Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 5, 2025

FROM CLASSICAL VISION TO DEEP LEARNING: A SURVEY OF COMPUTER VISION METHODS IN ROBOTICS

Shafiq Hussain[1], Muhammad Aakash Imtiaz[*2], Hassan Revel[3], Muhammad[4], Adeen Amjad[5], Aleena Jamil[6], Waqar Ahmad[7], Arslan Ali Mansab[8], Muhammad Hamza Akbar[9], Muhammad Waqas[10]

[1,*2,3,4,5,6,7,8,9,10]*Department of Computer Science, University of Sahiwal, Sahiwal, Pakistan*

[1]drshafiq@uosahiwal.edu.pk, [*2]akashimtiaz123@gmail.com, [3]hassanrevelai@gmail.com, [4]muhammed.beig@gmail.com, [5]adeen.amjad@uosahiwal.edu.pk, [6]aleena.jamil_vf@uosahiwal.edu.pk, [7]waqarahmad@uosahiwal.edu.pk, [8]arslansli@uosahiwal.edu.pk, [9]hamzaakbar@uosahiwal.edu.pk, [10]bssit.10.02@gmail.com

DOI: https://doi.org/10.5281/zenodo.18051202

Keywords
Computer Vision, Robotics, Scene Understanding, Semantic Perception, Object Detection, Semantic Segmentation, Deep Learning

Article History
Received: 25 March 2025
Accepted: 14 May 2025
Published: 27 May 2025

Corresponding Author:
Muhammad Aakash Imtiaz

Abstract
Understanding the scene is one of the fundamental principles of robotics autonomy. Robots need to abstract, locate, and infer objects in the environment they should eventually navigate, handle, or interact with safely. Traditionally, such capabilities have been achieved through highly crafted pipelines based on geometric reasoning and hand-engineered features. Over the past decade, however, deep learning has revolutionized computer vision so that robots can process rich scenes without explicit feature retrieval by training on data. In this survey, we present a literature review of computer vision algorithms used for scene understanding in robotics, from traditional to modern deep learning methods. We introduce a novel taxonomy that differentiates between geometric mapping, object detection, semantic segmentation, and high-level scene understanding; we survey existing methods for each category; and evaluate the accuracy, robustness, and computational requirements of such techniques through extensive experimentation. We also overview key benchmark datasets which have spurred development and emphasize the open issues and avenues for future work.

INTRODUCTION

The robots work in different, sometimes unpredictable environments. For tasks such as navigation, manipulation or hu-man–robot interaction, a robot needs to detect and understand its environment. This ability, known as scene understanding, entails constructing a representation of the environment that encompasses both geometric structure and semantic contexts, separating surfaces, objects, and relations between them. The first robotic vision systems made use of classical computer vision methods: features were extracted using hand-crafted operators and geometric reasoning was applied to recover structure and recognize objects. Instances are monocular SLAM based on sparse feature tracking [1] or object detectors relying on descriptors such as SIFT and SURF [2], [3]. These approaches worked quite well in structured setups, but failed in complex and cluttered scenes without interpreting high-level semantics. The rise of deep learning was a game changer. Convolutional neural networks (CNNs) learned on large-scale datasets like ImageNet [4] have exhibited the impressive advantages to

learn hierarchical visual representations. Classical methods were very quickly outdone on standard benchmarks by CNN-based detectors (e.g., R-CNN [5] and its descendents [6]) and segmentation networks (e.g., Fully Convolutional Networks [7]). In robotics, deep models allow for highly precise object detection and semantic segmentation, even in unstructured set-tings [8], [9]. More recently, transformer models has brought alternative self-attention-based models for vision [10], [11].

Despite these developments, classical approaches remain essential, in particular for geometric mapping [12] and localization, where accuracy of physical consistency is required [13]. A combination of learned perception with geometric reasoning is frequently the best approach.

Some surveys have covered specific parts of robotic scene understanding. Cadena et al. give an introduction to SLAM from the beginning until robustness and real time systems [13]. Tang et al. survey the deep learning in perception and navigation of autonomous system [14], while Chen et al. deep learning maps in visual localization and mapping. Grigorescu et al. summarize deep learning techniques in autonomous driving [8], and Ni et al. research deep learning methods for robot scene understanding [15]. Our review completes and extends these pieces of work by side-by-side comparison between classical- as well as deep-learning-based vision in a shared taxonomy, placing scene-studying tasks rather than localization-related usage at the center of our attention, giving performance comparisons on a comparative standard, open challenges and future directions.

The rest of this paper is organized as follows. Section II introduces a taxonomy of scene understanding techniques in robotics. In Section III, we recall classical computer vision techniques, while in Section IV we present recent deep learning methods such as CNN and scene graphs/toy models/transformers. We summarize benchmark datasets and evaluation measures in Section V. Section VI contrasts selected classical and deep approaches and discusses their applicability for different robotic domains. Section VII discusses the open challenges, and in Section VIII, we present some potential research directions. Section IX concludes the survey.

I. Taxonomy of Scene Understanding Methods

Room scene understanding is a multidisciplinary under-taking and includes both classical and learned methods. We classify methods along two dimensions: (i) approach paradigm capturing whether it corresponds to traditional computer vision or modern deep learning; and (ii) task describing the type of output or representation being generated. This taxonomy is illustrated in figure 1.

A. Approach Paradigm

Classical approaches are based on hand-crafted features (e.g.,edges and corners) and explicit models of geometry and appearance. Feature-based SLAM [1], [16], HOG-SVM pedestrian detectors [17] and CRF-based segmentation [18], [19] are typical examples of this paradigm. These techniques are usually finicky to set up, but offer interpret-ability and very fine geometric reasoning.

Deep learning techniques Deep learning uses data-driven models, in the form of deep neural networks, for learning features and representations in an end-to-end manner from the data. Such known embodiments include CNNbased object detectors [5], [6], [20]–[22], semantic segmentation networks [7], [23]–[26]. These models often attain better performance, but rely on large amounts of data and computational resources.

B. Task Categories

Our method categorizes the methods further according to the primary goal that they tackle:

- **Geometric mapping localization**: Computing the 3D layout of the environment and the robot's pose. Classical solutions include the sparse feature-based SLAM [1], [16] and direct methods [27], [28]. Deep learning provides learned depth estimation and feature descriptors [13].
- **Object detection and recognition**: Find

the positions of, and classify objects. Classical detectors employ descriptors such as SIFT, SURF or HOG [2], [3], [17], [29]. The family of deep detectors contains two-stage region-based CNNs (R-CNN, Faster R-CNN) [5], [6], single-stage detectors (YOLO, SSD) [20], [21], and in-stance segmentation methods such as Mask R-CNN [22].

- **Semantic segmentation** Pixelwise annotation of images into semantic classes. In previous work graph-cuts and CRFs have been employed [18], [19]. The accuracy in deep segmentation networks such as FCN [7], SegNet [23], U-Net [24], PSPNet [25], and DeepLabv3+ [26] has improved massively.
- **Scene representation and reasoning**: Creating structured scene representations (e.g., scene graphs) which capture objects along with spatial relations. Recent works generalize scene graphs to 3D and robotics [30], [31] as well as study transformer-based models for fully perceiving the world [10], [11]. Topological maps and semantic SLAM are used by traditional robotics [32], [33].

Many modern systems combine several different tasks (se-mantic SLAM fuses the mapping, detection and segmentation). Hence our taxonomy is a conceptual one not rigid bins.

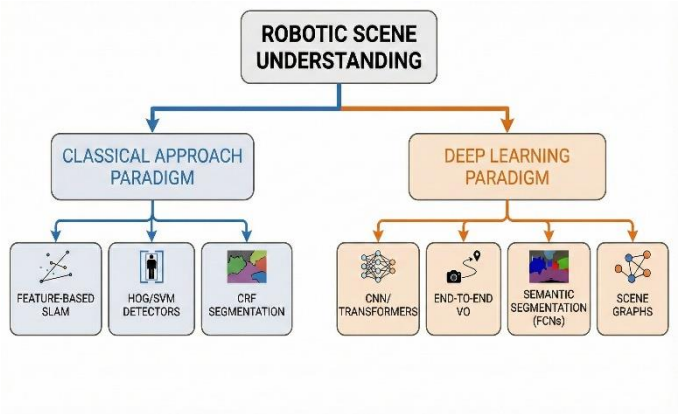

Fig. 1. Taxonomy of scene understanding methods in robotics, highlighting classical and deep learning approaches across key tasks.

II. Classical Vision Methods in Robotics

Early robotic perception was dominated by classical vision techniques. They rely on breaking down the vision problem into several steps that include feature extraction, fitting of models and classification. We provide a summary of the state-of-the-art approaches to mapping and localization, object recognition and semantic scene understanding.

A. Geometric Mapping and Localization

Visual SLAM estimates the pose of a robotic system while it builds a map of its surroundings. Considering SLAM with monocular cameras, Davison [1] first presented a real-time EKF-based monocular SLAM which tracks sparse corner features. Robust feature detectors and descriptors such as SIFT [2], SURF [3] and ORB [16], [34] have led to accurate feature matching over time sequences. ORB-SLAM2 [16] incorporates FAST keypoints, a bag-of-words loop closure detection and map optimization and is state of the art in terms of accuracy. In parallel to feature-based methods, there are direct counterparts that purely reduce photometric error among images even without feature extraction. LSD-SLAM [27] does semi-dense mapping while DTAM [28] constructs dense depth maps by solving a variational formulation. Classical SLAM still remains as a best rival for accurate, drift-free localization due to its explicit model of uncertainties [13].

B. Object Detection and Recognition

Prior to deep learning, object detection pinned hopes on hand-crafted features and classifiers. The Viola–Jones detector
[35] proposed the cascade of boosted Haar-like features for real-time face detection. Dalal and Triggs [17] introduced the linear SVM-based Histogram of Oriented Gradients (HOG) descriptor that showed remarkable performance in pedestrian detection. Felzenszwalb et al. proposed the Deformable Parts Model (DPM) [29] that encodes objects as a composition of HOG filters placed in a deformable configuration. Although successful in smaller-task setups, these methods performed poorly on large intra-class variation and cluttered scenes and were soon outperformed by CNN-based detectors.

C. Semantic Segmentation and Scene Labelling

Classical segmentation technique segment the image into region based on low-level information (color, texture, proximity) and classify it using probabilistic model. Shi and Malik [36] posed the segmentation problem as graph partition (normalized cuts). Boykov and Jolly [18] applied graph-cut optimization for interactive segmentation. In order to add semantics, models such as TextonBoost [19] trained boosted classifiers on textons and smoothed predictions in a Conditional Random Field (CRF). Such approaches, although worked well for limited datasets, they did not have the ability to learn complex visual appearance and contextual relationships that deep methods extract.

I. Robotic Vision with Deep Learning

Vision for robotics has been transformed by deep learning, providing accurate detection, segmentation and holistic scene understanding. We look at the major deep learning methods used for various tasks.

A. Deep Object Detection and Recognition

Deep object detectors have been facilitated since large image net kind datasets were trained using CNNs [4]. R-CNN
[5] classifies region proposals using one convolutional neural network (CNN) and has significantly improved detection performance. Faster R-CNN [6], which incorporates the Region Proposal Network in CNN can be trained end-to-end and has real-time performance. One-stage detectors such as YOLO [20] and SSD [21] regress both bounding boxes and scores at all positions, however sacrificing accuracy to gain speed. Mask R-CNN [22] is a generalization of Faster R-CNN that simultaneously predicts instance masks in addition to bounding boxes, allowing robots to see object shapes.

B. Semantic Segmentation and Scene Parsing

One of the prominent models used for deep scene perception is the Segmentation Network which assigns semantic labels to pixels. Instead of the fully connected layers used in classification CNNs, Fully Convolutional Networks (FCNs) [7] use convolutional and up sampling layers. SegNet [23] employs an encoder–decoder network architecture with pooling indices for fast, efficient feature up sampling. U-Net [24] adds skip connections between encoder and decoder layers, keeping the fine details. PSPNet [25] performs pyramid pooling to accumulate context in multi-scale, and DeepLabv3+ [26] is based on atrous convolution and a encoder–decoder architecture for high accuracy. Such networks can allow robots to generate dense semantic maps of their environment.

C. Scene Graphs and Relational Understanding

Scene graphs capture entities in a scene as nodes, and their connection (e.g., spatial or functional) as edges, thereby offering structured abstraction of a scene. Armeni et al. pro-posed 3D scene graphs that integrate semantics, geometry and room layouts [30], Kim et al. suggested spares 3D scene graphs designed for robot navigation [31]. In robotics, scene graph creation generally consists of CNN-based detection and pose estimation combined with a construction step from these detected objects. They support high-level reasoning, task planning, and multi-modal interaction.

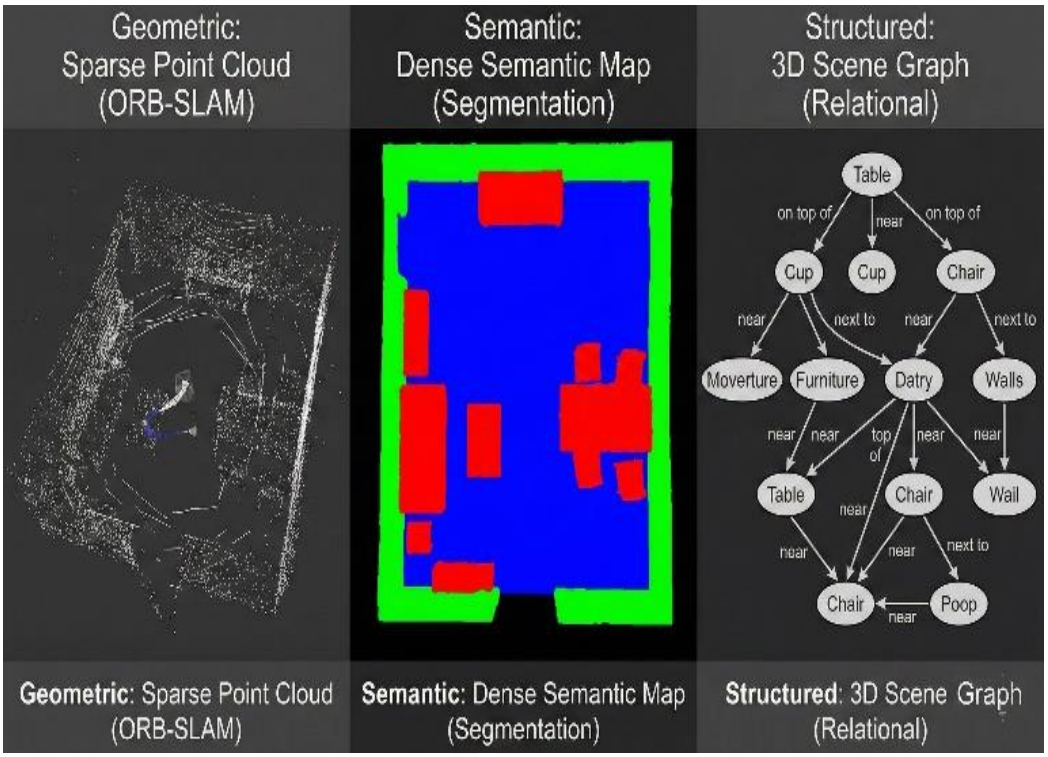

Fig. 2. Comparison of scene representations. Top: Sparse geometric feature map typical of classical SLAM [31]. Middle: Dense semantic segmentation providing class labels per pixel [22]. Bottom: 3D Scene Graph, abstracting the environment into objects and spatial relationships for high-level reasoning. [24]

D. Transformer-Based Perception Models

Originally conceived for natural language processing, Trans-former architectures have been repurposed for vision. The Vision Transformer (ViT) [10] regards an image as a sequence of patch tokens, and achieves strong performance by learn-ing global dependencies from self-attention. Detection Trans-former (DETR) [11] uses an encoder–decoder transformer to perform object detection, directly producing a set of objects without employing anchor boxes or non-maximum suppression. Transformers also enable a framework for multi-modal perception, to allow for image and depth map integration as well as other sensor data. While these models are not computationally efficient, they provide a new direction for generative holistic scene understanding in robotics.

I. Benchmark Datasets and Evaluation

In robotic vision, progress is spurred on by benchmark datasets that supply high-quality data and ground-truth annotations with which to evaluate new methods. We overview major datasets for scene understanding.

KITTI [37] consists of stereo images, LIDAR scans and the corresponding ground-truth for tasks such as stereo depth estimation, optical flow, SLAM, object detection and tracking. It proved useful for the evaluation of classical visual odometry as well as deep learning approaches.

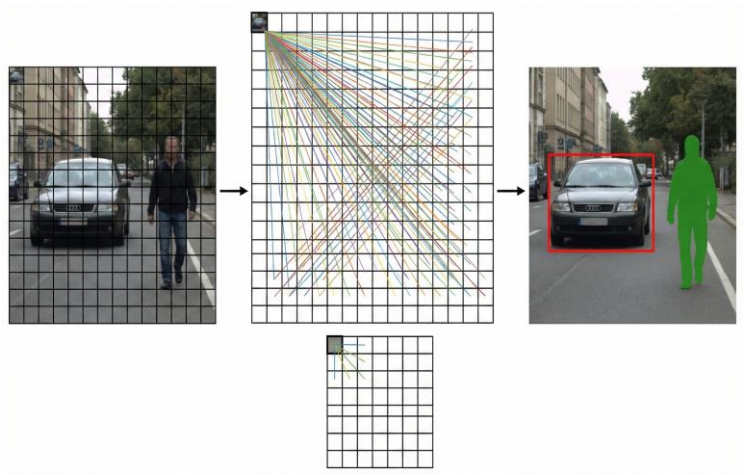

Fig. 3. Visualizing the Transformer paradigm in robotics. Unlike CNNs that process local neighborhoods, Vision Transformers (ViT) split images into patches and use self-attention mechanisms to model long-range dependencies across the entire scene, facilitating holistic scene understanding.

Cityscapes [38] contains high-resolution street scenes with pixel-level semantic mapping of 30 classes to support progress in the field of semantic urban scene understanding. This dataset is used for very deep segmentation network like PSPNet and DeepLabv3+.

For example, PASCAL VOC [39] is an object benchmark dataset that contains images annotated with bounding boxes and/or segmentation masks of objects from a set of 20 categories, on which classical detectors (e.g.,DPM), early deep detectors (R-CNN, Fast R-CNN) were evaluated.

MS COCO [40] augments PASCAL with 80 categories and denser context (instance segmentation, keypoint annotations). It is still a major training data set for today's detectors and segmentations models.

ImageNet [4] provides more than a million images labeled according to their class. Pre-training on ImageNet is a common practice that results in features transferable for robotics.

NYU Depth V2 and SUN RGB-D [41], [42] offer RGB-D images of indoor scenes with semantic labels for service robot to perform segmentation or 3D understanding.

ScanNet [43] provides 3D scans of indoor scenes with semantic labeling for 3D scene analysis and understanding research.

RGB-D images and 6-DoF pose annotations of warehouse items from the Amazon Picking Challenge datasets [44] in-spired us to use deep methods for object pose estimation and grasping.

Evaluation metrics vary by task. Object detection can be evaluated with mean average precision (mAP), segmentation via mean inter-section-overIou (IoU), and SLAM through trajectory error statistics [13]. Instead of listing metrics, in Table I we summarize popular public datasets for robotic vision research.

II. Comparative Analysis of Classical and Deep Methods

We then compare representative classical and deep visions approaches in this integrated view based on their accuracy, computational cost and applicability across various robotic domains. Table II presents some of the main algorithms considered in this survey. While deep learning is the contender to rule detection and segmentation, classical methods continue to be useful for accurate geometry estimation and low-latency applications.

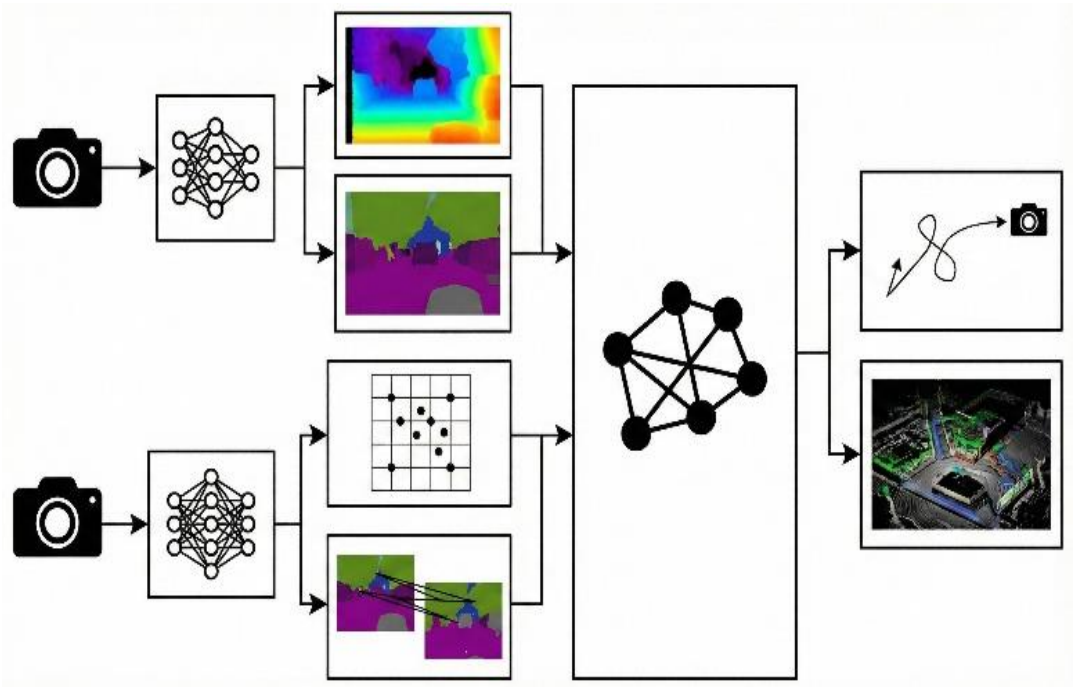

Fig. 4. Conceptual framework of a Hybrid Robotic Vision System. Deep neural networks give dense depth and semantic priors (bottom) that, combined with classical feature-based tracking (top), are fused in a common optimization backend (e.g., Factor Graph) for robust state estimation.

Generally, deep method highly out perform classic solutions in detection and segmentation benchmark based on mAP and IoU score [38]–[40]. Deep models are also able to generalize on complex environments when trained with large variety of datasets. Nonetheless, classical approaches are useful for high-precision geometry (like SLAM) and as ingredients in hybrid systems. For instance, ORB-SLAM2 employs hand -crafted features combined to deep loop -closure detection [16]. Hybrid methodologies that combine learned perception with classical optimization are likely to prevail in near-term robotics.

III. Open Challenges

Despite advancements similar to those described above, there are several remaining open challenges to achieve robust scene understanding for robotics:

- **Generalization and domain shift**: Deep models can experience catastrophic performance collapse when they are put into the real-world different from which the model is trained. Domain adaptation, few-shot learning and simulation-to-real transfer are the subject of intense research efforts [8], [9].
- **Real-time computation** Many of state-of-the-art models are Deploying them in embedded robotic platforms requires model compression, efficient architectures and hardware specialization.
- **Multi-modal integration:** Robots are forced to integrate vision with other sensing modalities (like LIDAR, depth, inertial measurements). Discovering ways to integrate heterogeneous sources of data robustly is still an open problem.
- **Dynamic and interactive scenes:** Scene understanding of moving agents (human, vehicles) and prediction of their behavior require high-level temporal reasoning.

TABLE I MAJOR BENCHMARK DATASETS FOR ROBOTIC VISION

Dataset	Year	Modalities	Primary Tasks
KITTI [37]	2012	Stereo, LiDAR	Stereo depth estimation, optical flow, object detection, track-ing, SLAM
Cityscapes [38]	2016	RGB images	Semantic segmentation, urban scene understanding
PASCAL VOC [39]	2010	RGB images	Object detection, instance segmentation
MS COCO [40]	2014	RGB images	Object detection, instance segmentation, keypoint detection, image captioning
ImageNet [4]	2009	RGB images	Large-scale image classification, representation pre-training
NYU Depth V2 [41]	2012	RGB-D	Indoor semantic segmentation, depth-based scene under-standing
SUN RGB-D [42]	2015	RGB-D	3D object detection, semantic segmentation
ScanNet [43]	2017	3D scans	3D semantic segmentation, surface reconstruction
Amazon Picking [44]	2017	RGB-D	6-DoF object pose estimation for robotic manipulation

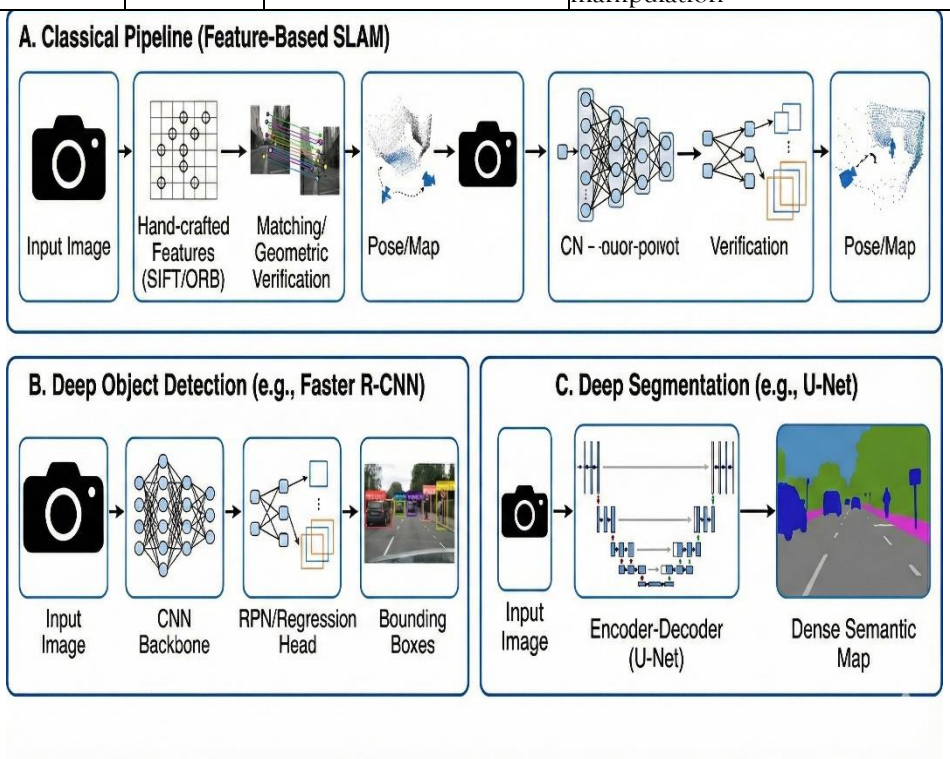

Fig. 5. Evolution of robotic vision architectures. (a) Classical Pipeline: Relies on hand-crafted feature extraction (e.g., SIFT, ORB) and geometric reasoning [cite: 1517]. (b) Deep Detection: Uses CNN backbones (e.g., ResNet) with region proposal networks to regress bounding boxes [cite: 1536]. (c) Deep Segmentation: Employs encoder-decoder structures (e.g., U-Net) to produce pixel-wise semantic labels [cite: 1542].

- **Interpretability and confidence**: Deep networks are frequently black boxes. It is important to be able to estimate uncertainty and generate interpretable explanations for safety and trust.
- **Lifelong learning and adaptivity**: Robots ought to be able to learn over time rather than being trained from scratch each time a new skill is taught, as well as adapting to changes with no catastrophic forgetting. It remains an open problem to integrate lifelong learning into deep vision.

IV. Future Directions

Attaining scene understanding in robotics will take scaling both new algorithms and integrated robotic systems. We high-light promising directions:

- **Hybrids** Sharing features of both neural networks and physical models can help in bridging the gap between these two paradigms. Examples of this include differen-tiable SLAM, learned feature descriptors incorporated in optimisation or an end-to-end neural SLAM.
- **Self-supervised and cross-modal learning:** A robot can gather huge amounts of unlabelled datawhile in operation. Self-supervision from geometry, temporal continuity or cross-modal consistency can alleviate the dependence on manual annotation.
- **Few-shot and on-the-fly learning:** Learning methods that can generalize to new objects or tasks from just a few examples will be important for robots in open-set domains.
- **Hierarchical reasoning and knowledge modulation**: Modulating the high-level reasoning by either using pre-processed images from a low-level perception model (e.g., object recognition) or updating the model using higher-level concepts can infer meaning of scenery at different levels beyond detection segmentation, such as inferring affordances and causality. The persistent world models, which accumulate the knowledge over the time will enable robots to track objects, detect changes and plan accordingly. Memory-enhanced networks and explicit mapping models might be involved.
- **Edge–cloud integration**: Decentralized perception archi-tectures in which robots work together and leverage edge cloud servers for offloading heavy computations may empower the large neural network model with a small on?board resource overhead.
- **New benchmarks and metrics**: New benchmarks must reflect the challenges of real-world robotics scenar-ios with long-term autonomy, dynamic scenes and multi-modal perception. Evaluation dreams need to in-clude safety/reliability as well as accuracy.

TABLE II Representative Classical and Deep Vision Methods for Robotic Scene Understanding

Method / Authors	Year	Type	Key Contribution
Davison [1]	2003	Classical SLAM	Real-time monocular EKF-SLAM tracking sparse features, pioneering visual SLAM for robotics
Lowe [2]	2004	Feature Descriptor	SIFT: scale-invariant keypoint detection and description, enabling robust matching across views
Bay et al. [3]	2006	Feature Descriptor	SURF: fast robust local descriptors using integral images for real-time performance
Harris & Stephens [34]	1988	Corner Detection	Harris corner detector identifying salient interest points for tracking and mapping
Mur-Artal & Tardo´s [16]	2017	Classical SLAM	ORB-SLAM2: feature-based monocular/stereo/RGB-D SLAM with loop closure and map optimisation

Engel et al. [27]	2014	Direct SLAM	LSD-SLAM: semi-dense direct SLAM using photometric error 986ptimization for large-scale mapping
Newcombe et al. [28]	2011	Direct SLAM	DTAM: dense tracking and mapping with GPU-accelerated depth estimation
Viola & Jones [35]	2004	Classical Detection	Haar-cascade face detector enabling fast, cascade-based object detection
Dalal & Triggs [17]	2005	Classical Detection	HOG descriptor with linear SVM for robust pedestrian detection
Felzenszwalb et al. [29]	2010	Classical Detection	Deformable Parts Model combining HOG features with latent SVM for objects with deformable structure
Shi & Malik [36]	2000	Classical Segmentation	Normalized cuts segmentation using graph partitioning for coherent region extraction
Boykov & Jolly [18]	2001	Classical Segmentation	Graph-cut 986ptimization for interactive binary segmentation of objects
Shotton et al. [19]	2009	Classical Segmentation	TextonBoost combining textons and boosting with CRF for multi-class segmentation
Girshick et al. [5]	2014	Deep Detection	R-CNN: region proposals fed into CNN for object detection; enabled deep detectors
Ren et al. [6]	2015	Deep Detection	Faster R-CNN: integrated region proposal network for real-time two-stage detection
Redmon et al. [20]	2016	Deep Detection	YOLO: single-stage detector achieving high frame rates suitable for real-time robotics
Liu et al. [21]	2016	Deep Detection	SSD: single-shot multibox detector combining multi-scale anchors with CNN features
He et al. [22]	2017	Deep Segmentation/Detection	Mask R-CNN: extends Faster R-CNN to predict instance masks in parallel with boxes
Long et al. [7]	2015	Deep Segmentation	FCN: transforms classification CNNs into fully convolutional networks for semantic segmentation
Badrinarayanan et al. [23]	2017	Deep Segmentation	SegNet: encoder–decoder network with pooling indices for efficient upsampling
Ronneberger et al. [24]	2015	Deep Segmentation	U-Net: skip-connected encoder–decoder for fine-grained segmentation with limited data
Zhao et al. [25]	2017	Deep Segmentation	PSPNet: pyramid scene parsing network capturing context at multiple scales
Chen et al. [26]	2018	Deep Segmentation	DeepLabv3+: atrous convolution and encoder–decoder for state-of-the-art segmentation
Armeni et al. [30]	2019	Scene Graph	3D scene graph combining semantic labels, geometry and hierarchy for holistic reasoning
Kim et al. [31]	2020	Scene Graph	Sparse semantic 3D scene graph tailored for robotic navigation and task planning
Dosovitskiy et al. [10]	2021	Transformer	Vision Transformer: applies self-attention to image patches, providing global context
Carion et al. [11]	2020	Transformer	DETR: end-to-end object detection transformer predicting sets of objects without anchors

V. CONCLUSION

This survey has discussed the history of computer vision algorithms for scene understanding in robotics, ranging from classical pipelines that are based on hand-crafted features and geometric interpretation to state-of-the-art approaches based on deep learning methods driven by large-scale databases and neural networks. We then introduced a taxonomy of methods by paradigm and task, reviewed landmark solutions for mapping, detection, semantic segmentation and holistic scene representation tasks, outlined benchmark datasets and contrasted classical versus deep approaches. Deep learning has opened the door to new possibilities in robotic perception, as robots are now able to perceive and segment complex scenes with human-like or near-human accuracy. Meanwhile, its classical counterparts retain indispensable roles in accurate geometric estimation and explainable behavior. In addition, future work should look into hybrid models fusing learning and structure, self-supervised learning saving on annotation cost, and safe adaptive systems for the open world. Only by coupling perception, reasoning and action can robotics move toward truly intelligent agents with robust scene understand-ing.

REFERENCES

[1] A. J. Davison, "Real-time simultaneous localisation and map-ping with a single camera," in Proc. IEEE International Con-ference on Computer Vision, 2003, pp. 1403–1410.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[3] H. Bay, T. Tuytelaars and L. Van Gool, "SURF: speeded up robust features," in Proc. European Conference on Computer Vision, 2006, pp. 404–417.

[4] J. Deng, W. Dong, R. Socher, L. Jia Li, K. Li and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[5] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.

[6] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in Proc. Advances in Neural Information Processing Systems, 2015, pp. 91–99.

[7] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[8] S. Grigorescu, B. Trasnea, T. Cocias and G. Macesanu, "A survey of deep learning techniques for autonomous driving," Journal of Field Robotics, vol. 37, no. 3, pp. 362–386, 2020.

[9] A. I. Ka´roly, P. Galambos, J. Kuti and I. J. Rudas, "Deep learning in robotics: survey on model structures and training strategies," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 51, no. 1, pp. 266–279, 2021.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An image is worth 16×16 words: transformers for image recognition at scale," in Proc. International Conference on Learning Representations, 2021.

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, "End-to-end object detection with transform-ers," in Proc. European Conference on Computer Vision, 2020, pp. 213–229.

[12] C. Chen, B. Wang, C. X. Lu, N. Trigoni and A. Markham, "Deep learning for visual localization and mapping: a survey," IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 12, pp. 17000–17019, 2024.

[13] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid and J. J. Leonard, "Past, present, and fu-ture of simultaneous localization and mapping: toward the robust-perception age," IEEE Transactions on Robotics, vol. 32, no. 6, pp. 1309–1332, 2016.

[14] Y. Tang, C. Zhao, J. Wang, C. Zhang, Q. Sun, W. Zheng, W. Du, F. Qian and J. Kurths, "Perception and navigation in autonomous systems in the era of learning: a survey," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 11, pp. 9604–9624, 2022.

[15] J. Ni, Y. Chen, G. Tang, J. Shi, W. Cao and P. Shi, "Deep learning-based scene understanding for autonomous robots: a survey," Intelligence & Robotics, vol. 3, no. 3, pp. 374–401, 2023.

[16] R. Mur-Artal and J. D. Tardo´s, "ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras," IEEE Transactions on Robotics, vol. 33, no. 5, pp. 1255–1262, 2017.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.

[18] Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in ND images," in Proc. IEEE International Conference on Computer Vision, 2001, pp. 105–112.

[19] J. Shotton, J. Winn, C. Rother and A. Criminisi, "TextonBoost for image understanding: multi-class object recognition and segmentation by jointly modelling texture, layout, and context," International Journal of Computer Vision, vol. 81, no. 1, pp. 2–23, 2009.

[20] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: unified, real-time object detection," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A. C. Berg, "SSD: single shot multibox detector," in Proc. European Conference on Computer Vision, 2016, pp. 21–37.

[22] K. He, G. Gkioxari, P. Dolla´r and R. Girshick, "Mask R-CNN," in Proc. IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[23] V. Badrinarayanan, A. Kendall and R. Cipolla, "SegNet: a deep convolutional encoder–decoder architecture for image segmen-tation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481–2495, 2017.

[24] O. Ronneberger, P. Fischer and T. Brox, "U-Net: convolu-tional networks for biomedical image segmentation," in Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015, pp. 234–241.

[25] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, "Pyramid scene parsing network," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.

[26] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff and H. Adam, "Encoder–decoder with atrous separable convolution for se-mantic image segmentation," in Proc. European Conference on Computer Vision, 2018, pp. 833–851.

[27] J. Engel, T. Scho¨ps and D. Cremers, "LSD-SLAM: large-scale direct monocular SLAM," in Proc. European Conference on Computer Vision, 2014, pp. 834–849.

[28] R. A. Newcombe, S. J. Lovegrove and A. J. Davison, "DTAM: dense tracking and mapping in real time," in Proc. IEEE International Conference on Computer Vision, 2011, pp. 2320–2327.

[29] P. F. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1627–1645, 2010.

[30] I. Armeni, Z.-Y. He, J. Gwak, A. R. Zamir, M. Fischer, J. Malik, and S. Savarese, "3D scene graph: a structure for unified semantics, 3D space, and camera," in Proc. IEEE International Conference on Computer Vision, 2019, pp. 5664–5673.

[31] U. Kim, J. Park, T. Song, and J.-H. Kim, "3-D scene graph: a sparse and semantic representation of physical environments for intelligent agents," IEEE Transactions on Cybernetics, vol. 50, no. 12, pp. 4921–4933, 2020.

[32] M. Sualeh and G. W. Kim, "Simultaneous localization and mapping in the epoch of semantics: a survey," International Journal of Control, Automation and Systems, vol. 17, no. 3, pp. 729–742, 2019. IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3354–3361.

[38] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes dataset for semantic urban scene understanding," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.

[39] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn and A. Zisserman, "The Pascal visual object classes (VOC) challenge," International Journal of Computer Vision, vol. 88, no. 2, pp. 303–338, 2010.

[40] T. Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dolla´r and C. L. Zitnick, "Microsoft COCO: common objects in context," in Proc. European Conference on Computer Vision, 2014, pp. 740–755.

[41] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGB-D images," in Proc. European Conference on Computer Vision, 2012, pp. 746–760.

[42] S. Song, S. Lichtenberg, and J. Xiao, "SUN RGB-D: a RGB-D scene understanding benchmark suite," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 567–576.

[43] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: richly annotated 3D reconstructions of indoor scenes," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2432–2443.

[44] A. Zeng, K. T. Yu, S. Song, D. Suo, E. Walker, A. Rodriguez, and J. Xiao, "Multi-view self-supervised deep learning for 6D pose estimation in the Amazon Picking Challenge," in Proc. IEEE International Conference on Robotics and Automation, 2017, pp. 1386–1393.

[45] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," Neural Computation, vol. 1, no. 4, pp. 541–551, 1989.

[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[47] S. Lowry, N. Su¨nderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke and M. Milford, "Visual place recognition: a survey," IEEE Transactions on Robotics, vol. 32, no. 1, pp. 1–19, 2016.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. International Conference on Learning Representations, 2015.

[49] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[50] S. Ghosh, N. Das, I. Das and U. Maulik, "Understanding deep learning techniques for image segmentation," ACM Computing Surveys, vol. 52, no. 4, p. 73, 2019.

[33] R. Mascaro´ and M. Chli, "Scene representations for robotic spatial perception," Annual Review of Control, Robotics, and Autonomous Systems, vol. 8, pp. 351–377, 2025.

[34] C. Harris and M. Stephens, "A combined corner and edge detector," in Proc. Alvey Vision Conference, 1988, pp. 147–151.

[35] P. Viola and M. J. Jones, "Robust real-time face detection," International Journal of Computer Vision, vol. 57, no. 2, pp. 137–154, 2004.

[36] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888–905, 2000.

[37] A. Geiger, P. Lenz and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in Proc.