Ni *et al*. *Intell Robot* 2023;3(3):374-401
DOI: 10.20517/ir.2023.22

Intelligence & Robotics

Review

Open Access

Deep learning-based scene understanding for autonomous robots: a survey

Jianjun Ni[1,2], Yan Chen[1,2], Guangyi Tang[1], Jiamei Shi[2], Weidong Cao[1,2], Pengfei Shi[1,2]

[1]School of Artificial Intelligence and Automation, Hohai University, Changzhou 213022, Jiangsu, China.
[2]College of Information Science and Engineering, Hohai University, Changzhou 213022, Jiangsu, China.

Correspondence to: Prof. Jianjun Ni, School of Artificial Intelligence and Automation, Hohai University, No.200, North Jinling Road, Xinbei District, Changzhou 213022, Jiangsu, China. E-mail: njjhhuc@gmail.com; ORCID: 0000-0002-7130-8331

How to cite this article: Ni J, Chen Y, Tang G, Shi J, Cao W, Shi P. Deep learning-based scene understanding for autonomous robots: a survey. *Intell Robot* 2023;3(3):374-401. http://dx.doi.org/10.20517/ir.2023.22

Received: 25 Apr 2023 **First Decision:** 5 Jul 2023 **Revised:** 15 Jul 2023 **Accepted:** 4 Aug 2023 **Published:** 15 Aug 2023

Academic Editor: Simon X. Yang, Hongtian Chen **Copy Editor:** Yanbin Bai **Production Editor:** Yanbin Bai

Abstract

Autonomous robots are a hot research subject within the fields of science and technology, which has a big impact on social-economic development. The ability of the autonomous robot to perceive and understand its working environment is the basis for solving more complicated issues. In recent years, an increasing number of artificial intelligence-based methods have been proposed in the field of scene understanding for autonomous robots, and deep learning is one of the current key areas in this field. Outstanding gains have been attained in the field of scene understanding for autonomous robots based on deep learning. Thus, this paper presents a review of recent research on the deep learning-based scene understanding for autonomous robots. This survey provides a detailed overview of the evolution of robotic scene understanding and summarizes the applications of deep learning methods in scene understanding for autonomous robots. In addition, the key issues in autonomous robot scene understanding are analyzed, such as pose estimation, saliency prediction, semantic segmentation, and object detection. Then, some representative deep learning-based solutions for these issues are summarized. Finally, future challenges in the field of the scene understanding for autonomous robots are discussed.

Keywords: Autonomous robots, scene understanding, deep learning, object detection, pose estimation

1. INTRODUCTION

In recent years, science and technology have developed rapidly, and the applications of autonomous robots become increasingly extensive [1–3]. With the development of the technologies, the tasks for autonomous robots have become more complicated and challenging. To complete these tasks, one of the main requirements for autonomous robots is the strong capability of the robot to effectively perceive and understand the complicated three-dimensional (3D) environment in which it is positioned.

The ability of an autonomous robot to perceive and understand its own environment, akin to human perception, serves as the foundation for further autonomous interaction with the environment and human users. This problem is also a prominent topic in the field of computer vision, which has made great progress, and lots of research findings have been used for practical applications of autonomous robots. Many research findings in this field are based on two-dimensional (2D) images. However, the real world is a 3D environment, and there remains ample room for future research on the perception and understanding of 3D environments. The environment perception is the basis of scene understanding, which can provide stable and accurate information for scene understanding. On the other hand, scene understanding can provide richer and higher-level information for environment perception. In this paper, we will mainly discuss the scene understanding problems.

There are lots of research results in this field. Nevertheless, a significant portion of current research is focused on more idealized situations. However, the real world is a complicated scene with a number of issues that affect the accuracy of environmental perception and understanding, such as image interference, clutter occlusion, etc. Consequently, it is crucial to study the essential technologies that enable autonomous robots to perceive and comprehend their environment within complex 3D space, addressing both theoretical underpinnings and practical implementation.

This paper provides a survey on the deep learning-based scene understanding for autonomous robots. We provide a brief overview of the research methodologies used to study the perception and comprehension of the robotic environment, and then we concentrate on deep learning-based approaches to these issues. Other relevant surveys in the field of deep learning-based scene understanding can be used as supplements to this paper (see e.g., [4,5] and [6,7]). The main differences between this paper and other surveys lie in its function as an overview of the state-of-the-art approaches in this field, owing to the continuous emergence of new approaches driven by the rapid development of deep learning-based scene understanding. In addition, this paper provides a selection of the latest related works from our research group.

The main contributions of this paper are summarized as follows: (1) The advancement of scene understanding for autonomous robots is thoroughly analyzed and reviewed; (2) A survey on the applications of deep learning methods in scene understanding for autonomous robots is given out; and (3) Some representative deep learning-based methods in the field of autonomous robot scene understanding are analyzed. At last, some possible future study directions in this field are discussed.

This paper is organized as follows. Section 2 provides a summary of the development of autonomous robots and their ability to perceive and comprehend their environment. In Section 3, the key issues of the scene understanding for autonomous robots are analyzed. Additionally, select representative deep learning-based methods based on deep learning techniques in the field of scene understanding are outlined and analyzed. The potential study directions of deep learning-based perception and comprehension of the environment for autonomous robots are given out in Section 4. Finally, conclusions are given out in Section 5.

2. BACKGROUND AND SIGNIFICANCE OF THE SCENE UNDERSTANDING

The global economy has witnessed rapid growth in recent years, paralleled by swift advancements in science and technology. The applications of robots are becoming more and more popular[8]. Autonomous robots are the representative of advanced technologies, which are the integration of the robotics, information technology, communication technology, and artificial intelligence. These robots have been more integrated into human society, not only creating huge economic benefits for society but also effectively improving individual living standards[9].

The autonomous robot industry is an important standard to evaluate the innovation and high-end manufacturing level of a country. The development of the autonomous robot has attracted growing attention from countries all over the world. A number of famous research institutions and companies across the globe have focused on the realm of autonomous robots.

The representative robotics research institutions include the Robotics and Mechatronics Center (RMC) of the German Aerospace Center, the Computer Science and Artificial Intelligence Laboratory (CSAIL) of Massachusetts Institute of Technology, the Humanoid Robotics Institute (HRI) of Waseda University, Shenyang Institute of Automation Chinese Academy of Sciences, the Robotics Institute of Shanghai Jiaotong University, and so on. There are lots of representative robotic enterprises, such as ABB (Switzerland), KUKA Robotics (Germany), Yaskawa Electric Corporation (Japan), iRobot (USA), AB Precision (UK), Saab Seaeye (Sweden), SIASUN (China), etc[10].

Due to the current technical limitations, the functions of common autonomous robots in daily life are still relatively simple. For example, the serving robot [see Figure 1A] and the sweeping robot [see Figure 1B] can only complete some simple tasks, such as moving according to the planned trajectory to the designated position. The expansion of the robot application range requires that the functions of robots are no longer limited to mechanized or programmed operations, narrow human-computer interactions, etc. There is an increasing need for autonomous robots to carry out more difficult tasks. Robots are anticipated to be able to do complicated tasks, such as picking up and dropping off goods or even operating tools autonomously by sensing their surroundings. Empowering autonomous robots with ample environmental perception and a comprehensive understanding of their intricate 3D surroundings stands as an essential prerequisite to satisfy the requirements for these more difficult jobs. For example, the logistics robot can make mobility control decisions after it can autonomously perceive and understand the traffic and road environment [see Figure 1C]. To operate effectively and securely in the unknown and complex underwater environment, the underwater search robot must be aware of its surroundings [see Figure 1D].

When an autonomous robot conducts a task in a complicated environment, it must first determine its current position and estimate its displacement pose change through a visual Simultaneous Localization and Mapping (SLAM) system. The robot also needs to assess the shape of the environment and comprehend the range of its surroundings. In addition, it is of utmost practical importance to research room layout estimation in complex cluttered environments. Next, the autonomous robot should perform the saliency detection, namely directing its attention toward the regions of interest, akin to human behavior. This is followed by target detection, a crucial step in identifying manipulable items and their locations within the environment. Notably, the study of functional availability detection of objects in 3D space is fundamentally important for robots to further perform complex operational tasks because autonomous robots need to understand the functional availability and even the usage of each part of the object to be interacted with. This facet is closely related to the 3D structure of the object. The main tasks of the scene understanding for the autonomous robot in a complicated environment are shown in Figure 2.

All of these tasks introduced above are the research topics in the scene understanding of autonomous robots.

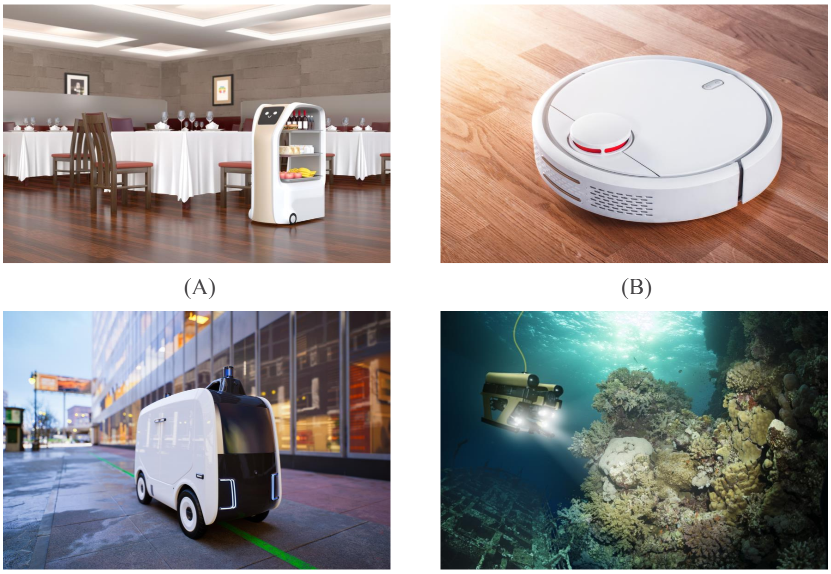

Figure 1. Applications of scene understanding for autonomous robots: (A) Service robots; (B) Sweeping robots; (C) Logistics robots; (D) Underwater search robots.

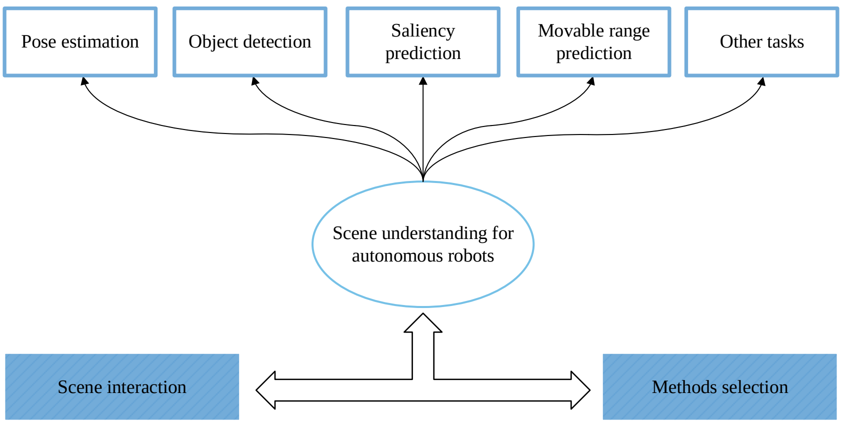

Figure 2. The main tasks of the scene understanding for the autonomous robot.

In a word, scene understanding of autonomous robots is to analyze a scene by considering the geometric and semantic context of its contents and the intrinsic relationships between them. The process mainly involves matching signal information from sensors observing the scene with a model that humans use to understand the scene. On this basis, scene understanding is the semantic extraction and addition of sensor data, which is used to describe the scene for autonomous robots.

In the early research of scene understanding, parts-based representations for object description and scene understanding were the mainstream methods. In these methods, the basic information and hidden deeper information of images are reflected by extracting the low-level and middle-level visual features. And these early methods often realize semantic classification through feature modeling. There are many traditional feature representation methods that have been used widely in scene understanding. Scale-Invariant Feature Transform (SIFT)[11,12] has rotation, scale, and affine invariant qualities. It has an excellent classification effect even for images with huge scale changes. GIST[13] is an image global description feature based on fusing contextual data and a spatial envelope model, which can extract the spatial structural data of pictures using the energy spectrum. LSA (Latent Semantic Analysis)[14] is used to address the issue of many words with a single meaning and multiple meanings of a word in text analysis. Other good manually designed features include Speeded Up Robust Features (SURF)[15], Histogram of Oriented Gradient (HOG)[16,17], and so on. Based on these features, a number of traditional image scene semantic classification techniques were developed. For example, Vailaya *et al.*[18] classified scene images using dual features of color moments and texture and derived global features of images in a Bayesian framework. Li *et al.*[19] presented the target library-based image classification technique by decomposing an image into a number of objects and identifying the semantics of each object to realize the semantic classification of images.

With the rapid and substantial growth of hardware computing power, deep learning methods have gained rapid development[20–22]. Data-driven methods, especially those based on deep neural networks, have been proven to have outstanding advantages in feature learning and visual data description. The scene understanding of autonomous robots based on deep learning has been developed rapidly. Compared to traditional scene understanding methods, the methods based on deep neural networks can more flexibly use the adaptively extracted features to perform tasks such as object detection, semantic segmentation, and more. As a result, they achieve far better performance.

Environment perception and understanding of autonomous robots in complex 3D scenes is a hot topic both in computer vision and the robotic field. However, there are some differences between normal computer vision and the scene understanding for autonomous robots. Firstly, normal computer vision usually obtains data from static images or videos for analysis and pays more attention to the detection, recognition, and positioning of objects. Scene understanding for autonomous robots usually requires the combination of multiple sensor data and needs to consider the dynamic changes in the environment and the 3D perception and understanding of the environment in order to carry out tasks such as path planning and obstacle avoidance. Furthermore, this process often entails interactions with both the environment and individuals, leading to decision-making based on the interaction output. In contrast, the normal computer vision does not require such interactivity.

A lot of challenging, realistic issues still need to be resolved, and various methods have been used in this field, such as traditional image processing methods, traditional artificial intelligence methods, and so on. Among these methods, deep learning-based methods have achieved great success in this field for their distinct advantages, such as high accuracy, strong robustness, and low cost. This paper will focus on the deep learning-based methods used in the field of scene understanding for autonomous robots.

3. DEEP LEARNING FOR SCENE UNDERSTANDING

Deep neural networks, which serve as the foundational network for image classification, target recognition, image segmentation, target tracking, and video analysis, are used in the deep learning-based vision system. The network parameters are trained through big data, and the feature extraction and classification are realized end-to-end, avoiding complex feature engineering design. Deep learning-based methods have strong feature representation capabilities that can be used to transform the original image into low-level spatial features, middle-level semantic features, and high-level target features. Then, through feature combination, classification and prediction tasks can be achieved efficiently. In addition, learning-based methods have strong generality and make it simpler to complete multi-task learning and multi-modal learning tasks that incorporate video, text, and speech. This helps to advance the development of scene understanding for autonomous robots.

As introduced in Section 2, lots of issues should be solved in the field of scene understanding for autonomous robots. In this section, the detailed applications grounded in various deep learning methods will be introduced. The main applications for scene understanding based on deep learning summarized here stem from the ex-

	3D object detection models	Pose estimation models	Semantic segmentation models	Saliency prediction models	Other application models
2018	SECOND F-pointNet	V2V-PoseNet	DenseASPP EncNet	ASNet	HRRS POL-SAR
2019	F-ConvNet Fast Point R-CNN	CDPN NOCS	DANet APCNet CANet	RGB-D-SOD AF-RGB-D	MIL
2020	SA-SSD TANet TGNet	DPVL G2L-Net PVN3D	EffcientFCN	CMP-SOI DevsNet	STFN Cam-Net RSSM-Net
2021	CenterPoint Part-A2	FFB6D	FuseSeg MaskFormer FANet	AMDFNet SSPNet STA3D	MSML
2022	RGBNet BADet	ROFT Voting and Attention Epro-PnP	FusionLane	ECANet TranSalNet	DA2Net
2023	DCLM		BCINet		MFGNet SAGN

Figure 3. Some deep learning-based models in the field of scene understanding in recent years.

tensive awareness of the authors, which can demonstrate the key issues and the latest advances in the field of scene understanding. The main applications of deep learning in scene understanding include object detection, semantic segmentation, pose estimation, and so on. These applications will be introduced in detail as follows. Figure 3 shows these deep learning-based models according to the time they are published. To describe easily without loss of generality, we do not distinguish the applications between normal computer vision and scene understanding for autonomous robots in this paper.

As we know, the datasets are very important for the scene understanding based on deep learning methods. Lots of works of literature have introduced various datasets in different tasks of scene understanding. So, before introducing the main applications of deep learning in this field, the most used datasets in the field of scene understanding are summarized and shown in Table 1.

3.1. 3D object detection

Object detection is an image segmentation based on geometric and statistical features of the object. It combines object segmentation and recognition into one task, with the aim of determining the location and class of object appearances. Currently, 2D object detection has been relatively mature, especially with the emergence of Faster Regions with convolutional neural network Features (Faster RCNN), which has brought it to an unprecedented boom. For example, in the previous work of our research group [31], a deep neural network-based SSD framework is proposed to improve the feature representation capability of feature extraction networks. However, in the application scenarios of driverless, robotics, and augmented reality, 2D object detection can only provide the confidence of the position and corresponding category of the object in a 2D image (see Figure 4), while the general 2D object detection cannot provide all the information needed for perceiving the environment.

In the real world, objects have 3D shapes, and most applications require information about the length, width,

Table 1. The most used datasets of deep learning in the field of scene understanding

KITTI [23]	The KITTI dataset contains 7,481 training samples and 7,518 test samples divided into three categories (i.e., Car, Pedestrian, and Cyclist). In addition, it is divided into three difficulty levels based on the scale, occlusion, and truncation levels of the objects in the context of autonomous driving (i.e., Easy, Moderate, and Hard)
nuScenes [24]	The nuScenes dataset consists of 1000 challenging driving video sequences, each about 20 seconds long, with 30k points per frame. It has 700, 150, and 150 annotated sequences for training, evaluation, and test segmentation, respectively
LINEMOD [25]	It is a dataset widely used for 6D object pose estimation. There are 13 objects in this dataset. For each object, there are about 1100-1300 images with annotations and only one object with annotation per image
FAST-YCB [26]	It consists of six realistic synthetic sequences, each containing the fast motion of a single object from the YCB model set in the desktop scene. Each sequence is rendered in bright static lighting conditions and provides 1280×720 RGB-D frames with accurate ground truth of 6D object pose and velocity
PASCAL VOC 2012 [27]	It is a benchmark dataset that initially contains 1464 images for training, 1449 for validation, and 1456 for testing. In the original PASCAL VOC 2012 dataset, there are a total of 20 foreground object classes and one background class
Cityscapes [28]	The dataset has 5,000 images captured from 50 different cities. Each image has 2048×10244 pixels, which have high-quality pixel-level labels of 19 semantic classes
DHF1K [29]	It contains the most common and diverse scenarios, with 1000 video samples and no publicly available ground-truth annotations. Only the first 700 annotated maps and videos are available in the DHF1K dataset, and the remaining 300 annotations are reserved for benchmarking
VOT-2017 [30]	The VOT-2017 dataset can be used for target tracking of different tasks and contains 60 short sequences labeled with six different attributes

(A) (B)

Figure 4. 2D object detection visualization: (A) in the bedroom; (B) in the kitchen [31].

height, and also the deflection angle of the target object. Therefore, research on methods related to 3D target detection is needed. In scene understanding for autonomous robots, object detection is a critical task to understand the position and class of the objects with which they interact. In real 3D complex scenes, the background information is very rich; therefore, object detection techniques can be used to understand the location and category of interactable objects by giving a 3D rectangular location candidate box and categorizing them according to their attribution possibilities.

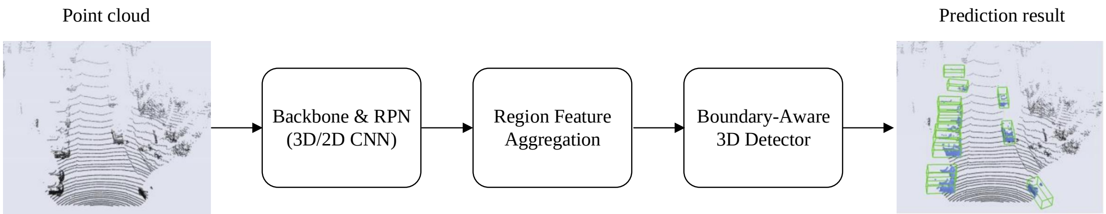

Figure 5. The network overview of the BADet, where RPN denotes the region proposal network [32].

3D object detection based on deep learning is a hot research topic in the field of environment perception and understanding. In the deep learning-based model, during the process of making the proposals in regional proposal networks in a bottom-up manner, the resulting proposals somehow deviate from the ground truth and appear densely in local communities. Due to the lack of a corresponding information compensation mechanism, the proposals generated by the general regional proposal networks give up a large amount of boundary information. To deal with this problem, Qian *et al.*[32] proposed BADet, a 3D object detection model from point clouds, which can efficiently model the local boundary correlations of objects through local neighborhood graphs and significantly facilitate the complete boundaries of each individual proposal.

BADet consists of three key components, namely, a backbone and region generation network, a region feature aggregation module, and a boundary-aware graph neural network. Its network overview is shown in Figure 5.

In the backbone and region proposal network (RPN) of BADet, the original point cloud is voxelized into a volume mesh for multi-scale semantic feature abstraction and 3D proposal generation with the help of the backbone and a series of 3D sparse convolutions. Specifically, let p be a point in a raw point cloud P with 3D coordinates (p_x, p_y, p_z) and reflectance intensities p_r, then

$$P = \left\{p^i = (p^i_x, p^i_y, p^i_z, p^i_r) \in \mathfrak{R}^4, i = 1, 2, \cdots, N\right\} \tag{1}$$

where N indicates the number of points within P. Let $[v_L, v_W, v_H] \in \mathfrak{R}^3$ be the quantization step, then the voxelized coordinates of p can be obtained, namely

$$V_p = \left(\left\lfloor \frac{p_x}{v_L} \right\rfloor, \left\lfloor \frac{p_y}{v_W} \right\rfloor, \left\lfloor \frac{p_z}{v_H} \right\rfloor\right) \tag{2}$$

where $\lfloor \cdot \rfloor$ is the floor function. Therefore, the point cloud P can be positioned into a feature map with a resolution of $L \times W \times H$, subject to the quantization step $[v_L, v_W, v_H]$.

In the region feature aggregation module of BADet, multi-level semantic features are leveraged to obtain more informative RoI-wise representations. In the boundary-aware graph neural network, neighboring 3D proposals are used as inputs for graph construction within a given cutoff distance. Specifically, the local neighborhood graph $G(V, E)$ can be constructed as

$$E = \left\{(i, j) \,\middle|\, \left\|x_i - x_j\right\|_2 < r\right\} \tag{3}$$

where V and E are the nodes and edges, respectively; r is the threshold; and x_i denotes the 3D coordinates of a node of graph G.

In[32], an overall loss L is used, namely

$$L = L_{rpn} + L_{gnn} + L_{offset} + L_{seg} \tag{4}$$

Table 2. The results of BADet on KITTI test server and nuScenes dataset [32]

	KITTI test server			nuScenes dataset	
	Easy	Moderate	Hard		
$AP_{3D}(\%)$	89.28	81.61	76.58	$mAP(\%)$	47.65
$AP_{BEV}(\%)$	95.23	91.32	86.48	$NDS(\%)$	58.84

AP_{3D} and AP_{BEV} mean the Average Precision (AP) with 40 recall positions on both BEV (Bird's Eye View) and 3D object detection leaderboard; mAP and NDS denote the mean Average Precision and nuScenes detection score, respectively.

where L_{rpn} and L_{gnn} are Focal Loss and Smooth-L1 Loss for the bounding box classification and regression, respectively; L_{offset} is the center offset estimation loss, which is used to obtain better boundary-aware voxelwise representations, and L_{seg} is the foreground segmentation loss.

To evaluate the performance of BADet, some comparison experiments are conducted on the KITTI and nuScenes datasets. The results of BADet are listed in Table 2.

The results in [32] show that BADet outperforms all its competitors with remarkable margins on KITTI BEV detection leaderboard and ranks 1st in "Car" category of moderate difficulty.

3D object detection methods have developed rapidly with the development of deep learning techniques. In recent years, many scholars have been exploring new results in this field. For example, Shi *et al.* [33] proposed the Part-A2 net to implement the 3D object detection using only LiDAR point cloud data. Li *et al.* [34] proposed the TGNet, a new graph convolution structure, that can effectively learn expressive and compositional local geometric features from point clouds.

According to the type of input data, 3D object detection can be divided into single-modal methods and multi-modal methods. Single-modal 3D object detection refers to the use of data collected by one kind of sensor as input. The advantage of single-modal 3D target detection is that the input data are simple and the processing flow is clear; the disadvantage is that the input data may not be sufficient to describe the target information in 3D space. Multi-modal 3D object detection refers to the use of multiple data collected by multiple types of sensors as inputs. The advantage of multi-modal 3D target detection is that the input data are rich, and the complementarity of different modal data can be utilized to improve the accuracy and robustness of the detection. The disadvantage is the complexity of the input data and the need to deal with inconsistencies between different modal data. In the following, a summary of the deep learning-based 3D object detection models presented in the last five years is illustrated in Table 3, where the type of the input data of each method is given out.

3.2. Pose estimation

Pose estimation is a crucial component of autonomous robot technology. The pose estimation task deals with finding the position and orientation of an object with respect to a specific coordinate system. The vision-based pose estimation approaches employ a number of feature extraction techniques to obtain the spatial positional information of the target from the image.

There are two classical methods for pose estimation, namely, the feature-based techniques and the template matching methods. The traditional feature-based technique primarily extracts features from images and creates a relationship between the 2D pixel points and 3D coordinate points in space. The differences in lighting and background complexity have a significant impact on the feature extraction process. In addition, the feature-based methods struggle to handle sparse target texture features. The template matching method can effectively solve the pose estimation problem for the targets with weak texture features in images. However, the accuracy

Table 3. A summary of the deep learning-based 3D object detection models presented in the last five years

Structure	Reference	Input data type	Performances
SECOND	Yan *et al.* (2018) [35]	Single-modal	AP of 83.13% on KITTI test set
F-pointNet	Qi *et al.* (2018) [36]	Multi-modal	AP of 81.20% on KITTI test set
F-ConvNet	Wang *et al.* (2019) [37]	Single-modal	AP of 85.88% on KITTI test set
Fast Point R-CNN	Chen *et al.* (2019) [38]	Single-modal	AP of 84.28% on KITTI test set
SA-SSD	He *et al.* (2020) [39]	Single-modal	AP of 88.75% on KITTI test set
TANet	Liu *et al.* (2020) [40]	Single-modal	3D mAP of 62.00% on KITTI test set
TGNet	Li *et al.* (2020) [34]	Single-modal	$MIoU$ of 68.17% on Paris-Lille-3D datasets
CenterPoint	Yin *et al.* (2021) [41]	Single-modal	mAP of 58.0% on nuScenes test set
Part-A2	Shi *et al.* (2021) [33]	Multi-modal	AP of 85.94% on KITTI test set
RGBNet	Wang *et al.* (2022) [42]	Multi-modal	mAP of 70.2% on ScanNetV2 val set
BADet	Qian *et al.* (2022) [32]	Single-modal	AP of 89.28% on KITTI test set
DCLM	Chen *et al.* (2023) [43]	Multi-modal	mAP of 65.6% on SUN RGB-D dataset

AP means the average precision. $MIoU$: the Mean Intersection over Union.

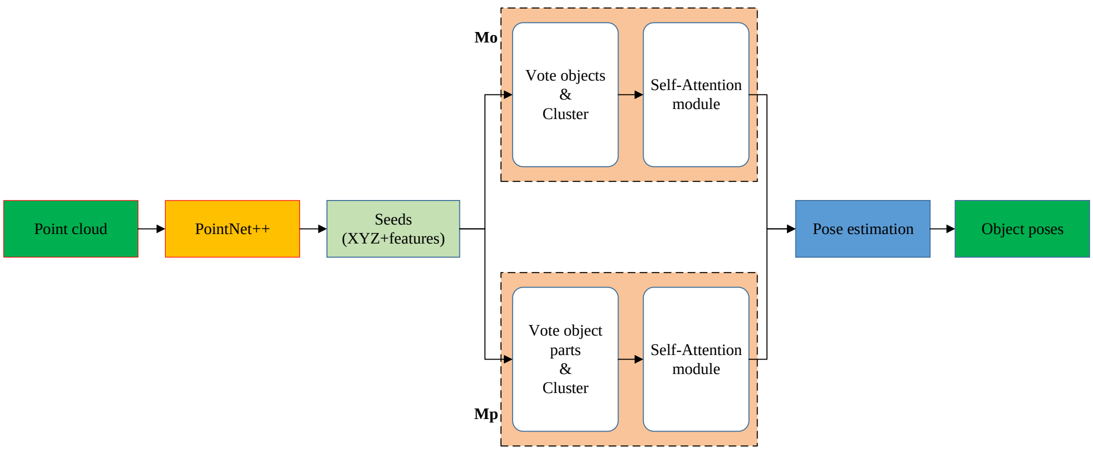

Figure 6. Network architecture of the Voting and Attention-based model [44].

of the template matching method is determined by the number of samples in the template library. While its accuracy improves with the number of the template libraries, it also causes a decrease in problem-solving efficiency, making it unable to meet real-time requirements.

The development of deep learning has influenced pose estimation, and there are numerous study findings in this field. For example, Hoang *et al.* [44] proposed the Voting and Attention-based model, which enhances the accuracy of object pose estimation by learning higher-level characteristics from the dependencies between the individual components of the object and object instances. The structure of this Voting and Attention-based network is shown in Figure 6.

As shown in Figure 6, there are four main parts in the Voting and Attention-based model, namely the feature extraction module based on PointNet++ architecture, part proposals learning module (M_p), object proposals learning module (M_o), and the voting module in both M_p and M_o based on VoteNet.

In the M_p module of the Voting and Attention-based model, the higher-order interactions between the proposed features can be explicitly modeled, which is formulated as non-local operations:

$$H_{part-part} = f\left(\theta\left(H\right)\phi\left(H\right)\right)g\left(H\right). \tag{5}$$

Table 4. The pose estimation results based on the Voting and Attention-based model on nine objects in the Siléane dataset and two objects in the Fraunhofer IPA dataset [44]

Objects	Siléane dataset									Fraunhofer IPA dataset		Mean
	Brick	Bunny	C. stick	C.cup	Gear	Pepper	Tless 20	Tless 22	Tless 29	Gear shaft	Ring screw	
AP	0.48	0.61	0.60	0.52	0.64	0.39	0.44	0.37	0.46	0.65	0.67	0.53

where $\theta(\cdot)$, $\phi(\cdot)$, $andg(\cdot)$ are the learnable transformation on the input feature map H, and $f(\cdot)$ is the encoding function of the relationship between any two parts.

In addition, the compact generalized non-local network (CGNL) [45] is used as the self-attentive module in $\mathrm{M_p}$. Specifically, the CGNL-based self-attentive module takes K clusters $C = (C_1, C_2, ...C_K)$ as input. Then, votes from each cluster are processed by the Multi-Layer Perceptron (MLP) and passed to CGNL. The self-attention mechanism allows features from different clusters to interact with each other and find out who they should pay more attention to.

Similarly, in the $\mathrm{M_o}$ module of the Voting and Attention-based model, the instance-to-instance correlation is modeled. Firstly, K clusters from the high-dimensional features and a set of object centers are generated. Then, CGNL is used to model the rich interdependencies between clusters in feature space. The output is a new feature mapping:

$$H_{obj-obj} = CGNL\left(max\left(MLP\left(v_i\right)\right)\right), i = 1, \ldots, n \tag{6}$$

where v_i is the i-th vote.

Finally, the new feature maps $H_{part-part}$ and $H_{obj-obj}$ are aggregated to the global information by an MLP layer after a max-pooling and concatenation operations.

In the Voting and Attention-based model, a multi-task loss is used for joint learning, namely

$$L = \lambda_1 L_{part-vote} + \lambda_2 L_{object-vote} + \lambda_3 L_{pose} \tag{7}$$

where λ_1, λ_2, and λ_3 are the weights of each task. The losses include voting partial loss $L_{part-vote}$, object voting loss $L_{object-vote}$, and pose loss L_{pose}.

The pose estimation results based on the Voting and Attention-based model on nine objects in the Siléane dataset and two objects in the Fraunhofer IPA dataset are listed in Table 4, and some qualitative results are shown in Figure 7.

The results in Table 4 and Figure 7 show that the Voting and Attention model is very effective in improving the accuracy of the pose estimation, which can obtain an average precision of 53%.

In addition to the above voting-based model, there are lots of research results in pose estimation based on deep learning methods. For example, Chen *et al.* [46] presented a probabilistic PnP (EPro-PnP) model for general end-to-end pose estimation, which is based on the method of locating 3D objects from a single RGB image via Perspective-n-Points (PnP). The EPro-PnP model can realize reliable end-to-end training for a PnP-based object pose estimation network by back-propagating the probability density of the pose to learn the 2D-3D association of the object.

Currently, there are five main types of methods for pose estimation, including feature-based methods, regression-based methods, projection-based methods, representation learning methods, and graph neural network methods. The feature-based method refers to restoring camera pose by establishing feature correspondence between

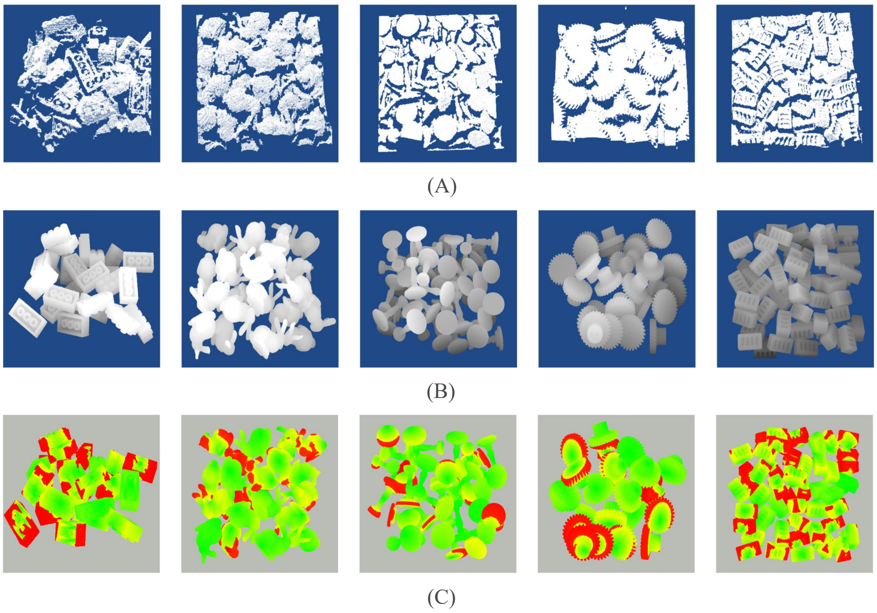

Figure 7. Visualization for pose estimation results based on the Voting and Attention-based model: (a) 3D point cloud input; (b) True values of the poses; (c) Results obtained by the method in [44]. The different color means the visualization of point-wise distance error, ranging from 0 (green) to greater than 0.2 times the diameter of the object (red).

Table 5. A summary of the deep learning-based poss estimation models in the last five years

Structure	Reference	Type of the method	Performances
V2V-PoseNet	Moon *et al*. (2018) [47]	Regression-based	Top1 in the HANDS 2017 frame-based dataset.
CDPN	Li *et al*. (2019) [48]	Feature-based	ADD of 89.86% on the LINEMOD dataset
NOCS	Wang *et al*. (2019) [49]	Projection-based	mAP of 88.4% for 3D IoU on Occluded LINEMOD dataset
DPVL	Yu *et al*. (2020) [50]	Representation learning	Mean ADD of 91.5% on the LINEMOD dataset
G2L-Net	Chen *et al*. (2020) [51]	Graph neural network	Mean ADD of 98.7% on the LINEMOD dataset
PVN3D	He *et al*. (2020) [52]	Projection-based	ADD of 99.4% on the LineMOD dataset
FFB6D	He *et al*. (2021) [53]	Feature-based	Mean ADD of 99.7% on the LINEMOD dataset
ROFT	Piga *et al*. (2022) [26]	Feature-based	$ADD-AUC$ of 76.59% on the FAST-YCB dataset
Voting and Attention	Hoang *et al*. (2022) [44]	Feature-based	AP of 53% on the Siléane dataset and Fraunhofer IPA dataset
EPro-PnP	Chen *et al*. (2022) [46]	Projection-based	ADD of 95.80% on the LineMOD Dataset

ADD means average distance metric. $ADD-AUC$ means area under the curve.

images and scenes. A regression-based method uses a regressor to predict the camera pose. A projection-based method utilizes projection transformation to estimate the pose of a target from an image or video. A representation learning method utilizes deep neural networks to learn high-resolution representations of objects from images or videos, which can improve the accuracy and interpretability of pose estimation. Graph neural network methods use graph neural networks to learn structured representations of objects from images or videos, which can improve robustness of pose estimation. In the following, a summary of the deep learning-based pose estimation models presented in the last five years is illustrated in Table 5, where the type of each method is given out.

3.3. Semantic segmentation

Semantic segmentation is a refined version of image classification. For an image, traditional image classification is to detect and recognize the objects that appear in the image, while semantic segmentation is to classify every pixel point in the image. In the field of autonomous robot environment perception and understanding,

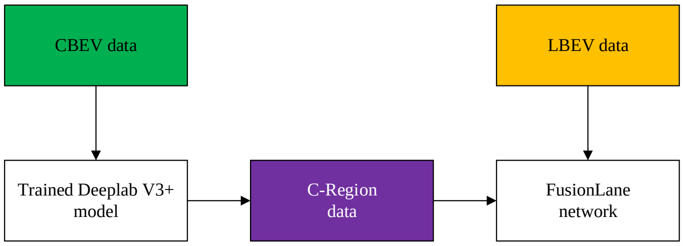

Figure 8. The workflow of the FusionLane model [60]. CBEV denotes the camera bird's eye view data. LBEV denotes the points cloud bird's eye view data. C-Region denotes the obtained semantic segmentation result on CBEV.

semantic segmentation is used to label each pixel in an image with its corresponding semantically meaningful category. Semantic segmentation can help robots recognize and understand surrounding objects and scenes. It is very useful for the semantic segmentation for the robot to find a specific object in the environment. For example, in the field of logistics robotics, semantic segmentation can help the autonomous robots perceive and understand road conditions, traffic signs, pedestrians, and vehicles, which can improve the safety and efficiency of logistics robotics.

The traditional semantic segmentation algorithms are mainly grayscale segmentation, conditional random fields, etc. Grayscale segmentation algorithms recursively segment images into sub-regions until labels can be assigned and then combine adjacent sub-regions with the same labels by merging them. The conditional random field is a type of statistical modeling method for structured prediction.

With the continuous development of deep learning techniques, deep learning has been widely applied in semantic segmentation tasks and achieved impressive results. There are a series of classical deep learning-based models for semantic segmentation, such as Full convolution network (FCN) [54], SegNet [55], DeepLab series [56,57], RefineNet [58], DenseASPP [59], etc. Recently, some improvements have been proposed based on those classical models. For example, Yin *et al.* [60] presented a multi-sensor fusion for lane marking semantic segmentation (FusionLane) based on the DeepLabV3+ network. The workflow of the FusionLane model is shown in Figure 8.

As shown in Figure 8, firstly, the DeepLabV3+ network is used to achieve semantic segmentation on camera BEV (CBEV) data (called as C-Region). Then, the C-Region and LiDAR point cloud BEV (LBEV) data are input into the FusionLane model to realize the lane marking semantic segmentation. Unlike other methods that mainly focus on the analysis of camera images, the semantic segmentation data used in FusionLane is a BEV image converted from the LiDAR point cloud instead of the images captured by the camera to obtain the accurate location information of the segmentation results.

The network contains two data input branches: the camera data and the point cloud data. The data from the two branches need to be preprocessed to meet the network input requirements. For the camera data, the front view is converted into CBEV. In CBEV, one pixel represents an area of $5cm \times 5cm$ in real space. Then, the CBEV image is semantically segmented using the trained DeepLabV3+ network to obtain the C-Region input data. For the point cloud data, it is projected into the 3D BEV with three channels. The values of the three channels are calculated as follows:

$$F(x, y) = \frac{\sum_1^n i}{n} \times 255 \tag{8}$$

$$S(x, y) = \frac{\sum_1^n (h+2)}{n} \times 255 \tag{9}$$

Table 6. Some comparison experiment results of the semantic segmentation on the KITTI dataset [60]

Methods	Background	Solid line	Dotted line	Arrow	Prohibited area	Stop line	Other point	MIoU	PA (%)
DeepLabv3+(LBEV)	0.9419	0.2587	0.2648	0.2793	0.1915	0.3586	0.2770	0.3674	91.31
DeepLabv3+(CBEV)	0.9106	0.6287	0.7012	0.5821	0.6935	0.5294	-	0.6743	85.76
FusionLane	**1.0000**	**0.7477**	**0.7838**	**0.7526**	**0.7979**	**0.9053**	**0.9867**	**0.8535**	**99.92**

IoU: the evaluation metrics include the Intersection over Union on each category; $MIoU$: the Mean Intersection over Union; PA: the Pixel Accuracy.

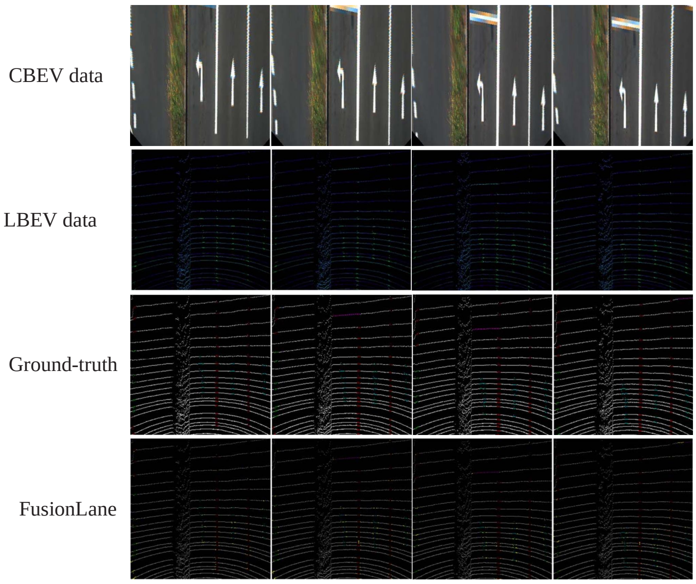

Figure 9. The segmentation results based on the FusionLane network for some scenarios [60].

$$T(x,y) = 255 \times \frac{2}{\pi} \times arctan\sqrt{\frac{\sum_1^n \left(h - \sum_1^n \frac{h}{n}\right)^2}{n}} \tag{10}$$

where $F(x,y)$, $S(x,y)$, and $T(x,y)$ denote the values of the first channel, the second channel, and the third channel, respectively; $i \in [0,1]$ is the reflection intensity value of each point falling within the grid corresponding to the pixel; $h \in [-2,-1]$ is the height value of each laser spot falling within the grid, and $arctan$ is used as the normalization function.

In the FusionLane model of[60], an encoder-decoder network model is proposed, and the LSTM structure is added to the network to assist the semantic segmentation of the lane marking. At last, the KITTI dataset is used to test the performance of the FusionLane model, which is processed and divided into seven categories. The experimental results are listed in Table 6, and some segmentation results based on the FusionLane network are shown in Figure 9.

The results in Table 6 show that DeepLabV3+ has a low IoU for all scenarios except "Background". However, it can be seen that the FusionLane model achieves the best results in all metrics compared to the traditional DeepLabV3+ model. The results in Table 6 and Figure 9 show that relying on a single kind of sensor, whether camera or LiDAR, cannot give sufficiently accurate semantic segmentation results. Effective fusion of data from different sensors can be considered a viable approach to solving the problem.

Table 7. A summary of the deep learning-based models of semantic segmentation in the last five years

Structure	Reference	Network structure	Performances
DenseASPP	Yang *et al*. (2018)[63]	Encoder-decoder	$MIoU$ score of 80.6% on Cityscapes datasets
EncNet	Zhang *et al*. (2018)[64]	Attention mechanism	$MIoU$ score of 85.9% on PASCAL VOC 2012
DANet	Fu *et al*. (2019)[65]	Graph neural network	$MIoU$ score of 81.5% on Cityscapes test set
APCNet	He *et al*. (2019)[66]	Attention mechanism	A new record 84.2% on PASCAL VOC 2012 test set
CANet	Zhang *et al*. (2019)[67]	Attention mechanism	$MIoU$ score of 57.1% on PASCAL-5i test set
EfficientFCN	Liu *et al*. (2020)[68]	GAN	$MIoU$ score of 55.3% on PASCAL Context test set
FuseSeg	Sun *et al*. (2021)[62]	Encoder-decoder	$MIoU$ score of 54.5% on the dataset released in[69]
MaskFormer	Cheng *et al*. (2021)[70]	Transformer learning	$MIoU$ score of 55.6% on the ADE20K dataset
FANet	Hu *et al*. (2021)[61]	Encoder-decoder	$MIoU$ score of 75.5% on Cityscapes test set
FusionLane	Yin *et al*. (2022)[60]	Encoder-decoder	$MIoU$ score of 85.35% on KITTI test set
BCINet	Zhou *et al*. (2023)[71]	Encoder-decoder	$MIoU$ score of 52.95% on the NYUv2 dataset

In addition to the above DeepLab-based model, there are lots of good semantic segmentation models based on deep learning methods. For example, Hu *et al.*[61] presented the FANet model, which is based on an improved self-attention mechanism, to capture the rich spatial context at a small computational cost. Sun *et al.*[62] proposed the FuseSeg model, a new RGB and thermal data fusion network, to achieve superior semantic segmentation performance in urban scenes.

More and more scholars have researched many results in this field. According to the type of network structure, semantic segmentation can be divided into encoder-decoder structure, attention mechanism, graph neural network, generative adversarial network (GAN), and transfer learning. The semantic segmentation method based on encoder-decoder utilizes the encoder-decoder structure to learn and predict the semantic category of each pixel from an image. The method based on GAN uses a generator and a discriminator to conduct confrontation learning. Attention mechanism is a technique that simulates the process of human visual attention. It can calculate the correlation between different positions or channels, give different weights, and highlight the parts of interest while suppressing irrelevant parts. A graph neural network is a deep neural network that can process graph-structured data, which can update the features of nodes and edges through graph convolution operations. Transfer learning is a machine learning technology that can use the knowledge of one domain (source domain) to help the learning of another domain (target domain), thus reducing the dependence on the labeled data of the target domain. A summary of the deep learning-based 3D semantic segmentation models presented in the last five years is illustrated in Table 7, where the type of the network structure of each method is given out.

3.4. Saliency prediction

The human visual system selectively attends to salient parts of a scene and performs a detailed understanding of the most salient regions. The detection of salient regions corresponds to important objects and events in a scene and their mutual relationships. In the field of scene understanding for autonomous robots, the task of the saliency prediction is to mimic the characteristics of human vision to focus on obvious or interested targets by acquiring 3D environment information containing color and depth through sensors. In detail, the saliency prediction needs to identify and segment the most salient objects from the acquired 3D environment information and pay attention to the focal objects.

The traditional saliency prediction problem is commonly known as the task of capturing rare and unique elements from images. Traditionally, salient prediction methods can be classified into three types: (1) Block-based detection models. In this type of method, the linear subspace method is used instead of actual image segmentation, and the significant regions are selected by measuring the feature pair ratio and geometric prop-

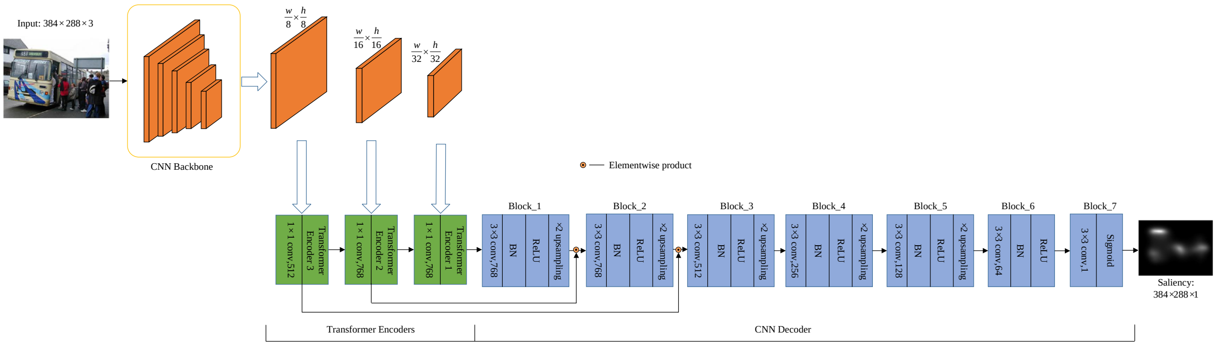

Figure 10. The schematic overview of the TranSalNet network [72].

erties of the region. (2) Region-based detection models. This type of method divides the image into multiple regions, and the saliency of each region is regarded as the sum of the product of its contrast and the weight of all the other regions. (3) Detection model based on external cues of the image. This model utilizes accurate annotations (ground-truth) obtained from the training set, video sequences, similar images, and other sources to make the results more accurate. The performance of the saliency prediction based on similar images will be improved if a large number of data sets are available. In general, the traditional methods use a large amount of saliency a priori information for saliency detection, mainly relying on hand-crafted features. These hand-crafted features have some shortcomings; for example, they may not be able to describe complex image scenes and object structures, cannot adapt to new scenes and objects, and have poor generalization ability. So, the saliency detection based on traditional methods has hit a bottleneck.

Recently, deep learning-based methods have been used widely in various image tasks (e.g., target detection, semantic segmentation, edge detection, etc.), which provide new ideas for saliency prediction and show surprising effect enhancement in some studies. For example, Lou *et al.* [72] proposed the TranSalNet network model. Its basic workflow is shown in Figure 10.

As shown in Figure 10, the convolutional neural network (CNN)-based encoding is used to extract features for saliency prediction. The outputs of the CNN encoding are three sets of multi-scale feature maps with $\frac{w}{8} \times \frac{h}{8}$, $\frac{w}{16} \times \frac{h}{16}$, and $\frac{w}{32} \times \frac{h}{32}$, respectively. Then, these feature maps are input into the transformer encoders to enhance the long-range and contextual information. At last, a CNN decoder is used to fuse the enhanced feature maps from the three transformer encoders. The CNN decoder used in [72] is a full CNN network with seven blocks. The processes from block1 to block6 are as follows:

$$X_i^f = \begin{cases} X_i^c, & i = 1 \\ \mathrm{ReLU}(\mathrm{Upsamle}(\hat{X}_{i-1}^f) \odot X_i^c), & i = 2,3 \\ \mathrm{Upsamle}(\hat{X}_{i-1}^f), & i = 4,5,6 \end{cases} \tag{11}$$

where X_i^f and $\hat{X}_i^f$ are the input and output of the i-th block. The output of the block7 $\hat{y}$ is the predicted saliency map, namely

$$\hat{y} = \mathrm{Sigmoid}(\mathrm{Conv}_{3\times3}(X_6^f)) \tag{12}$$

where $\mathrm{Sigmoid}(\cdot)$ is the sigmoid activation function; $\mathrm{Conv}_{3\times3}$ denotes the 3×3 convolution operation; and X_6^f is the output of the block6.

In the TranSalNet network model, a linear combination of four losses is used as the loss function, namely

$$L = \omega_1 L_{NSS} + \omega_2 L_{KLD} + \omega_3 L_{CC} + \omega_4 L_{SIM} \tag{13}$$

Table 8. Some results of the TranSalNet network model on MIT1003 and CAT2000 datasets [72]

Model name	MIT1003						CAT2000					
	Perception metrics			Non-perception metrics			Perception metrics			Non-perception metrics		
	CC	SIM	NSS	sAUC	AUC	KLD	CC	SIM	NSS	sAUC	AUC	KLD
TranSalNet-Res	0.7595	0.6145	2.8501	0.7546	0.9093	0.7779	0.8786	0.7492	2.4154	0.6054	0.8811	0.5036
TranSalNet-Dense	0.7743	0.6279	2.9214	0.7547	0.9116	0.7862	0.8823	0.7512	2.4290	0.6099	0.8820	0.4715

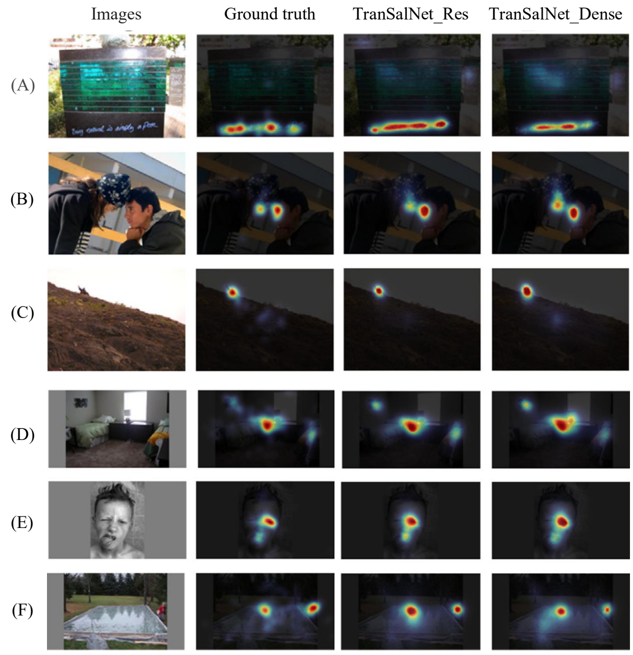

Figure 11. Results of saliency maps generated by TranSalNet_Res and TranSalNet_Dense [72]. The images from (a) to (c) are from the MIT1003 dataset, and the images from (d) to (f) are from the CAT2000 dataset.

where L_{NSS} is the Normalized Scanpath Saliency loss; L_{KLD} is the Kullback–Leibler divergence loss; L_{CC} is the Linear Correlation Coefficient loss; and L_{SIM} is the Similarity loss. ω_1, ω_2, ω_3, and ω_4 are the weights of each loss.

Some results of the TranSalNet network model are listed in Table 8, where TranSalNet_Res and TranSalNet_Dense denote the CNN encoders used in the TranSalNet network, ResNet_50 and DenseNet_161, respectively. Here, two public datasets are as follows: (1) MIT1003 [73]: This dataset contains 300 natural images and eye movement data from 39 observers and is the most influential and widely used dataset in the field of image human eye focus detection. (2) CAT2000 [74]: This dataset includes 4000 images, 200 in each of 20 categories, covering different types of scenes such as cartoon, art, object, low-resolution image, indoor, outdoor, chaotic, random, and line drawings. Some saliency maps generated by the two models are shown in Figure 11.

The results in Table 8 and Figure 11 prove that the TranSalNet architecture presented in [72] is effective in the saliency prediction tasks. In addition, the results in Table 8 and Figure 11 show that the performance of the TranSalNet could be further enhanced by replacing ResNet-50 with DenseNet-161.

In addition to the above TranSalNet model, there are other saliency prediction models based on deep learning, which also have obtained good results in this field. For example, Zou *et al.* [75] proposed the STA3D model, where the S3D network is used as an encoder and the prediction network with spatial dimensional upsam-

Table 9. A summary of the deep learning-based saliency prediction models in the last five years

Structure	Reference	Type of methods	Performances
ASNet	Wang *et al.* (2018) [76]	Gradient-based	MAE scores of 0.072 on the PASCAL-S dataset
RGB-D-SOD	Huang *et al.* (2019) [77]	Perturbation-based	AUC of 0.874 on the NJU400 dataset
AF-RGB-D	*et al.* (2019) [78]	SHAP value-based methods	MAE scores of 0.0462 on the STEREO dataset
CMP-SOI	Zhang *et al.* (2020) [79]	Gradient-based	AUC_J of 0.8839 on the ODI dataset
DevsNet	Fang *et al.* (2020) [80]	Gradient-based	MAE scores of 0.016 on the UVSD dataset
AMDFNet	Li *et al.* (2021) [81]	Gradient-based	MAE scores of 0.019 on the RGBD135 dataset
SSPNet	Lee *et al.* (2021) [82]	Gradient-based	EAO of 0.285 on the VOT-2017 dataset
STA3D	Zou *et al.* (2021) [75]	Gradient-based	AUC_J of 0.927 on the Hollywood2-actions dataset
ECANet	Xue *et al.* (2022) [83]	Attention mechanism-based	AUC_J of 0.903 on the DHF1K dataset
TranSalNet	Lou *et al.* (2022) [72]	Transformer learning-based	AUC of 0.9116 on the MIT1003 dataset

MAE means mean absolute error. AUC_J means the area under the receiver operating characteristic curve. EAO means expected average overlap.

pling and temporal dimensional compression is used as a decoder, to solve the difficulty of video significance prediction in the continuous frame with a fixed offset.

At present, there are five types of methods for saliency prediction, including gradient-based methods, perturbation-based methods, SHAP value-based methods, attention mechanism-based methods, and transfer learning-based methods. The gradient-based method utilizes the gradient information of neural networks to calculate the contribution of each pixel in the input image to the output saliency map. The perturbation-based method evaluates the importance of each pixel by randomly or regularly perturbing the input image. The method based on SHAP values utilizes shapely additive explanations to quantify the impact of each pixel on the output saliency map. The saliency prediction, based on attention mechanisms, utilizes an attention mechanism to simulate the process of human visual attention, thereby improving the accuracy and interpretability of saliency prediction. Transfer learning is used to solve the problem of data shortage and domain differences in saliency prediction, which can improve the generalization ability and adaptability of saliency prediction. A summary of the deep learning-based models in the last five years is illustrated in Table 9, where the type of each method is given out.

3.5. Other applications

In addition to the applications mentioned above, there are many other applications of deep learning methods in autonomous robot environment perception and understanding, such as image enhancement [84,85], visual SLAM [1,86], scene classification [87,88], moving object detection [89,90], and layout estimation [91,92]. In this section, some recent jobs of our group related to this review will be introduced in detail as follows.

3.5.1. Visual SLAM

When a robot enters an unknown environment, vision SLAM technology can be used to solve the problem of the robots about where they are. It estimates the current position, pose, and travel trajectory of the robot in the 3D scene by the changes in the visual data acquired during the robot's travel. In order to implement vision SLAM, there are three main methods: feature-based methods, direct methods, and semi-direct methods.

With feature-based visual SLAM methods, feature points are found and matched. Then, the poses of robots are calculated, and maps are built from geometric relationships. Scalar Transformation (SIFT) [11], Accelerated Robust Feature (SURF) [93], and Fast Rotational Abbreviation (ORB) [94] are the most frequently used feature extraction techniques. The most widely used method for visual SLAM is ORB-SLAM [95,96]. To overcome the problem of high computational complexity in the traditional ORB-SLAM, Fu *et al.* [97] proposed the Fast ORB-SLAM that is light-weight and efficient as it tracks keypoints between adjacent frames without computing descriptors.

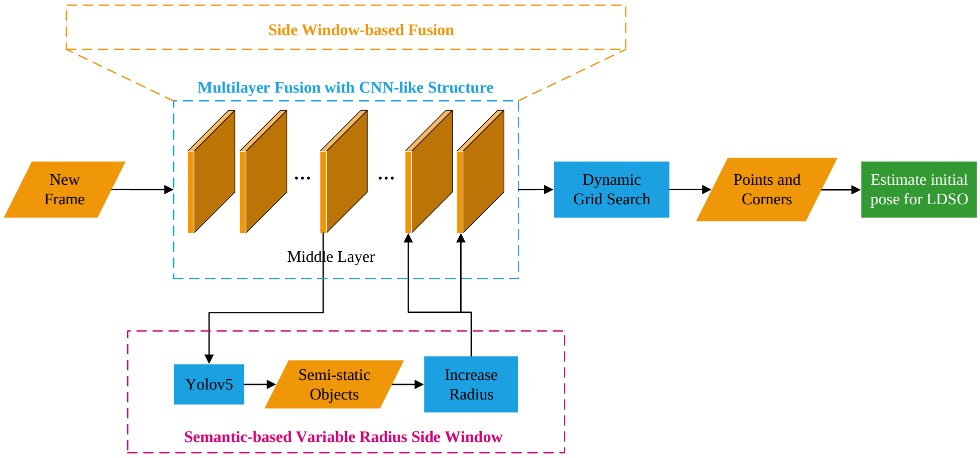

Figure 12. The framework of the improved LDSO method based on the variable radius side window [102].

Direct methods do not rely on one-to-one matching of points. These types of methods minimize the photometric error function of the pixels by extracting pixels with significant gradients and optimizing the interframe pose. The classical direct methods include Large Scale Direct Monocular SLAM (LSD-SLAM) [98], Direct Sparse Range (DSO) [99], etc. Recently, Wang *et al.* [100] introduced a new ceiling-view visual odometry method that introduces plane constraints as additional conditions and achieves better accuracy.

Semi-direct methods, such as SVO [101], employ a similar structure to the feature-based methods, which combine the tracking of the direct method with the motion optimization of feature-based methods. Both feature-based and semi-direct methods rely on highly repeatable low-level geometric feature extractors. Both of them are inappropriate for surfaces with little texture or many repetitive features.

Direct methods, on the other hand, can be applied to a wider variety of scenes. However, compared to feature-based methods, direct methods are less robust. The performance of the direct visual SLAM system under the influence of various camera imaging perturbations will be reduced obviously. To deal with this problem, our group proposed an improved Direct Sparse Odometry with Loop Closure (LDSO) method [102], which is shown in Figure 12.

In the framework of the improved LDSO shown in Figure 12, the region surrounding each pixel is divided into blocks when a new frame is introduced, using the side window approach. Then, a CNN structure is created by this multiple-layer superposition of pixel information fusion [22,103]. The middle layer shows the presence of semi-static items. In the later layers, the radius of the side windows of the pixels belonging to the semi-static objects is increased. Points with an adequate gradient intensity and corners are chosen using dynamic grid searches. The robustness of the system is increased by the addition of points in direct SLAM. To accomplish edge protection, the fusion method is used with a side window mechanism. Finally, to lessen the weight of semi-static objects, the radius of the adjustment side windows is modified in accordance with the semantic information based on a pre-trained Yolov5 model [104].

In the experiments to test the performance of the improved LDSO in [102], two public datasets are used: the KITTI dataset (outdoor datasets) and the TUM RGB-D dataset (indoor datasets). To test the improved LDSO under different camera sensor noises, Gaussian noise and Salt-and-Pepper noise are added to the two datasets. Some results of visual SLAM based on the improved LDSO are shown in Table 10 and Table 11, where $RMSE_{ATE}$ means the root mean squared error of absolute trajectory error. The comparison results on the KITTI dataset with Salt-and-Pepper noise are not given out because the general LDSO is entirely inoperable on the datasets

Table 10. $RMSE_{ATE}$ on the KITTI dataset with Gaussian noise [102]

Method	Gaussian noise											
	KITTI_00	KITTI_01	KITTI_02	KITTI_03	KITTI_04	KITTI_05	KITTI_06	KITTI_07	KITTI_08	KITTI_09	KITTI_10	Average
LDSO [105]	22.543	23.052	169.247	-	-	44.010	58.729	53.481	130.993	-	16.277	64.792
Improved LDSO	17.772	13.023	120.380	2.133	1.093	5.740	13.491	1.973	102.206	52.664	14.042	31.320

'-' means tracking failure. The average value is calculated based on the number of successes.

Table 11. $RMSE_{ATE}$ on the TUM RGB-D dataset [102]

Method	Gaussian noise					Salt-and-pepper noise				
	fr1_xyz	fr2_xyz	fr2_rpy	fr1_desk	fr1_desk2	fr1_xyz	fr2_xyz	fr2_rpy	fr1_desk	fr1_desk2
LDSO [105]	-	0.096	-	0.518	-	-	-	-	0.841	-
Improved LDSO	0.156	0.01	0.06	0.801	0.756	0.129	0.011	0.058	0.796	0.871

'-' means tracking failure.

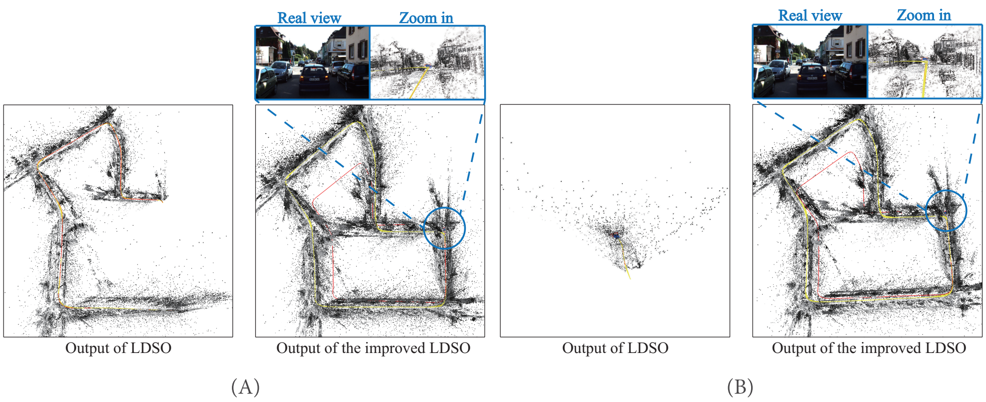

Figure 13. Sample outputs of the sequence 'KITTI_07' in the KITTI dataset [102]: (A) and (B) are the outputs on the sequence with Gaussian noise and Salt-and-Pepper noise, respectively.

under Salt-and-Pepper noise. The results of the point cloud map constructed on the sequence 'KITTI_07' in the KITTI dataset are shown in Figure 13. The results in Tables 10 and 11 show that the improved LDSO in [102] can work efficiently in both the indoor and the outdoor datasets under different noises, while the general LDSO will fail to track (see Figure 13).

3.5.2. Scene classification

Scene classification is one of the key technologies of scene understanding for autonomous robots, which can provide the basis for decision-making of the robots. The task of the scene classification for an autonomous robot refers to the information of its surroundings obtained by the on-board sensors, and then the state of the current position is recognized.

Lots of researchers have conducted studies on scene classification. For example, Tang *et al.* [106] proposed an adaptive discriminative region learning network for remote sensing scene classification, which locates discriminative regions effectively for solving the problems of scene classification, such as scale-variation of objects and redundant and noisy areas. Song *et al.* [107] used an ensemble alignment subspace adaptation method for the cross-scene classification. It can settle the problem of both foreign objects in the same spectrum and different spectra. Zhu *et al.* [108] proposed a domain adaptation cross-scene classification approach to simultaneously

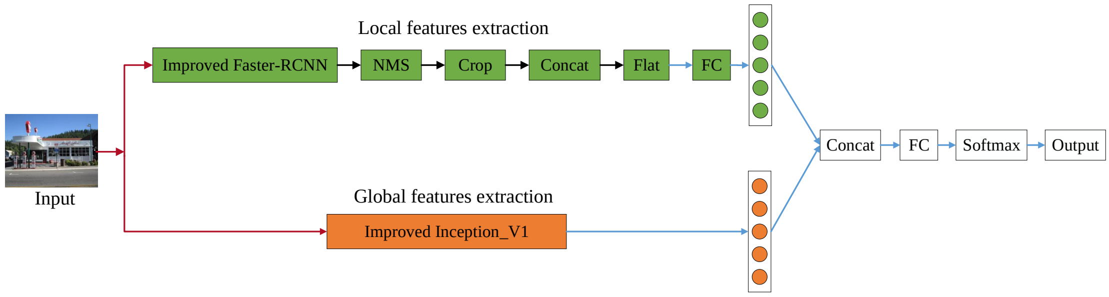

Figure 14. The structure of the proposed deep network for the road scene classification [109].

classify the target common categories and detect the target private categories based on the divergence of different classifiers.

The methods for the scene classification can be divided into two main types. One of them is based on the underlying visual features. This type of method has some shortcomings. For example, the accuracy of the scene classification is low when only the low-level visual features are used to represent the contents of the scene. The other type of the scene classification method is based on the deep learning technologies. To deal with the problem of the scene classification of the road scene, our group presented an improved deep network-based model [109]. The structure of the proposed model is shown in Figure 14.

As shown in Figure 14, there are four main parts in the proposed scene classification model, namely, (1) The improved Faster RCNN-based local feature extraction module; (2) The improved Inception_V1-based global feature extraction module; (3) The feature fusion module; (4) The classification network.

In the improved Faster RCNN-based local feature extraction module, the VGG16 Net is used to get the feature map of the whole image first. Then, a residual attention module is used to further deal with redundant information in images. The operation on the feature map based on the residual attention module is:

$$F_{output}(i,j) = F_{input}(i,j) \otimes a_{ij} + F_{input}(i,j) \tag{14}$$

where F_{output} and F_{input} are the output and the input feature value of the residual attention module, respectively; a_{ij} is the attention weight; $\otimes$ is the dot product operation.

The output of the residual attention module is input into the RPN to generate region proposals. The output Region-of-Interests (ROIs) of the RPN is processed by a ROI pooling network to get a fixed-size proposal feature map, which is finally input into a fully connected layer for the object classification and generating the positions of the objects.

In the global feature extraction module, the Inception_V1 is used as the baseline network, which has nine Inception blocks. One Inception block has four branches. To deal with the shortcomings of the general Inception_V1 [110], the Inception_V1 is improved in the proposed model in [109], where a mixed activation function is presented by alternately using the ELU and Leaky ReLU functions for the Inception networks. The Leaky ReLU function is denoted by:

$$y_i = \begin{cases} x_i, & \text{if } x_i \geq 0 \\ \alpha x_i, & \text{if } \alpha x_i < 0 \end{cases} \tag{15}$$

where α is a fixed parameter.

Table 12. The experimental results of scene classification based on different deep networks [109]

Network	Total accuracy	Standard deviation	On sunny days	On rainy days	At night
AlexNet [113]	84.20%	5.22%	90.20%	81.70%	80.70%
EfficientNet [114]	87.07%	8.31%	96.30%	80.00%	85.30%
Inception_V1 [110]	90.53%	2.51%	93.40%	88.70%	89.50%
Ours	**94.76%**	**1.62%**	**96.50%**	**93.30%**	**94.50%**

The ELU function is denoted by:

$$y_i = \begin{cases} x_i, \text{ if } x_i \geq 0 \\ e^{x_i} - 1, \text{ if } x_i < 0 \end{cases} \tag{16}$$

In the feature fusion module, the local feature vectors and the global feature vectors are appended to get the fused feature F, namely

$$F = [L, G] \tag{17}$$

where $L = [l_1, l_2, \cdots, l_N]$ and $G = [g_1, g_2, \cdots, g_N]$ denote the local feature vectors and the global feature vectors, respectively; N is the feature dimension.

At last, the fused feature vector is input to the classification network for the scene classification. The loss function used in this classification network is as follows:

$$Loss_{cls} = \frac{1}{S} \sum_i \left(- \sum_{j=1}^{C} y_{ij} \log \left(p_{ij} \right) \right) \tag{18}$$

where C is the number of scene classification; S is the number of the samples; p_{ij} is the probability that the i-th sample belongs to the j-th category, and y_{ij} is the indicator variable.

To test the performance of the proposed road scene classification model, our group proposed a special dataset based on two public datasets: KITTI [111] and Place365 [112]. The results of the comparison experiments are listed in Table 12, and some scene classification results based on different models are shown in Figure 15.

It can be seen that our proposed model can improve the accuracy to 94.76%, which is 4.67% (Relative value) higher than the general Inception_V1 (the second-best model). In addition, our proposed model has good scene classification performance under some challenging tasks, such as the task on a rainy day or at night (see Table 12 and Figure 15 for details).

4. FUTURE DIRECTIONS

With the developments of the artificial intelligence technologies and deep learning methods, great progress has been made in the research of scene understanding for autonomous robots. However, there are still a lot of difficulties in using deep learning to perceive and understand the surroundings for autonomous robots. There are some problems that should be further studied as follows:

(1) Light-weight models: With the continuous improvement of the computing power of hardware devices, the scene understanding method based on deep learning technology has achieved great success. However, it is difficult to run large-scale models on autonomous robots with limited processing, memory, and power resources. How to design a practical light-weight deep learning model while keeping the desired accuracy is a challenging task. Meanwhile, it also needs to develop efficient compact representation models for 3D data.

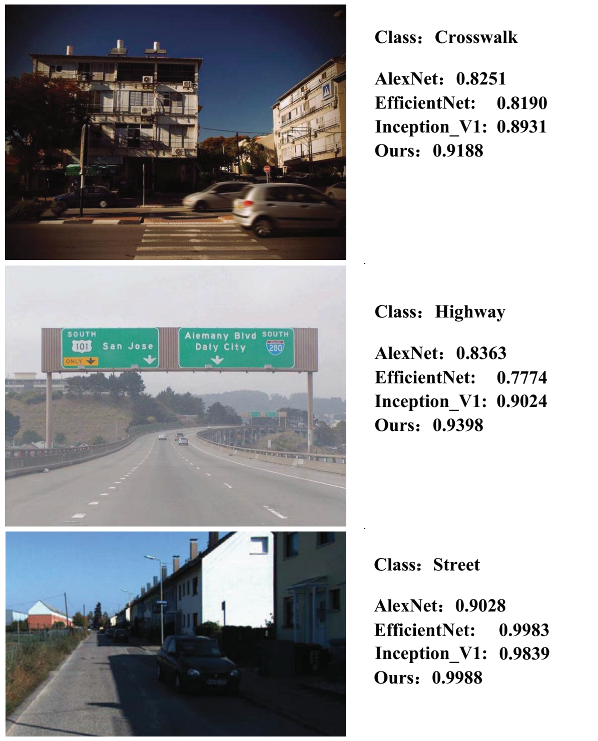

Figure 15. Some scene classification results based on different models [109].

(2) Multi-task learning: A valuable but less explored direction for scene understanding is to jointly train models on multiple terminal tasks. For example, semantic contour detection technology could jointly detect target contours and recognize the semantic information of the contours. This multi-task learning method is useful for model learning without decreasing the performance of any single task.

(3) Transfer learning: Common tasks such as object detection, semantic segmentation, and scene classification usually have many annotated examples for training. However, there is a lack of large datasets for tasks such as layout estimation, affordance prediction, and physics-based reasoning. How to optimally fine-tune an existing model to the desired task so that the knowledge is properly transferred from the source domain to the target domain is a good research direction in this field.

(4) Multi-modal fusion: Building a cross-modal adaptive fusion network will allow us to more fully fuse the sparse information in the point cloud space with the dense information in the image space. Based on these multi-modal fusion methods, the accuracy of the scene understanding can be further improved. In this field, how to fuse different modal information efficiently is a good research direction.

(5) The specific datasets: To improve the performance of the deep learning-based models, some specific datasets should be constructed for the applications of the robots in different environments. For example, how to make the autonomous underwater vehicle (AUV) work efficiently is still a challenging task. The main reason is that the underwater environments are complex; for example, the illumination is low, and the reference objects are fewer. To build a specific dataset for special robots is arduous, but it is very meaningful.

(6) Application extensions: With the popularization of robot applications and the important role that robots play in various fields, we need to take a step forward in researching the applications of scene understanding for autonomous robots. In addition to the applications mentioned above, such as target detection and pose estimation, we need to focus on more application extensions, such as physics-based reasoning, affordance

prediction, full 3D reconstruction, etc.

The scene understanding of autonomous robots is the first prerequisite for autonomous robots to complete complex tasks. On this basis, robots can become smarter to further improve social productivity, produce huge social benefits, and improve people's life quality. Therefore, there are many problems that need to be solved efficiently. The deep learning-based methods for the robotic scene understanding are still on the way.

5. CONCLUSIONS

This study analyzes the most recent advancements in deep learning-based environment perception and understanding methods for autonomous robots. Firstly, this paper provides a summary of recent advances in the ability of autonomous robots to perceive and understand their environments. The typical application techniques for perceiving and understanding the surroundings by autonomous robots are discussed. Then, the research and application of deep learning-based methods in the field of scene understanding for autonomous robots are further discussed in this study, which also presents exemplary techniques for the use of robot environment perception and understanding. Lastly, the main issues and difficulties of deep learning-based autonomous robot scene understanding are examined.

It is obvious that the deep learning method will become one of the most popular research topics in the field of autonomous robot scene understanding, including theoretical and applied research. Deep learning-based technologies will further improve the intelligence and autonomy of robots. With a better perception and understanding of the environment, the robots will be able to solve complex tasks instead of just performing some simple and single commands. At present, many fundamental problems of robot scene understanding based on deep learning have been explored with exciting results, which show the potential of deep learning. But there are still many questions that need to be further studied.

DECLARATIONS

Authors' contributions

Funding acquisition: Ni J

Project administration: Ni J, Shi P

Writing-original draft: Chen Y, Tang G

Writing-review and editing: Shi J, Cao W

Availability of data and materials

Not applicable.

Financial support and sponsorship

This work has been supported by the National Natural Science Foundation of China (61873086) and the Science and Technology Support Program of Changzhou (CE20215022).

Conflicts of interest

The authors declared that they have no conflicts of interest related to this work.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

REFERENCES

1. Ni J, Wang X, Gong T, Xie Y. An improved adaptive ORB-SLAM method for monocular vision robot under dynamic environments. *Int J Mach Learn Cyber* 2022;13:3821-36. DOI
2. Li J, Xu Z, Zhu D, et al. Bio-inspired intelligence with applications to robotics: a survey. *Intell Robot* 2021;1:5883. DOI
3. Ni J, Tang M, Chen Y, Cao W. An improved cooperative control method for hybrid unmanned aerial-ground system in multitasks. *Int J Aerosp Eng* 2020. DOI
4. Zhao ZQ, Zheng P, Xu ST, Wu X. Object Detection with Deep Learning: A Review. *IEEE Trans Neural Netw Learn Syst* 2019;30:3212-32. DOI
5. Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Martinez-Gonzalez P, Garcia-rodriguez J. A survey on deep learning techniques for image and video semantic segmentation. *Appl Soft Comput* 2018;70:41-65. DOI
6. Wang L, Huang Y. A survey of 3D point cloud and deep learning-based approaches for scene understanding in autonomous driving. *IEEE Intell Transport Syst Mag* 2022;14:135-54. DOI
7. Naseer M, Khan S, Porikli F. Indoor scene understanding in 2.5/3D for autonomous agents: a survey. *IEEE Access* 2019;7:1859-87. DOI
8. Zhu M, Ferstera A, Dinulescu S, et al. A peristaltic soft, wearable robot for compression therapy and massage. *IEEE Robot Autom* 2023;8:4665-72. DOI
9. Sun P, Shan R, Wang S. An intelligent rehabilitation robot with passive and active direct switching training: improving intelligence and security of human-robot interaction systems. *IEEE Robot Automat Mag* 2023;30:72-83. DOI
10. Wang TM, Tao Y, Liu H. Current researches and future development trend of intelligent robot: a review. *Int J Autom Comput* 2018;15:525-46. DOI
11. Lowe DG. Distinctive image features from scale-invariant keypoints. *Int J comput vis* 2004;60:91-110. DOI
12. Zhou H, Yuan Y, Shi C. Object tracking using SIFT features and mean shift. *Comput Vis Image Underst* 2009;113:345-52. DOI
13. Oliva A, Torralba A. Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int J comput vis* 2001;42:145-75. DOI
14. Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. *Mach Learn* 2001;42:177-96. DOI
15. Sarhan S, Nasr AA, Shams MY. Multipose face recognition-based combined adaptive deep learning vector quantization. *Comput Intell Neurosci* 2020;2020:8821868. DOI
16. Liu B, Wu H, Su W, Zhang W, Sun J. Rotation-invariant object detection using Sector-ring HOG and boosted random ferns. *Vis Comput* 2018;34:707-19. DOI
17. Wang X, Han TX, Yan S. An HOG-LBP human detector with partial occlusion handling. In: 2009 IEEE 12th International Conference on Computer Vision; 2009 Sep 29 - Oct 02; Kyoto, Japan. IEEE; 2010. p. 32-39. DOI
18. Vailaya A, Figueiredo MA, Jain AK, Zhang HJ. Image classification for content-based indexing. *XX* 2001;10:117-30. DOI
19. Li LJ, Su H, Xing EP, Fei-Fei L. Object bank: a high-level image representation for scene classification & semantic feature sparsification. In: Proceedings of the 23rd International Conference on Neural Information Processing Systems; 2010. p. 1378–86. Available from: https://proceedings.neurips.cc/paper/2010/hash/140f6969d5213fd0ece03148e62e461e-Abstract.html [Last accessed on 8 Aug 2023]
20. Zhang L, Li W, Yu L, Sun L, Dong X, Ning X. GmFace: an explicit function for face image representation. *Displays* 2021;68:102022. DOI
21. Ning X, Gong K, Li W, Zhang L, Bai X, et al. Feature refinement and filter network for person re-identification. *IEEE Trans Circuits Syst Video Technol* 2021;31:3391-402. DOI
22. Ni J, Chen Y, Chen Y, Zhu J, Ali D, Cao W. A survey on theories and applications for self-driving cars based on deep learning methods. *Appl Sci* 2020;10:2749. DOI
23. Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence; 2012 Jun 16-21; RI, USA. IEEE; 2012. p. 3354-61. DOI
24. Caesar H, Bankiti V, Lang AH, et al. nuScenes: a multimodal dataset for autonomous driving. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13-19; Seattle, WA, USA. IEEE; 2020. p. 11618-28. DOI
25. Hinterstoisser S, Lepetit V, Ilic S, et al. Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes. In: Lee KM, Matsushita Y, Rehg JM, Hu Z, editors. Computer Vision - ACCV; 2013. p. 548-62. DOI
26. Piga NA, Onyshchuk Y, Pasquale G, Pattacini U, Natale L. ROFT: Real-time tptical flow-aided 6D object pose and velocity tracking. *IEEE Robot Autom Lett* 2022;7:159-66. IEEE Robotics and Automation Letters DOI
27. Everingham M, Gool LV, Williams CKI, Winn JM, Zisserman A. The pascal visual object classes (VOC) challenge. *Int J Comput Vis* 2010;88:303-38. DOI
28. Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27-30; Las Vegas, NV, USA. IEEE; 2016. p. 3213-23. DOI
29. Wang W, Shen J, Guo F, Cheng MM, Borji A. Revisiting video saliency: a large-scale benchmark and a new model. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18-23; Salt Lake City, UT, USA. IEEE; 2018. p. 4894-903. DOI
30. Kristan M, Leonardis A, Matas J, et al. The visual object tracking VOT2017 challenge results. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW); 2017 Oct 22-29; Venice, Italy. IEEE; 2017. p. 1949-72. DOI
31. Ni J, Shen K, Chen Y, Yang SX. An improved SSD-like deep network-based object detection method for indoor scenes. *IEEE Trans Instrum Meas* 2023;72:1-15. IEEE Transactions on Instrumentation and Measurement 2023;72:1-15. DOI
32. Qian R, Lai X, Li X. BADet: boundary-aware 3D object detection from point clouds. *Pattern Recognit* 2022;125:108524. DOI
33. Shi S, Wang Z, Shi J, Wang X, Li H. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation

network. *IEEE Trans Pattern Anal Mach Intell* 2021;43:2647-64. DOI

34. Li Y, Ma L, Zhong Z, Cao D, Li J. TGNet: geometric graph CNN on 3-D point cloud segmentation. *IEEE Trans Geosci Remote Sens* 2020;58:3588-600. DOI
35. Yan Y, Mao Y, Li B. SECOND: sparsely embedded convolutional detection. *Sensors* 2018;18:3337. DOI
36. Qi CR, Liu W, Wu C, Su H, Guibas LJ. Frustum pointnets for 3D object detection from RGB-D data. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018 Jun 18-23; Salt Lake City, UT, USA; IEEE; 2018. p. 918-27. DOI
37. Wang Z, Jia K. Frustum convNet: sliding frustums to aggregate local point-wise features for amodal 3D object detection. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2019 Nov 03-08; Macau, China. IEEE; 2019. p. 1742-49. DOI
38. Chen Y, Liu S, Shen X, Jia J. Fast point R-CNN. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27 - Nov 02; Seoul, Korea. IEEE; 2019. p. 9774-83. DOI
39. He C, Zeng H, Huang J, Hua XS, Zhang L. Structure aware single-stage 3D object detection from point cloud. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13-19; Seattle, WA, USA. IEEE; 2020. p. 11870-9. DOI
40. Liu Z, Zhao X, Huang T, Hu R, Zhou Y, Bai X. TANet: robust 3D object detection from point clouds with triple attention. In: 34th AAAI Conference on Artificial Intelligence, AAAI; 2020 Feb 7-12; New York, NY, United states. California: AAAI; 2020. p. 11677-84. DOI
41. Yin T, Zhou X, Krahenbuhl P. Center-based 3D Object Detection and Tracking. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2021. Virtual, Online, United states; 2021. pp. 11779 – 11788. DOI
42. Wang H, Shi S, Yang Z, et al. RBGNet: ray-based Grouping for 3D Object Detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18-24; New Orleans, LA, USA. IEEE; 2022. p. 1100-09. DOI
43. Chen Y, Ni J, Tang G, Cao W, Yang SX. An improved dense-to-sparse cross-modal fusion network for 3D object detection in RGB-D images. *Multimed Tools Appl* 2023. DOI
44. Hoang DC, Stork JA, Stoyanov T. Voting and attention-based pose relation learning for object pose estimation from 3D point clouds. *IEEE Robot Autom Lett* 2022;7:8980-7. DOI
45. Yue K, Sun M, Yuan Y, Zhou F, Ding E, Xu F. Compact generalized non-local network. arXiv.[Preprint.] November 1, 2018. Available from: https://arxiv.org/abs/1810.13125 [Last accessed on 8 Aug 2023]
46. Chen H, Wang P, Wang F, Tian W, Xiong L, Li H. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. arXiv. [Preprint.] August 11, 2022 Available from: https://arxiv.org/abs/2203.13254 [Last accessed on 8 Aug 2023]
47. Moon G, Chang JY, Lee KM. V2V-poseNet: voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18-23; Salt Lake City, UT, USA.IEEE; 2018. p. 5079-88. DOI
48. Li Z, Wang G, Ji X. CDPN: coordinates-based disentangled pose network for real-time RGB-based 6-DoF object pose estimation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27 - Nov 02; Seoul, Korea (South). IEEE; 2020. p. 7677-86. DOI
49. Wang H, Sridhar S, Huang J, Valentin J, Song S, Guibas LJ. Normalized object coordinate space for category-level 6D object pose and size estimation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15-20; Long Beach, CA, USA. IEEE; 2020. p. 2637-46. DOI
50. Yu X, Zhuang Z, Koniusz P, Li H. 6DoF object pose estimation via differentiable proxy voting loss. arXiv. [Preprint.] Febuary 11, 2020. Available from: https://arxiv.org/abs/2002.03923 [Last accessed on 8 Aug 2023]
51. Chen W, Jia X, Chang HJ, Duan J, Leonardis A. G2L-net: global to local network for real-time 6D pose estimation with embedding vector features. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020 Jun 13-19; Seattle, WA,USA. IEEE; 2020. p. 4232-41. DOI
52. He Y, Sun W, Huang H, Liu J, Fan H, Sun J. PVN3D: a deep point-wise 3D keypoints voting network for 6DoF pose estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13-19 Seattle, WA, USA. IEEE; 2020. pp. 11629-38. DOI
53. He Y, Huang H, Fan H, Chen Q, Sun J. FFB6D: a full flow bidirectional fusion network for 6D pose estimation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20-25; Nashville, TN, USA. IEEE; 2021. p. 3002-12. DOI
54. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. arXiv. [Preprint.] November 14, 2014. Available from: https://arxiv.org/abs/1411.4038 [Last accessed on 8 Aug 2023]
55. Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell* 2017;39:2481-95. DOI
56. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Semantic image segmentation with deep convolutional nets and fully connected CRFs. arXiv. [Preprint.] December 22, 2014. Available from: https://arxiv.org/abs/1412.7062 [Last accessed on 8 Aug 2023]
57. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans Pattern Anal Mach Intell* 2017;40:834-48. DOI
58. Lin G, Milan A, Shen C, Reid I. Refinenet: multi-path refinement networks for high-resolution semantic segmentation. In: 2017IEEE conference on computer vision and pattern recognition (CVPR); 2017. p. 1925-34. DOI
59. Zeng L, Zhang S, Wang P, Li Z, Hu Y, Xie T. Defect detection algorithm for magnetic particle inspection of aviation ferromagnetic parts based on improved DeepLabv3+. *Meas Sci Technol* 2023;34:065401. Measurement Science and Technology 2023;34:065401. DOI

60. Yin R, Cheng Y, Wu H, Song Y, Yu B, Niu R. Fusionlane: multi-sensor fusion for lane marking semantic segmentation using deep neural networks. *IEEE Trans Intell Transport Syst* 2022;23:1543-53. DOI
61. Hu P, Perazzi F, Heilbron FC, et al. Real-time semantic segmentation with fast attention. *IEEE Robot Autom Lett* 2021;6:263-70. DOI
62. Sun Y, Zuo W, Yun P, Wang H, Liu M. FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion. *IEEE Trans Automat Sci Eng* 2021;18:1000-11. DOI
63. Yang M, Yu K, Zhang C, Li Z, Yang K. DenseASPP for semantic segmentation in street scenes. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18-23; Salt Lake City, UT, USA. IEEE; 2018. p. 3684-92. DOI
64. Zhang H, Dana K, Shi J, et al. Context encoding for semantic segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18-23; Salt Lake City, UT, USA. IEEE; 2018. p. 7151-60. DOI
65. Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15-20; Long Beach, CA, USA. IEEE; 2020. p. 3141-9. DOI
66. He J, Deng Z, Zhou L, Wang Y, Qiao Y. Adaptive pyramid context network for semantic segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15-20; Long Beach, CA, USA. IEEE; 2020. p. 7511-20. DOI
67. Zhang C, Lin G, Liu F, Yao R, Shen C. CANet: class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15-20; Long Beach, CA, USA. IEEE; 2020. p. 5212-21. DOI
68. Liu J, He J, Zhang J, Ren JS, Li H. EfficientFCN: holistically-guided decoding for semantic segmentation. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. Computer Vision – ECCV 2020. Cham: Springer; 2020. p. 1-17. DOI
69. Ha Q, Watanabe K, Karasawa T, Ushiku Y, Harada T. MFNet: towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS); 2017 Sep 24-28; Vancouver, BC, Canada. IEEE; 2017. p. 5108-15. DOI
70. Cheng B, Schwing AG, Kirillov A. Per-pixel classification is not all you need for semantic segmentation. *Signal Process Image Commun* 2021;88:17864-75. DOI
71. Zhou W, Yue Y, Fang M, Qian X, Yang R, Yu L. BCINet: bilateral cross-modal interaction network for indoor scene understanding in RGB-D images. *Inf Fusion* 2023;94:32-42. DOI
72. Lou J, Lin H, Marshall D, Saupe D, Liu H. TranSalNet: towards perceptually relevant visual saliency prediction. *Neurocomputing* 2022;494:455-67. DOI
73. Judd T, Ehinger K, Durand F, Torralba A. Learning to predict where humans look. In: 2009 IEEE 12th International Conference on Computer Vision; 2009 Sep 29 - Oct 02; Kyoto, Japan. IEEE; 2010. p. 2106-13. DOI
74. Ishikura K, Kurita N, Chandler DM, Ohashi G. Saliency detection based on multiscale extrema of local perceptual color differences. *IEEE Trans Image Process* 2018;27:703-17. DOI
75. Zou W, Zhuo S, Tang Y, Tian S, Li X, Xu C. STA3D: spatiotemporally attentive 3D network for video saliency prediction. *Pattern Recognit Lett* 2021;147:78-84. DOI
76. Wang W, Shen J, Dong X, Borji A. Salient object detection driven by fixation prediction. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18-23; Salt Lake City, UT, USA. IEEE; 2018. p. 1711-20. DOI
77. Huang R, Xing Y, Wang Z. RGB-D salient object detection by a CNN with multiple layers fusion. *IEEE Signal Process Lett* 2019;26:552-6. IEEE Signal Processing Letters DOI
78. Wang N, Gong X. Adaptive fusion for RGB-D salient object detection. *IEEE Access* 2019;7:55277-84. DOI
79. Zhang J, Yu M, Jiang G, Qi Y. CMP-based saliency model for stereoscopic omnidirectional images. *Digit Signal Process* 2020;101:102708. DOI
80. Fang Y, Zhang C, Min X, et al. DevsNet: deep video saliency network using short-term and long-term cues. *Pattern Recognit* 2020;103:107294. DOI
81. Li F, Zheng J, fang Zhang Y, Liu N, Jia W. AMDFNet: adaptive multi-level deformable fusion network for RGB-D saliency detection. *Neurocomputing* 2021;465:141-56. DOI
82. Lee H, Kim S. SSPNet: learning spatiotemporal saliency prediction networks for visual tracking. *Inf Sci* 2021;575:399-416. DOI
83. Xue H, Sun M, Liang Y. ECANet: explicit cyclic attention-based network for video saliency prediction. *Neurocomputing* 2022;468:233-44. DOI
84. Zhang N, Nex F, Kerle N, Vosselman G. LISU: low-light indoor scene understanding with joint learning of reflectance restoration. *SPRS J Photogramm Remote Sens* 2022;183:470-81. DOI
85. Tang G, Ni J, Chen Y, Cao W, Yang SX. An improved cycleGAN based model for low-light image enhancement. *IEEE Sensors J* 2023. DOI
86. He J, Li M, Wang Y, Wang H. OVD-SLAM: an online visual SLAM for dynamic environments. *IEEE Sensors J* 2023;23:13210-9. DOI
87. Lu X, Sun H, Zheng X. A feature aggregation convolutional neural network for remote sensing scene classification. *IEEE Trans Geosci Remote Sens* 2019;57:7894-906. DOI
88. Ma D, Tang P, Zhao L. SiftingGAN: generating and sifting labeled samples to improve the remote sensing image scene classification baseline *in vitro*. *IEEE Geosci Remote Sens Lett* 2019;16:1046-50. IEEE Geoscience and Remote Sensing Letters DOI
89. Zhang X, Qiao Y, Yang Y, Wang S. SMod: scene-specific-prior-based moving object detection for airport apron surveillance systems. *IEEE Intell Transport Syst Mag* 2023;15:58-69. DOI
90. Tang G, Ni J, Shi P, Li Y, Zhu J. An improved ViBe-based approach for moving object detection. *Intell Robot* 2022;2:13044. DOI
91. Lee CY, Badrinarayanan V, Malisiewicz T, Rabinovich A. Roomnet: end-to-end room layout estimation. arXiv. [Preprint.] August 7,

2017. Available from: https://arxiv.org/abs/1703.06241 [Last accessed on 8 Aug 2023]

92. Hsiao CW, Sun C, Sun M, Chen HT. Flat2layout: Flat representation for estimating layout of general room types. arXiv. [Preprint.] May 29, 2019. Available from: https://arxiv.org/abs/1905.12571 [Last accessed on 8 Aug 2023]

93. Sarhan S, Nasr AA, Shams MY. Multipose face recognition-based combined adaptive deep learning vector quantization. *Comput Intell Neurosci* 2020;2020:8821868. DOI

94. Rublee E, Rabaud V, Konolige K, Bradski G. ORB: an efficient alternative to SIFT or SURF. In: 2011 International conference on computer vision; 2011 Nov 06-13; Barcelona, Spain. IEEE; 2012. p. 2564-71. DOI

95. Wang K, Ma S, Ren F, Lu J. SBAS: salient bundle adjustment for visual SLAM. *IEEE Trans Instrum Meas* 2021;70:1-9. DOI

96. Ni J, Gong T, Gu Y, Zhu J, Fan X. An improved deep residual network-based semantic simultaneous localization and mapping method for monocular vision robot. *Comput Intell Neurosci* 2020;2020:7490840. DOI

97. Fu Q, Yu H, Wang X, et al. Fast ORB-SLAM without keypoint descriptors. *IEEE Trans Image Process* 2022;31:1433-46. DOI

98. Engel J, Schöps T, Cremers D. LSD-SLAM: large-scale direct monocular SLAM. In: Fleet D, Pajdla T, Schiele B, Tuytelaars, editors. Computer Vision – ECCV 2014. Cham: Springer; 2014. p. 834-49. DOI

99. Engel J, Koltun V, Cremers D. Direct sparse odometry. *IEEE Trans Image Process* 2018;40:611-25 DOI

100. Wang Y, Zhang S, Wang J. Ceiling-view semi-direct monocular visual odometry with planar constraint. *Remote Sens* 2022;14:5447. DOI

101. Forster C, Zhang Z, Gassner M, Werlberger M, Scaramuzza D. SVO: semidirect visual odometry for monocular and multicamera systems. *IEEE Trans Robot* 2017;33:249-65. DOI

102. Chen Y, Ni J, Mutabazi E, Cao W, Yang SX. A variable radius side window direct SLAM method based on semantic information. *Comput Intell Neurosci* 2022;2022:4075910. DOI

103. Liu L. Image classification in htp test based on convolutional neural network model. *Comput Intell Neurosci* 2021;2021:6370509. DOI

104. Zheng D, Li L, Zheng S, et al. A defect detection method for rail surface and fasteners based on deep convolutional neural network. *Comput Intell Neurosci* 2021;2021:2565500. DOI

105. Gao X, Wang R, Demmel N, Cremers D. LDSO: direct sparse odometry with loop closure. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2018 Oct 01-05; Madrid, Spain. IEEE; 2019. p. 2198-204. DOI

106. Tang C, Zheng X, Tang C. Adaptive discriminative regions learning network for remote sensing scene classification. *Sensors* 2023;23:1-5. DOI

107. Song Y, Feng W, Dauphin G, Long Y, Quan Y, Xing M. Ensemble alignment subspace adaptation method for cross-scene classification. *IEEE Geosci Remote Sensing Lett* 2023;20:1-5. DOI

108. Zhu S, Wu C, Du B, Zhang L. Adversarial divergence training for universal cross-scene classification. *IEEE Trans Geosci Remote Sens* 2023;61:1-12. DOI

109. Ni J, Shen K, Chen Y, Cao W, Yang SX. An improved deep network-based scene classification method for self-driving cars. *IEEE Trans Instrum Meas* 2022;71:1-14. DOI

110. Mohapatra RK, Shaswat K, Kedia S. Offline handwritten signature verification using CNN inspired by inception V1 architecture. In: 2019 Fifth International Conference on Image Information Processing (ICIIP); 2019 Nov 15-17; Shimla, India. IEEE; 2020. p. 263-7. DOI

111. McCall R, McGee F, Mirnig A, et al. A taxonomy of autonomous vehicle handover situations. *Transp Res Part A Policy Pract* 2019;124:507-22. DOI

112. Wang L, Guo S, Huang W, Xiong Y, Qiao Y. Knowledge guided disambiguation for large-scale scene classification with multi-resolution CNNs. *IEEE Trans Image Process* 2017;26:2055-68. DOI

113. Hosny KM, Kassem MA, Fouad MM. Classification of skin lesions into seven classes using transfer learning with AlexNet. *J Digit Imaging* 2020;33:1325-34. DOI

114. Alhichri H, Alsuwayed A, Bazi Y, Ammour N, Alajlan NA. Classification of remote sensing images using EfficientNet-B3 CNN model with attention. *IEEE Access* 2021;9:14078-94. DOI