Deep Learning for Perceptive and Autonomous Robotic Systems

Anh Nguyen

21 May 2021

Abstract. Autonomous robots must do more than recognise objects or reconstruct their surroundings. They must interpret how the environment can be used, understand human intentions, combine information from multiple sensors, and translate perception into safe physical action. This paper presents a connected body of my research addressing these challenges across robotic manipulation, navigation, wearable systems, and robot-assisted intervention. Beginning with geometric and contextual understanding of three-dimensional scenes, the work progressed towards object affordance detection, task-oriented manipulation, vision-language learning, multimodal navigation, and learning-enabled medical robotics. Despite the diversity of applications, the central research question remains consistent: how can robots transform complex and uncertain sensory observations into meaningful, reliable, and executable actions?

1. From Geometric Perception to Action-Oriented Understanding

The research begins with the problem of organising unstructured three-dimensional sensor data. The survey of 3D point-cloud segmentation methods [1] established segmentation as a foundation for autonomous mapping, navigation, and interaction. Contextual labelling with conditional random fields then introduced relationships between neighbouring points and semantic classes, enabling local geometric evidence to be interpreted within a broader scene context [2]. Together, these studies addressed a fundamental requirement: before a robot can act intelligently, it must first impose meaningful structure on an uncertain physical world.

Semantic recognition alone, however, does not explain how an object can be used. This limitation motivated a shift from object-centred perception towards action-centred perception through affordance learning. Early work used convolutional neural networks to predict functional object regions directly from visual observations [3]. Preparatory object reorientation extended this idea by recognising that manipulation success depends not only on choosing the right object, but also on arranging it into a configuration suitable for the intended task [4].

The subsequent object-based affordance framework combined object detection, convolutional segmentation, and dense conditional random fields to improve functional reasoning in cluttered scenes [5]. AffordanceNet then unified object localisation, classification, and affordance segmentation in a single end-to-end architecture [6]. The same action-oriented view also informed foothold reasoning for humanoid locomotion, where curved 3D surface patches were evaluated according to their suitability for stable contact [7]. Across these works, perception evolved from identifying what is present to determining what actions the environment makes possible.

2. Learning Robot Intent from Demonstrations and Language

Action-oriented perception was next extended from static objects to dynamic human demonstrations. The video-to-command framework learned to translate sequences of human activity into compact commands for robotic manipulation [8]. Visual features captured what was happening in each frame, while recurrent models represented the temporal structure of the demonstrated task and generated corresponding command words. This established a direct bridge between observing human behaviour and producing instructions that a robot could execute.

V2CNet strengthened this connection by jointly learning command generation and fine-grained action recognition [9]. The two tasks provided complementary supervision: action recognition helped distinguish visually similar manipulations, while command generation translated those distinctions into an interpretable representation of intent. In parallel, object captioning and natural-language retrieval showed how language can describe objects through their appearance, attributes, and functional properties rather than through fixed category labels alone [10]. These studies anticipated later vision-language robotic systems by treating language as an interface between human intention, visual perception, and physical execution.

The broader programme was consolidated in the doctoral work on scene understanding for autonomous manipulation [11]. Its central argument was that advances in computer vision become most valuable to robotics when their outputs are grounded in physical tasks. Affordance detection, action understanding, language generation, and robotic demonstration were therefore developed as connected components of an autonomous manipulation system rather than as isolated recognition problems.

3. Robust Autonomy in Complex Environments

Moving beyond structured manipulation required perception methods that remained reliable under rapid motion, difficult illumination, uncertain terrain, and incomplete sensor observations. Real-time six-degree-of-freedom relocalisation for event cameras addressed this challenge using stacked spatial LSTM networks [12]. By learning from asynchronous visual events, the work explored a sensing and modelling strategy suited to situations in which conventional frame-based cameras may struggle.

BeetleBot provided a physical platform for integrating perception, localisation, navigation, and artificial intelligence in realistic environments [13]. This platform-level perspective was complemented by the deep multimodal fusion network for autonomous navigation [14], which combined visual, range, and three-dimensional information. The key insight was that robust autonomy need not depend on one superior sensor. It can instead emerge from learning how complementary sensing modalities compensate for one another as environmental conditions change.

4. From General Robotics to Human-Centred and Medical Systems

The same principles were subsequently transferred to systems in which physical uncertainty and human safety are especially important. In a multi-degree-of-freedom shoulder sensing exosuit, learning-based nonlinearity compensation was used to estimate human motion for real-time teleoperation [15]. This work demonstrated how data-driven perception can create an intuitive and responsive connection between human movement and robotic action, even when compliant sensing introduces complex nonlinear behaviour.

For surgical robotic tools, purely analytical models can generalise well but may simplify important nonlinear effects, while purely data-driven models can capture complexity but depend strongly on their training data. The hybrid kinematic model combined analytical knowledge with data-driven learning to improve robotic control [16]. This reflects a broader principle across the research programme: machine learning should complement physical and geometric knowledge rather than automatically replace it.

The endovascular robotics studies completed the transition from perception to collaborative autonomy. Optical-flow-guided warping was integrated into a real-time catheter segmentation framework to exploit temporal continuity in fluoroscopic image sequences [17]. Reliable catheter localisation provided an essential perceptual foundation for learning-enabled intervention. Generative adversarial imitation learning was then applied to collaborative robot-assisted catheterisation, enabling policies to be learned from expert demonstrations [18]. Together, these studies connected visual tracking, temporal learning, robot control, human expertise, and procedural safety within a clinically motivated application.

5. A Coherent Research Trajectory

Although the publications span different robots and application domains, they form a coherent progression. The early work asked how raw geometry could be organised and labelled. The affordance studies then asked how perception could reveal opportunities for action. Video and language models addressed how robots could infer human intent and express task meaning. Multimodal navigation expanded these capabilities to uncertain and unstructured environments, while the wearable and medical robotics studies translated them into systems that interact directly with people and must therefore operate with greater precision, responsiveness, and safety.

Across this trajectory, the role of deep learning also becomes more mature. It first appears as a tool for semantic prediction, then as a mechanism for learning temporal intent and fusing heterogeneous sensors, and finally as one component within hybrid systems that incorporate geometry, analytical modelling, human demonstrations, and physical constraints. The unifying contribution is therefore not a single neural network, but an approach to robotics in which learning is consistently grounded in the requirements of embodied action.

6. Conclusion

A capable autonomous robot must connect three levels of understanding: geometry describes the physical world; semantics and affordances explain its meaning; and learned action models determine how the robot should respond. The reviewed work contributes methods at each level while repeatedly grounding them in physical robotic systems. Collectively, the publications present a pathway from robots that merely observe their surroundings towards robots that can interpret, communicate, navigate, manipulate, and collaborate safely with people.

References

[1] A. Nguyen and B. Le, "3D Point Cloud Segmentation: A Survey."

[2] A. Nguyen and B. Le, "Contextual Labeling 3D Point Clouds with Conditional Random Fields," in Intelligent Information and Database Systems, 2014.

[3] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Detecting Object Affordances with Convolutional Neural Networks," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016.

[4] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Preparatory Object Reorientation for Task-Oriented Grasping," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016.

[5] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Object-Based Affordances Detection with Convolutional Neural Networks and Dense Conditional Random Fields," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017.

[6] T. T. Do, A. Nguyen, and I. Reid, "AffordanceNet: An End-to-End Deep Learning Approach for Object Affordance Detection," in IEEE International Conference on Robotics and Automation (ICRA), 2018.

[7] D. Kanoulas, C. Zhou, A. Nguyen, G. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Vision-Based Foothold Contact Reasoning Using Curved Surface Patches," in IEEE-RAS International Conference on Humanoid Robotics (Humanoids), 2017.

[8] A. Nguyen, D. Kanoulas, L. Muratore, D. G. Caldwell, and N. G. Tsagarakis, "Translating Videos to Commands for Robotic Manipulation with Deep Recurrent Neural Networks," in IEEE International Conference on Robotics and Automation (ICRA), 2018.

[9] A. Nguyen, T. T. Do, I. Reid, D. G. Caldwell, and N. G. Tsagarakis, "V2CNet: A Deep Learning Framework to Translate Videos to Commands for Robotic Manipulation," 2019.

[10] A. Nguyen, Q. D. Tran, T. T. Do, I. Reid, D. G. Caldwell, and N. G. Tsagarakis, "Object Captioning and Retrieval with Natural Language," in IEEE/CVF International Conference on Computer Vision Workshops, 2019.

[11] A. Nguyen, "Scene Understanding for Autonomous Manipulation with Deep Learning," arXiv:1903.09761, 2019.

[12] A. Nguyen, T. T. Do, D. G. Caldwell, and N. G. Tsagarakis, “Real-Time 6DoF Pose Relocalization for Event Cameras with Stacked Spatial LSTM Networks,” in IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[13] A. Nguyen, E. Tjiputra, and Q. D. Tran, “BeetleBot: A Multi-Purpose AI-Driven Mobile Robot for Realistic Environments,” in UKRAS20 Conference: Robots into the Real World, 2020.

[14] A. Nguyen, N. Nguyen, K. Tran, E. Tjiputra, and Q. D. Tran, “Autonomous Navigation in Complex Environments with Deep Multimodal Fusion Network,” in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020.

[15] R. J. Varghese, A. Nguyen, E. Burdet, G.-Z. Yang, and B. P. L. Lo, “Nonlinearity Compensation in a Multi-DoF Shoulder Sensing Exosuit for Real-Time Teleoperation,” in IEEE International Conference on Soft Robotics (RoboSoft), 2020.

[16] F. Cursi, A. Nguyen, and G.-Z. Yang, “Hybrid Data-Driven and Analytical Model for Kinematic Control of a Surgical Robotic Tool,” arXiv:2006.03159, 2020.

[17] A. Nguyen et al., “End-to-End Real-Time Catheter Segmentation with Optical Flow-Guided Warping during Endovascular Intervention,” in IEEE International Conference on Robotics and Automation (ICRA), 2020.

[18] W. Chi et al., “Collaborative Robot-Assisted Endovascular Catheterization with Generative Adversarial Imitation Learning,” in IEEE International Conference on Robotics and Automation (ICRA), 2020.