From Artificial Intelligence to Embodied Intelligence: The Evolution from Abstract Problem-Solving to 3D Scene Understanding

Jingyuan Huo

{ Jingyuan.huo@my.jcu.edu.au }

James Cook University (Singapore), Singapore

Abstract. Classical artificial intelligence has long centered on symbolic reasoning frameworks and statistical pattern recognition, with most prior studies confined to deterministic, closed settings that decouple cognitive computation from real-world physical interaction. This paper systematically explores the pivotal paradigm shift toward embodied intelligence, a framework where cognition is fundamentally grounded in physical or simulated three-dimensional spaces via tight, continuous perception-action coupling. It generalizes the core methodological innovations across data acquisition pipelines, 3D representation learning, and embodied learning algorithms, and validates relevant empirical findings through standard benchmark tasks including robot navigation, dexterous manipulation, and object-centric interaction tasks. The paper evaluates seminal enabling technologies such as RGB-D data sensing, learning-based 3D perception, and high-fidelity simulation platforms, along with matched data augmentation approaches. It also discusses emerging fundamental conflicts between end-to-end and modular architectures, and between geometric and semantic 3D representations. Finally, it outlines frontier research trends including multimodal foundation models, differentiable physics, and sim-to-real transfer techniques, while clearly defining key unresolved bottlenecks: sample efficiency deficits, causal reasoning limitations, and cross-scenario transfer validity gaps.

Keywords: Embodied intelligence; 3D scene understanding; Sim-to-real transfer; Multimodal learning; Robotic perception.

1 Introduction

Symbolic reasoning, logical inference and statistical recognition of patterns were the primary focus of the Traditional artificial intelligence (AI). This problem-solving task has been in control and well-defined environments [1]. Success has been seen in areas such as natural language processing (NLP), playing games and data analytics for quick decision making using the same approaches. Despite this, they always lack the skills to reason, perceive and act effectively within complex and dynamic real-world environments, situations in which direct physical contact interaction and sensory earthing are essential [2].

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365341

Therefore, there is a paradigm shift in AI study with an emphasis on coupling tightly between perception, cognition, and action within direct contact, which is a physical or simulated environment represented by Embodied intelligence [3]. In this pattern, intelligent agents continuously acquire knowledge via progressive interaction with their surroundings and environment, which is dependent on multimodal sensory inputs, given examples include vision, depth, and spatial cues. In conjunction with that, it gives a fundamental capability supporting embodied intelligence, which is 3D scene understanding. Hence, making it possible for the agents to recognize objects, how objects and locations connect in the space to infer spatial relationships, and understand more about geometry and semantics in three-dimensional space [4].

3D scene understanding has been progressively accelerated by recent improvements in deep learning, robotics, computer vision and many more. Indoor and outdoor representations have been richer, which have been enabled by Learning-based models and Large-scale 3D datasets, including navigation, manipulation and human-robot interactions as supporting tasks [5][6].

2. Review logic and baseline methodology

Low degrees of perception-cognition-action interaction, absence of explicit spatial grounding and limited capacity to represent three-dimensional geometry and semantic relations in the environment characterize the traditional form of disembodied artificial intelligence [7]. These limitations have been widely observed to be internal hindrances to effective intelligence in real-life scenario particularly in those situations where the necessity to communicate, adapt and think spatially in real-time is involved.

The study of embodied intelligence has developed in response to these shortcomings and aims at the integration of perception, cognition and action through direct interaction with the physical or simulated environment. In this paradigm, the use of 3D scene understanding has an enabling role as it allows intelligent agents to perceive spatial structure, detect objects in three-dimensional space and represent semantic and geometric relationships that are necessary to execute navigation and manipulation processes [4]. Recent developments in deep learning, robotics, and huge 3D data sets have prompted a higher frequency of the development of embodied systems to behave in complicated indoor and outdoor environments [8].

3. Related previous methods

3.1. Baseline: traditional ai, early robotics and behaviour-based systems

Traditional artificial intelligence was mostly developed in a disembodied paradigm, in which intelligence was defined as abstract computation detached from any physical interaction with the environment. The initial methods focused on symbolic reasoning, decision making based on rules and pattern recognition, and statistical operations with well-defined inputs [1]. Action and perception were independent variables, making it impossible for such systems to work strongly in dynamic and unstructured situations. In early robotics, this separation was further reflected in simplified perception pipelines and pre-programmed control strategies. The poor interconnection between perception, cognition and action has been identified as a root cause of

hindrance to autonomous development, especially when it comes to performing tasks that involve continuous interaction and spatial judgment [7].

Embodied and behaviour-based robotics redefined intelligence as an emergent behaviour of an agent in its interaction with the environment. The subsumption architecture provided by Brooks showed how centralized symbolic planning might be substituted with the use of layered reactive behaviours, to allow real-time responsiveness [1]. Complementary views of embodied cognition focus on the importance of physical morphology and environmental affordances to intelligent behaviour [2]. While these approaches improved robustness, they relied heavily on handcrafted models and classical pipelines such as SLAM and motion planning, motivating the integration of learning-based perception.

3.2 3D data acquisition, representations and scene understanding

Advances in 3D data acquisition and representation enabled the transition toward embodied intelligence. Dense reconstruction methods based on multi-view geometry and RGB-D sensing provided spatial structure, while large-scale annotated datasets such as ScanNet and Matterport3D enabled supervised learning of geometry and semantics at scale [5, 9]. Such developments enabled learning-based 3D scene-understanding tasks that include semantic segmentation, object detection and scene parsing, which transform raw geometry information into structured representations and these can be used in navigation and manipulation. Object-centric 3D perception and manipulation are illustrated in Figure 1.

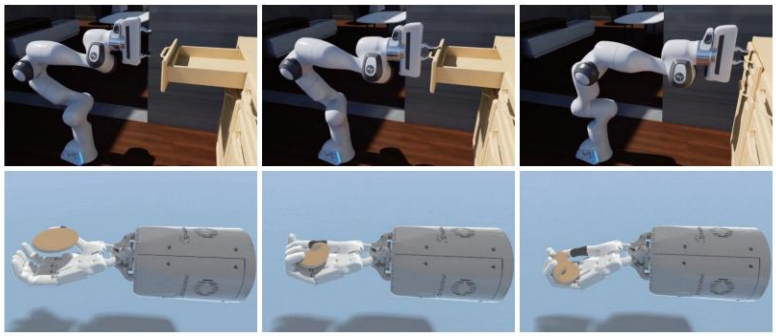

Fig. 1. Object-centric 3D perception and manipulation [3].

One of the main methodological differences is related to the choice of representation. Voxel grids, point clouds, meshes, and implicit neural representations all consist of different trade-offs between fidelity, scale, and computational cost. It has been shown that the representation choice has a direct effect on the learning structures and the downstream embodied performance [8].

3.3 Embodied learning and data-centric methods

Embodied intelligence is turning out to be more dependent on reinforcement learning and imitation learning in the endeavour to close the perception-action loop, as shown in Figure 2. Reinforcement learning allows agents to learn navigation and interaction policies right through experience in both the virtual and real worlds [6]. Imitation learning, on the other hand, improves the efficiency and stability of the samples in manipulation and control problems,

employing the expert demonstrations [10]. Recent works highlight the importance of object-oriented and interaction-based representation learning to aid the generalizable embodied behaviour in a variety of tasks and settings [3].

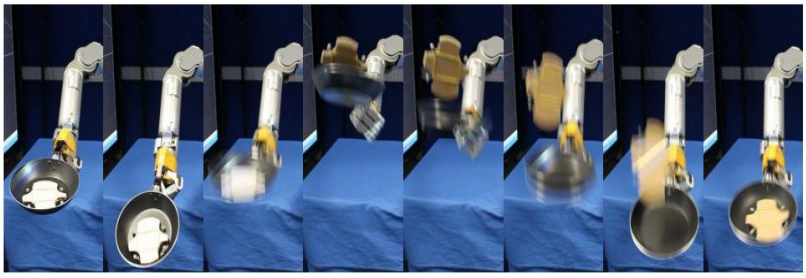

Fig. 2. Reinforcement learning perception–action loop in a pancake-flipping robot [9]

Because large-scale, fully annotated 3D data are expensive to obtain, data-centric learning strategies and synthetic data augmentation play an increasingly important role in embodied perception. A principled framework for such augmentation is Vicinal Risk Minimization, operationalized through the Mixup strategy, which generates new training samples by linear interpolation between pairs of examples and labels [11]. This approach encourages smoother decision boundaries and improves generalization in high-dimensional visual learning problems. Formally, empirical risk minimization (ERM) and Mixup is written as shown in Equations (1), (2) and (3), where f denotes the model, x_i and y_i are the input and label of the i-th sample, n is the number of samples, λ is a mixing coefficient sampled from a Beta distribution, and $\tilde{x}$ and $\tilde{y}$ are the interpolated input and label respectively.

$$R(f) = (1/n) \sum_{i=1}^{n} \ell(f(x_i), y_i) \tag{1}$$

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j \tag{2}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j \tag{3}$$

The methods conspicuously promote robustness to noise, occlusions, and domain changes in embodied settings, where visual and geometric states are significantly different in dissimilar settings. These data-centric approaches conceptualize generalization at a representational level and have been turned into part and parcel of scaled embodied learning pipelines.

3.4 Limitations and synthesis

Despite advances, there are internal tensions in the design of embodied intelligence systems. A key debate is between end-to-end and modular approaches. End-to-end reinforcement learning allows for joint optimization of perception control, yet is hampered by poor sample efficiency and low interpretability. Conversely, more stable and data-efficient modular systems that separate mapping, planning, and control are possible, particularly with learned components [12], albeit with fixed intermediate representations that incur a cost in flexibility.

A secondary tension is the geometric-semantic trade-off. Geometric reconstructions guarantee spatial accuracy at the cost of task-relevant information, whereas semantic maps generate

object-driven abstractions at the cost of metric accuracy. Hybrid architectures that optimize both features facilitate improved navigation and interaction while making them more complex and computationally costly. This unresolved contradiction reflects deeper questions about optimal internal representation for embodied reasoning.

Finally, sample efficiency and sim-to-real transfer are currently unresolved. Simulation allows data collection at scale, but policies can often fail in the real world because of visual, physical, and sensory differences. Domain randomization addresses this by exposing agents to a variety of appearance variations during training so that their resilience is improved in the deployment environment [13], but this adds to the complexity of training and does not fully solve the dynamics and contact mechanics mismatches. These contradictions suggest that despite advancements in understanding 3D scenes and learning-based control, the field is still missing a cohesive architectural or representational paradigm.

4 Experimental studies and empirical evidence

The embodied intelligence is fundamentally different from the traditional approaches to artificial intelligence, where the experimental aspect is evaluated by continuous interaction in three-dimensional settings instead of on discrete abstract tasks. In this respect, the evaluation metrics are more concerned with the success rate of the tasks performed, the efficiency of learning, its strength and generalization in different settings [7, 3]. Table 1 summarizes representative experimental paradigms, evaluation tasks, and key findings that illustrate the empirical transition from traditional disembodied AI to embodied intelligence grounded in 3D scene understanding.

Table 1. Summary of Experimental Progress from Traditional AI to Embodied Intelligence in 3D Scene Understanding

Paradigm	Representative Systems	Representative Tasks	Key Experimental Findings
Traditional (Disembodied) AI	Symbolic planners; rule-based AI [1])	Logical reasoning, planning in static environments	High performance in entirely observable and unmoving situations nonetheless it lacks strength and flexibility in dynamic or partially observable ones.
Behaviour-Based Robotics	Subsumption architecture [1]; embodied cognition models [2]	Reactive navigation, obstacle avoidance	Better real-time responsiveness but with a low level of scalability and poor spatial reasoning because of no learned internal representations.
Learning-Based 3D Scene Understanding	ScanNet, Matterport3D [5, 9]	Semantic segmentation, object detection, scene reconstruction	Formatted 3D representations are much better than raw sensory input in improving semantic understanding and spatial reasoning.

Embodied Navigation with 3D Perception	Target-driven navigation [6]	Goal-directed indoor navigation	Explicit 3D scene representation agents are more successful in their navigation, but they need high sample complexity.
Embodied Manipulation and Learning	Imitation and reinforcement learning in robotics [10]; object-centric embodied learning [3]	Object manipulation, interaction	Imitation learning has higher sample efficiency while object-centric representations are more cross-environmentally generalized.

Table 1 demonstrates that experimental assessment has begun to transition beyond stagnant symbolic tests and into embodied tasks that involve perception of space, interaction, and generalization. In navigation and manipulation problems, systems that explicitly represent 3D scenes reliably out-compete raw sensory or symbolic representations [6], though sample efficiency and sim-to-real transduction may be issues.

Recent studies are based on standardized simulation environments such as AI2-THOR, and provide photo-realistic, interacting indoor scenes with physics, permitting reproducible comparisons of representations, learning policies, and exploration policies [14].

Relevant to the aforementioned, previous disembodied AI was enhanced in controlled, fully observable environments but failed in dynamic or partially observable ones [1]. Behavior-based robotics were responsive but not scalably spatially reasoning with no learned representations [2].

Semantic segmentation, object detection, and scene reconstruction were advanced with large-scale 3D datasets such as ScanNet and Matterport3D [5, 9]. Experimental findings support the idea that agents employing structured 3D representations always outperform unstructured sensory agents in body challenges [6].

Reinforcement learning makes goal-aligned navigation possible, though it would need large samples [6]. Mimicking learning enhances efficiency and stability, especially in the area of accurate manipulation [10]. Modern object-based embodied systems show better generalization and inter-environment transfer [3].

5 Emerging trends

Current developments have pointed to a move towards foundation models in which vision, language, and action are represented in the same framework. Large-scale cross-modal alignment between text and image models, like CLIP, makes it possible to create language-conditioned perception in embodied systems and to specify tasks [14].

Simultaneously, neural scene representations like Neural Radiance Fields (NeRF) provide a continuous and differentiable alternative to more traditional voxel or mesh-based 3D representations, which can be used to perform high-fidelity view synthesis, as well as make geometric decisions [15]. Such representations are being investigated more and more as models of the inner world to plan and interact.

The collective trends appear to point towards large-scale, multimodal and physically-grounded foundation models of embodied intelligence.

6 Conclusion

The world has moved past abstract, symbolic systems to embodied intelligence that needs 3D scene understanding. Early AI was good at controlled tasks but poor in dynamic environments that required perception and spatial reasoning. This promoted the transition to embodied cognition, whereby intelligence arises as a result of perception-action-environment interactions.

The development of 3D data capture, representation learning, and annotated datasets has improved the capability of agents to learn spatial structures, object relations, and semantic contexts. Agents with structured 3D representations are more successful at navigation and manipulation tasks than agents limited to raw sensory information. Adaptive, goal-directed behaviours are further facilitated by reinforcement and imitation learning.

Nonetheless, there are ongoing obstacles: excessive data loads, ineffective physical and causal reasoning, scaling problems, and simulation-to-real transfer gaps. Future studies need to include sample efficient learning, unified multimodal representations, integrated physical-causal models, and robust real-world testing to build strong, generalizable embodied AI systems.

References

[1] Brooks, R.A.: Intelligence without representation. Artificial Intelligence. 47(1-3), pp. 139-159 (1991)

[2] Pfeifer, R., Bongard, J.: How the Body Shapes the Way We Think: A New View of Intelligence. The MIT Press, Cambridge, MA (2006)

[3] Zheng, Y., Yao, L., Su, Y., Zhang, Y., Wang, Y., Zhao, S., Zhang, Y., Chau, L.-P.: A Survey of Embodied Learning for Object-centric Robotic Manipulation. Machine Intelligence Research. 22(4), pp. 588-626 (2025)

[4] Chang, A., Dai, A., Funkhouser, T., Halber, M., Niebner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from RGB-D Data in Indoor Environments. In: International Conference on 3D Vision (3DV). Qingdao, China, pp. 667-676 (2017)

[5] Zhu, Y., Mottaghi, R., Kolve, E., Lim, J.J., Gupta, A., Fei-Fei, L., Farhadi, A.: Target-driven visual navigation in indoor scenes using deep reinforcement learning. In: IEEE International Conference on Robotics and Automation (ICRA). Singapore, pp. 3357-3364 (2017)

[6] Vernon, D., Metta, G., Sandini, G.: A Survey of Artificial Cognitive Systems: Implications for the Autonomous Development of Mental Capabilities in Computational Agents. IEEE Transactions on Evolutionary Computation. 11(2), pp. 151-180 (2007)

[7] Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M.: Deep Learning for 3D Point Clouds: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 43(12), pp. 4338-4364 (2021)

[8] Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Niessner, M.: ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA, pp. 2432-2443 (2017)

[9] Kormushev, P., Calinon, S., Caldwell, D.: Reinforcement Learning in Robotics: Applications and Real-World Challenges. Robotics. 2(3), pp. 122-148 (2013)
[10] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond Empirical Risk Minimization. In: International Conference on Learning Representations (ICLR). Vancouver, Canada (2018)
[11] Gupta, S., Tolani, V., Davidson, J., Levine, S., Sukthankar, R., Malik, J.: Cognitive Mapping and Planning for Visual Navigation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA (2017)
[12] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Vancouver, Canada, pp. 23-30 (2017)
[13] Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Gordon, D., Zhu, Y., Gupta, A., Farhadi, A.: AI2-THOR: An Interactive 3D Environment for Visual AI. arXiv preprint (2017)
[14] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models from Natural Language Supervision. In: International Conference on Machine Learning (ICML), pp. 8748-8763 (2021)
[15] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In: European Conference on Computer Vision (ECCV). Glasgow, UK, pp. 405-421 (2020)