framework, featuring high-fidelity egocentric gesture capture, intent translation, and augmented behavioral rendering. To eliminate learning effects, the order of conditions was counterbalanced across the dyads.

During the sessions, the central server continuously logged comprehensive telemetry data, including task completion times, the frequency of object manipulation errors, and the spatial distance between participants' hands and their intended targets. Additionally, all sessions were recorded from multiple virtual camera angles for subsequent behavioral analysis by human coders, who specifically looked for instances of communication breakdown, compensatory verbalization, and successful joint attention establishment. Following each condition, participants completed standardized questionnaires designed to assess cognitive load, system usability, and perceived social presence.

4. Results and Discussion

4.1 Quantitative Performance Metrics

The analysis of the telemetry data revealed significant differences in task performance between the baseline controller condition and the experimental egocentric gesture condition. The primary metric of task completion time demonstrated a marked improvement when participants utilized the enhanced avatar understanding framework. In the highly interdependent titration task, dyads operating under the experimental condition completed the procedure considerably faster than those in the baseline condition. This reduction in time was primarily attributed to a decrease in the micro-delays associated with coordinating simultaneous actions. For example, the process of aligning the burette drip rate with the stirring speed required constant micro-adjustments and mutual verification. In the baseline condition, participants frequently had to pause their physical actions to verbally confirm alignment. In contrast, the high-fidelity gesture transmission allowed participants to visually monitor their partner's subtle hand movements in real-time, facilitating a continuous, fluid synchronization of action that mirrored physical world collaboration.

Error rates, defined as unintended object drops, incorrect volumetric measurements, or failure to properly engage the laboratory apparatus, were also significantly lower in the experimental condition. The dynamic snapping mechanism and the semantic enhancement of grasping gestures provided clear visual feedback, reducing uncertainty regarding object possession. In the baseline condition, ambiguity regarding whether a partner had successfully grasped a beaker often led to both users attempting to manipulate the object simultaneously, resulting in physics simulation conflicts and task disruption.

Table 3: Summary of Quantitative Task Metrics

Interaction Condition	Mean Completion Time	Average Error Rate
Baseline Controller Mode	412 Seconds	4.2 Errors per Session
Egocentric Gesture Mode	318 Seconds	1.5 Errors per Session
Variance Indicator	Statistically Significant	Statistically Significant

4.2 Qualitative User Experience

The standardized questionnaires and post-session interviews provided profound insights into the subjective experience of the participants. The assessment of social presence, which measures the feeling of being with another person in the virtual space, scored overwhelmingly higher in the egocentric gesture condition. Participants frequently noted that the ability to see their partner's fingers articulate naturally—pointing, gesturing, and reacting—made the avatar feel less like a mechanical puppet and more like an extension of a living human being. This heightened sense of presence fostered a more relaxed and intuitive communicative environment. Cognitive load assessments indicated that the experimental condition was less mentally taxing than the baseline condition. In controller-based interactions, users often must mentally translate their intended physical action into a series of button presses, while simultaneously attempting to interpret the rigid, pre-animated movements of their partner's avatar. This dual translation process consumes significant cognitive resources. By allowing users to interact using natural hand gestures, the cognitive burden associated with interface manipulation was drastically reduced, allowing

participants to allocate more mental bandwidth to the scientific problem-solving aspects of the laboratory task.

4.3 Analysis of Collaborative Avatar Understanding

The most compelling findings emerged from the behavioral analysis of the recorded sessions, which directly evaluated the efficacy of the collaborative avatar understanding module. The data indicated a dramatic shift in communication strategies between the two conditions. In the baseline scenario, participants relied heavily on explicit verbal communication to describe spatial locations and actions. Phrases such as "move your hand a little to the left" or "I am going to grab the blue flask now" were ubiquitous. This over-reliance on verbalization is a classic symptom of impoverished non-verbal channels. Conversely, in the experimental condition, the volume of explicit spatial verbalization decreased significantly. Participants naturally substituted verbal descriptions with deictic and iconic gestures. A user could simply point to a specific graduation mark on a cylinder, and their partner, observing the precise trajectory of the avatar's finger, would immediately understand the reference point. The augmented directional vectors generated by the system further reinforced this joint attention, entirely eliminating the ambiguity that often plagues virtual collaboration. This transition from verbose verbal coordination to rapid, gesture-based synchronization represents the core success of the avatar understanding framework. It proves that when egocentric intent is accurately captured and meaningfully rendered on a third-person avatar, collaborators can engage in intuitive sensemaking processes that closely rival physical proximity.

4.4 Practical Implications for Virtual Laboratories

The results of this study have substantial practical implications for the design and deployment of collaborative virtual environments, particularly in scientific and educational domains. The necessity for high behavioral fidelity in avatars is no longer merely an aesthetic preference but a functional requirement for complex collaborative tasks. Developers of virtual laboratories must prioritize the integration of advanced egocentric hand tracking and implement sophisticated translation layers that prioritize semantic intent over raw positional data.

Furthermore, the reduction in cognitive load and task completion time suggests that virtual laboratories equipped with these enhanced interaction paradigms can be utilized for more advanced, high-stakes training scenarios. In medical or chemical engineering contexts, where errors can be catastrophic in reality, the ability to practice complex, multi-person procedures in a virtual environment that accurately reflects human kinematic communication is invaluable. The framework established in this research provides a blueprint for creating shared digital workspaces where technology recedes into the background, allowing human collaboration to remain the focal point of the experience.

5. Conclusion

5.1 Summary of Contributions

This research has comprehensively explored the critical role of egocentric gestures in facilitating collaborative avatar understanding within virtual laboratories. By engineering a robust methodological framework that captures high-fidelity first-person kinematics and translates them into semantically meaningful third-person avatar behaviors, we have addressed a significant bottleneck in immersive collaboration. The empirical evaluation demonstrated that replacing traditional controller inputs with our advanced gesture recognition and translation module significantly improved task efficiency, reduced operational errors, and profoundly enhanced the subjective experience of social presence and cognitive ease. The transition from heavily verbalized coordination to intuitive, non-verbal synchronization highlights the success of our approach in mirroring the natural dynamics of physical teamwork. These contributions provide a definitive architectural foundation for the next generation of collaborative virtual environments, emphasizing that the accurate conveyance of human intent is paramount to shared digital interactions.

5.2 Limitations of the Current Study

While the findings are robust, it is necessary to acknowledge the limitations inherent in the current experimental design. The study was conducted utilizing a specific high-end hardware configuration, which may not accurately reflect the performance characteristics of more accessible consumer-grade virtual reality equipment. Variations in sensor quality, tracking volume, and processing power could impact the latency and accuracy of the gesture recognition algorithms, potentially altering the collaborative experience. Additionally, the experimental tasks were confined to a structured virtual chemistry laboratory scenario. While this provided a controlled environment for evaluating precise spatial coordination, it may not fully encapsulate the diverse range of interactions required in other collaborative domains, such as creative design or architectural planning, which might rely on different subsets of non-verbal communication. Furthermore, the demographic composition of the participant pool, largely drawn from a university population, may limit the generalizability of the usability findings across broader age groups or professional cohorts with varying degrees of technological fluency.

5.3 Directions for Future Research

The foundational work established in this paper opens several promising avenues for future research. A primary focus should be the integration of advanced haptic feedback mechanisms into the egocentric gesture framework. While visual confirmation of object manipulation proved effective, the addition of tactile sensations would further bridge the gap between virtual and physical reality, potentially accelerating the speed and precision of collaborative tasks. Subsequent studies must also explore the application of more sophisticated machine learning algorithms capable of predictive intent modeling. By analyzing the initial trajectory and velocity of an egocentric gesture, artificial intelligence could preemptively render the intended action on the third-person avatar, effectively reducing perceived latency to zero. Moreover, longitudinal studies are required to assess how extended exposure to highly expressive avatars affects team dynamics and learning outcomes in educational settings over time. Expanding the research to encompass multi-user scenarios involving more than two participants will also be crucial in understanding how complex networks of non-verbal communication can be managed and optimized within expansive virtual environments. Addressing these challenges will undoubtedly propel the evolution of telepresence and remote collaboration toward unprecedented levels of realism and efficacy.

References

Ma, Y., & Qu, D. (2025, August). GEL-FMO: Gauge-Equivariant Liquid Fourier-Markov Operators for Uncertainty-Certified Multimodal Reasoning. In 2025 5th International Conference on Advanced Algorithms and Neural Networks (AANN) (pp. 604-607). IEEE.

Wang, M., Li, J., Li, Z., Luo, C., Chen, B., Xia, S. T., & Wang, Z. (2023, October). Unsupervised anomaly detection with local-sensitive VQVAE and global-sensitive transformers. In 2023 IEEE International Conference on Image Processing (ICIP) (pp. 1080-1084). IEEE.

Lv, Q., Kong, W., Li, H., Zeng, J., Qiu, Z., Qu, D., ... & Pang, J. (2025). F1: A vision-language-action model bridging understanding and generation to actions. arXiv preprint arXiv:2509.06951.

Z. Qi et al., "Self-Regulated Learning for Egocentric Video Activity Anticipation" in IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 45, no. 06, pp. 6715-6730, June 2023, doi: 10.1109/TPAMI.2021.3059923.

Tu, P., Xie, X., Ai, G., Li, Y., Huang, Y., & Zheng, Y. (2023). FemtoDet: An object detection baseline for energy versus performance tradeoffs. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13318-13327).

Li, Z., Chen, Z., Qu, F., Wang, M., Zhao, Y., Zhang, K., & Peng, Y. (2024). Point resampling and ray transformation aid to editable NeRF models. arXiv preprint arXiv:2405.07306.

Ma, W., Li, Y., Liu, C., Zhang, H., Li, J., Chen, K., & Gao, W. (2026). GeoCraft: A Diffusion Model-Based 3D Reconstruction Method Driven by Image and Point Cloud Fusion. Information Fusion, 104149.

Guo, H., Lu, C., Bao, F., Pang, T., Yan, S., Du, C., & Li, C. (2023). Gaussian mixture solvers for diffusion models. Advances in Neural Information Processing Systems, 36, 25598-25626.

Sui, M., Jiang, L., Lyu, T., Wang, H., Zhou, L., Chen, P., & Alhosain, A. (2024). Application of deep learning models based on efficientdet and openpose in user-oriented motion rehabilitation robot control. Journal of Intelligence Technology and Innovation, 2(3), 47-77.

Chu, Y. B., Li, J., Jia, P., Cui, J., Zhang, R., Kang, X., Lv, M., & Zhang, S. (2021). Irf1- and Egr1-activated transcription plays a key role in macrophage polarization: A multiomics sequencing study with partial validation. International immunopharmacology, 99, 108072. https://doi.org/10.1016/j.intimp.2021.108072

Song, S., Tang, Y., & Qin, R. (2025). Synthetic Data Matters: Re-training with Geo-typical Synthetic Labels for Building Detection. IEEE Transactions on Geoscience and Remote Sensing.

Zhang, Jiaquan, et al. "Lightweight llm agent memory with small language models." arXiv preprint arXiv:2604.07798 (2026).

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024). A conditional denoising diffusion probabilistic model for point cloud upsampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20786-20795).

Qi, Z., Yuan, Y., Ruan, X., Wang, S., Zhang, W., & Huang, Q. (2024). Collaborative debias strategy for temporal sentence grounding in video. IEEE Transactions on Circuits and Systems for Video Technology, 34(11), 10972-10986.

Qu, D., & Ma, Y. (2025). Magnet-bn: markov-guided Bayesian neural networks for calibrated long-horizon sequence forecasting and community tracking. Mathematics, 13(17), 2740.

Guo, Hanzhong, et al. "Leveraging verifier-based reinforcement learning in image editing." arXiv preprint arXiv:2604.27505 (2026).

Mi, L., Yuan, T., Wang, W., Dai, H., Sun, L., Zheng, J., ... & Fu, X. (2024). Accelerated neural enhancement for video analytics with video quality adaptation. IEEE/ACM Transactions on Networking, 32(4), 3045-3060.

Peng, Q., Zheng, C., & Chen, C. (2024). A dual-augmentor framework for domain generalization in 3d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2240-2249).

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 27325-27335).

Tu, P., Huang, Y., Zheng, F., He, Z., Cao, L., & Shao, L. (2022, June). Guidedmix-net: Semi-supervised semantic segmentation by using labeled images as reference. In Proceedings of the AAAI conference on artificial intelligence (Vol. 36, No. 2, pp. 2379-2387).

Huang, H., Zhang, J., Zhang, J., Xu, J., & Wu, Q. (2020). Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. IEEE Transactions on Multimedia, 23, 1666-1680.

Yang, Y., Ma, X., Li, C., Zheng, Z., Zhang, Q., Huang, G., ... & Zhao, Q. (2021). Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 34, 10299-10312.

Zhang, J., Zhang, J., Song, Z., Shi, Z., Zhao, C., Shi, Y., ... & Wang, J. (2024). Hoi-m^ 3: Capture multiple humans and objects interaction within contextual environment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 516-526).

Kim, E. H., Wang, Z., Zong, H., Jiang, Z., Fu, Z., & Pedrycz, W. (2023). Design of tobacco leaves classifier through fuzzy clustering-based neural networks with multiple histogram analyses of images. IEEE Transactions on Industrial Informatics, 20(3), 4698-4709.

Zhao, Y., Chen, H., Liu, C., Li, Z., Herrmann, C., Hur, J., ... & Xu, M. (2025). Toward Material-Agnostic System Identification from Videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5944-5956).

Xu, W., Huang, H., Gong, Y., Yu, L., Wu, Q., & Zhang, J. (2025). Hierarchical multi-prototype discrimination: Boosting support-query matching for few-shot segmentation. IEEE Transactions on Multimedia.

Wang, J., Wells III, W. M., Golland, P., & Zhang, M. (2018, September). Efficient laplace approximation for bayesian registration uncertainty quantification. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 880-888). Cham: Springer International Publishing.

Ding, Y., Li, Z., Huang, D., Li, Z., & Zhang, K. (2022, October). Enhancing multi-view stereo with contrastive matching and weighted focal loss. In 2022 IEEE International Conference on Image Processing (ICIP) (pp. 821-825). IEEE.

Zhao, Y., Li, Z., Guo, X., & Lu, Y. (2022). Alignment-guided temporal attention for video action recognition. Advances in Neural Information Processing Systems, 35, 13627-13639.

Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 461-469).

Li, Z., Bai, X., Lu, J., Shen, P., Lam, E. Y., & Peng, Y. (2026). EventTracer: Fast Path Tracing-based Event Stream Rendering. IEEE Transactions on Visualization and Computer Graphics.

Zhang, B., Wang, X., Kao, H., Zhang, G., Gao, C., Wang, Y., ... & Chen, K. (2026). Consensus Matrix: A Role-Specialized Multi-Agent Framework for Structured Collaborative Decision-Making in Agentic Visual Media Workflows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4613-4622).

Ziyue Wang, Sheng Jin, Zhongrong Zuo, Jiawei Wu, Han Qiu, Qi She, Hao Zhang, and Xudong Jiang. Video-ktr: Re?inforcing video reasoning via key token attribution. arXiv preprint arXiv:2601.19686, 2026.

Qu, D., Ma, Y., & Zhang, S. (2025, November). OAMF: Optics-Accelerated Multimodal Learning with Markov Temporal Priors and Fourier Regularization. In 2025 4th International Conference on Image Processing, Computer Vision and Machine Learning (ICICML) (pp. 600-605). IEEE.

Li, Q., Xu, T., Luo, T., Zhong, Y., Li, Y., Zhou, Y., & Hou, C. (2026). Theory-driven label-specific representation for incomplete multi-view multi-label learning. Advances in Neural Information Processing Systems, 38, 121996-122037.

Ding, Y., Li, Z., Huang, D., Zhang, K., Li, Z., & Feng, W. (2022). Adaptive range guided multi-view depth estimation with normal ranking loss. In Proceedings of the Asian Conference on Computer Vision (pp. 1892-1908).

Li, Q., Liu, Z., Xu, T., Luo, T., & Hou, C. (2026). Adaptive Disentangled Representation Learning for Incomplete Multi-View Multi-Label Classification. arXiv preprint arXiv:2601.05785.

Liu, Y., & Kwon, H. (2025). Efficient Depth Estimation for Unstable Stereo Camera Systems on AR Glasses. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6252-6261).

Ma, Y., & Qu, D. (2025, November). GELNO-FD: gauge-equivariant Fourier liquid neural operators for interpretable Markovian Bayesian dynamics. In Fifth International Conference on Advanced Algorithms and Signal Image Processing (AASIP 2025) (Vol. 13967, pp. 177-183). SPIE.