IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 11, 2020, accepted September 23, 2020, date of publication October 5, 2020, date of current version October 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3028740

Object Detection Recognition and Robot Grasping Based on Machine Learning: A Survey

QIANG BAI[1], SHAOBO LI[1,2,3], JING YANG[1,3], (Member, IEEE), QISONG SONG[1], ZHIANG LI[1], AND XINGXING ZHANG[1]

[1]School of Mechanical Engineering, Guizhou University, Guiyang 550025, China
[2]Key Laboratory of Advanced Manufacturing Technology, Ministry of Education, Guizhou University, Guiyang 550025, China
[3]Guizhou Province Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China

Corresponding author: Shaobo Li (lishaobo@gzu.edu.cn)

This work was supported in part by the National Key Technologies Research and Development Program of China under Grant 2018AAA0101800, in part by the National Natural Science Foundation of China under Grant 51475097 and Grant 91746116, in part by the Ministry of Industry and Information Technology of the People's Republic of China Talents under Grant [2016]213, and in part by the Science and Technology Project of Guizhou Province Talents under Grant [2015]4011 and Grant [2016]5013.

ABSTRACT With the rapid development of machine learning, its powerful function in the machine vision field is increasingly reflected. The combination of machine vision and robotics to achieve the same precise and fast grasping as that of humans requires high-precision target detection and recognition, location and reasonable grasp strategy generation, which is the ultimate goal of global researchers and one of the prerequisites for the large-scale application of robots. Traditional machine learning has a long history and good achievements in the field of image processing and robot control. The CNN (convolutional neural network) algorithm realizes training of large-scale image datasets, solves the disadvantages of traditional machine learning in large datasets, and greatly improves accuracy, thereby positioning CNNs as a global research hotspot. However, the increasing difficulty of labeled data acquisition limits their development. Therefore, unsupervised learning, self-supervised learning and reinforcement learning, which are less dependent on labeled data, have also undergone rapid development and achieved good performance in the fields of image processing and robot capture. According to the inherent defects of vision, this paper summarizes the research achievements of tactile feedback in the fields of target recognition and robot grasping and finds that the combination of vision and tactile feedback can improve the success rate and robustness of robot grasping. This paper provides a systematic summary and analysis of the research status of machine vision and tactile feedback in the field of robot grasping and establishes a reasonable reference for future research.

INDEX TERMS Machine learning, recognition, grasping, robot, tactile feedback, vision.

I. INTRODUCTION

Vision is the main way in which humans to receive all types of information, followed by tactile feedback. One goal of researchers is to equip robots with vision systems that have high accuracy and robustness, similar to human beings, to help people complete all types of work. Thus, machine vision has always been an important research topic in the field of artificial intelligence and robotics. With the rapid development of machine learning, machine vision has been widely and successfully applied in various image processing tasks, such as defect detection, target detection, medical image judgment [1]–[14], etc. To this end, researchers hope to achieve great breakthroughs in machine vision to allow for precise recognition, positioning and grasp strategy generation and the realization of stable grasping of robots, which could lead to wide application.

Although the above papers provide a wide range of research and surveys of machine learning and machine vision in plain image processing, there are very few surveys of machine learning used for object detection recognition and robot grasping. Accurate and fast object recognition and grasping based on vision are the basis of robot applications in both industry and real-life scenarios. This paper mainly summarizes the research achievements of six mainstream methods in object detection recognition, positioning,

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Zhou.

grasp strategy generation and grabbing, including traditional machine learning, deep learning, unsupervised learning, self-supervised learning, reinforcement learning and visual-tactile fusion. Machine learning is the inevitable product of artificial intelligence development to a certain stage and has been put forward and developed for decades. The most substantial advantage of traditional machine learning (support vector machine (SVM), random forest, decision tree, clustering, and Bayesian algorithms) is that it requires only a small amount of data and has strong interpretability and fast running speed [15]–[17]. However, with the increase in the amount of data, the performance of these algorithms becomes limited and stagnated instead of continuing to improve [18], [19]. For a long time after the birth of the neural network algorithm in the 1980s, SVMs and other machine learning algorithms had an advantage. However, the gradient vanishing problem of these algorithms has led to difficulties in deep network training [20], [21] and revealed limitations in the number of samples and computing power. In 2012, the success of the Alex network led to the comeback of the deep neural network [22]. It is widely used in various fields of machine vision, and its performance continues to increase with the increase in datasets, avoiding the disadvantages of traditional machine learning in large datasets. Deep learning needs numerous labeled data, but it is not easy to label all of the data, which has led to the emergence of unsupervised and self-supervised learning algorithms. Unsupervised learning mainly addresses situations in which the input data is not labeled and the output is not determined [23], [24]. This approach classifies the samples according to the similarity. However, unsupervised learning has no label data at all, which may lead to slow speed and low precision [25]. Self-supervised learning uses the input data to generate supervisory information and benefits almost all types of downstream tasks [26], [27]. With Google's successful application of reinforcement learning in the Go game, reinforcement learning has attracted the worldwide attention of researchers. Reinforcement learning considers sequence problems and has a long-term perspective on long-term returns [28], while supervised learning generally considers one-off problems and focuses on only short-term and immediate returns. This long-term perspective of reinforcement learning is very important for determining the optimal solution to many problems. The key point of the above algorithm is to process the image collected by the camera, realize the object detection recognition positioning and grasp strategy generation and then guide the robot to complete the capture. However, noncontact object perception always has inherent defects, especially in unstructured environments and real-life scenes, and it is difficult to accurately predict the weight, shape and grasping strategy of the object [29]. Based on the above situation, adding pressure sensors to the dexterous hand to provision it with tactile feedback and combine it with vision has become a new direction in robot grasping research [30], [31].

This paper is organized as follows. The first part introduces the advantages and disadvantages of the six methods and the main content of this paper. The second part discusses the research achievements of several mainstream traditional machine learning methods in image processing, object recognition and guided robot grasping. The third part summarizes the performance of the convolutional neural network (CNN) algorithm in object detection recognition position and grasp strategy generation. In the fourth part, aiming to address the difficulty of acquiring label data, the paper describes the performance of unsupervised learning, self-supervised learning and reinforcement learning in the fields of vision and grabbing. The fifth part discusses the inherent defects of vision and summarizes the research achievements of robot tactile feedback and the combination of vision and tactile. In the sixth part, the future development prospects of machine vision in robot object recognition and grasping are proposed based on the above analysis. Finally, conclusions are drawn in the seventh part.

II. CLASSICAL MACHINE LEARNING

It has been nearly 70 years since Arthur Samuel put forward the concept of ''machine learning'' in 1952. In the 1980s, machine learning became an independent discipline and developed rapidly. Since 2006, due to the demand of big data analysis, neural networks based on machine learning have attracted more attention and become the basis of deep learning theory. Currently, the research of machine learning is mainly divided into two directions: the first is traditional machine learning, which mainly studies the learning principle and pays attention to exploring the learning mechanism of humanoids [32]–[36]; the second is the research of machine learning in big data environments, which mainly focuses on how to use information effectively and how to acquire hidden, effective and understandable knowledge from massive amounts of data [37]–[41]. From the perspective of methodology, machine learning can be divided into linear models and nonlinear models. Linear models are relatively simple, but they are the basis of nonlinear models, and many nonlinear models are transformed from linear models [42]–[46]. Nonlinear models can be divided into traditional machine learning models (SVM, KNN, decision tree, etc.) and deep learning models. Fig. 1 lists the currently mature traditional machine learning algorithms and briefly describes their principles and characteristics [47]–[51]. It is found that the functions of different algorithms are varied, indicating that each algorithm has different application scenarios. Although deep learning plays a dominant role in the field of machine vision, deep learning is data-driven and has poor performance in small datasets [52]–[54]. However, traditional machine learning can adapt to a variety of datasets; especially in scenarios with small amounts of data (such as the medical field), machine learning has better performance [55], [56]. In this case, the advantages of traditional machine learning algorithms are highlighted. Alternately, the traditional machine learning model is small, and the requirement of computer hardware is not high, which yields a strong speed advantage in the field of manipulator grasping-based vision [57]–[59]. According to

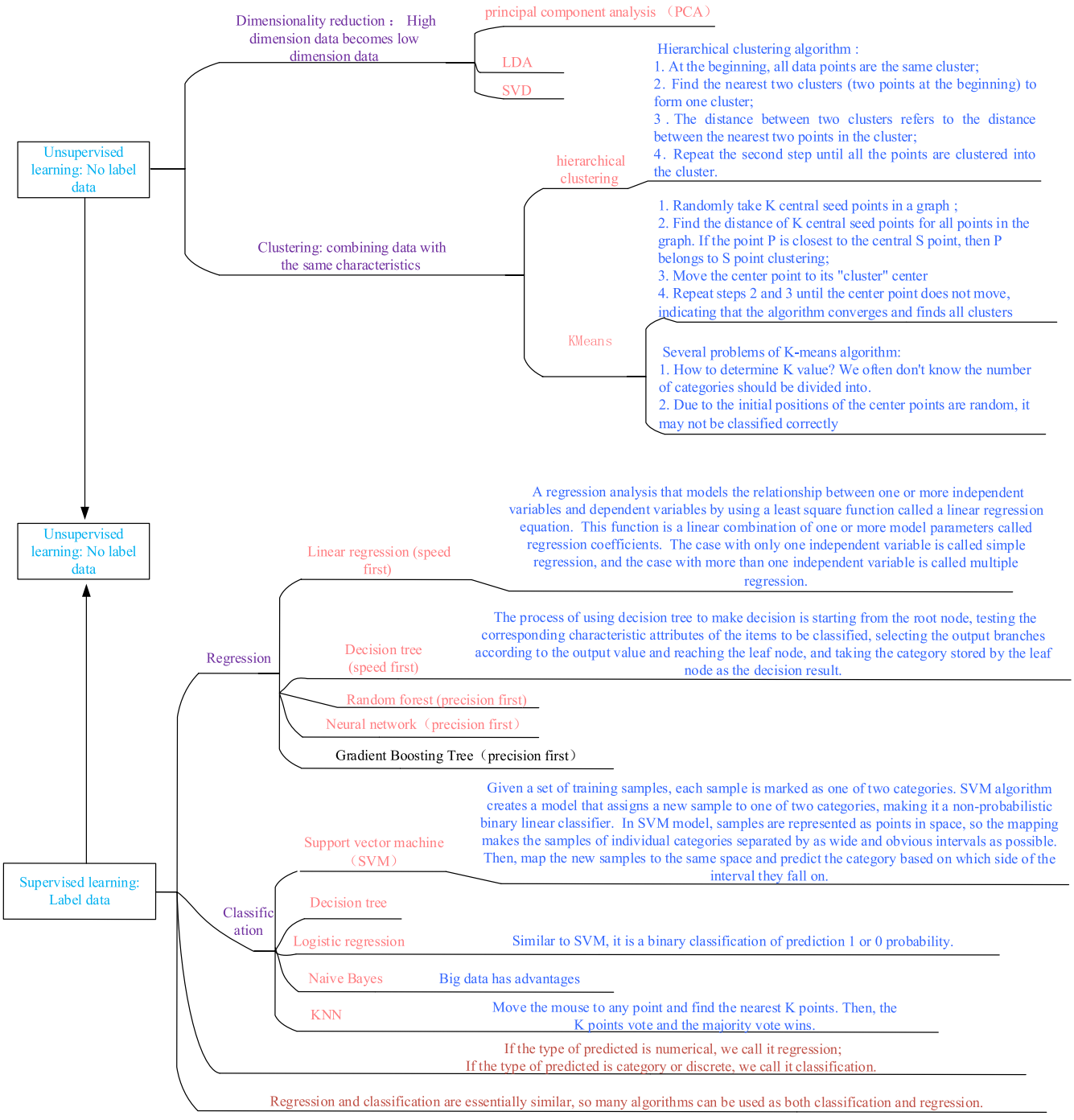


FIGURE 1. **Introduction of traditional machine learning.**

the characteristics of different machine learning algorithms, they can be applied in all aspects of manipulator grasping to improve the accuracy and robustness.

A. SUPPORT VECTOR MACHINE (SVM)

The SVM has strong generalization performance and can address machine learning problems in high-dimensional datasets and small samples, so it is widely used in the field of image processing. Based on RGB images and point cloud images, Yuan *et al.* [12] used the SVM-rank algorithm to recognize object features and generate the grabbing strategy and then realize the accurate grabbing of objects by a five-fingered dexterous hand. Ergene and Durdu [60] used the bag of words (BoW) method and an SVM to achieve feature extraction and object classification based on the grid and then guide the manipulator to achieve the classification

and grabbing of a pen, water cup and stapler. The accuracy was 83%. Hu *et al.* [61] developed an operation and grasp control system based on sensor-motor fusion for a robot hand-eye system, proposed a motion recognition method of a multifinger manipulator based on an AdaBoost-SVM, and proved the high response and flexibility of this method. Valente *et al.* [62] used the competitive Hopfield neural network to collect several points on the edge of the object to build an approximate polygon, used the radial bases function-global ridge regression (RBF) network to process the polygon, and selected the appropriate grasping points to guide the grasping of the manipulator.

SVM is a type of supervised learning method that has the advantages of good classification performance and simple structure but is difficult to train on large datasets and has poor performance on multiclassification problems. According to related research [12], [60], [61], SVMs have the disadvantages of complex feature work and poor generalization performance in target recognition, location and grasping. However, the improved SVM can be used in the robot grasping control algorithm and achieves good results.

B. CLUSTERING ALGORITHM

The clustering algorithm has the advantages of simplicity and easy implementation and can utilize large datasets, so it is widely used. Hannat *et al.* [63] presented a real-time method for visual categorization to achieve robot grasping. This method uses the speeded up robust feature (SURF) points to describe the feature data of objects and uses the K-means algorithm to extract the vocabulary. The results of our object recognition experiments show an average accuracy between 95% and 100%. Harada *et al.* [64] first clustered the polygon model of the object and the surrounding environment and then separated the environment and the object through different clustering algorithms to achieve successful grasping and stable placement. Verma *et al.* [65] proposed that the algorithm of density clustering and homography transformation can obtain the maximum stable extremal approach of the object and then realize the accurate positioning of the object, which provides powerful assistance for the successful grasping of the manipulator. Zhang and Shen [66] extracted effective local features from photos of the object. After clustering, these key points of each image are mapped into a uniform dimension histogram vector, and the histogram is used as the input vector of the multiclass SVM algorithm to establish the training classifier model and realize the real-time recognition of moving objects. Kouskouridas *et al.* [67] combined shape retrieval technology with a classification and clustering algorithm for attitude estimation of objects. Wiesmann *et al.* [68] proposed an event-driven embedded system for feature extraction and object recognition during robot grasping. Skotheim *et al.* [69] proposed a flexible 3D object positioning system that can make the manipulator assemble, grasp and place in a 3D environment. The system is improved based on a robust clustering algorithm and attitude verification algorithm, which significantly improves the accuracy and robustness of the system.

The clustering algorithm is a type of unsupervised algorithm with a long history, and it is widely used because it does not need training datasets and has a simple structure and fast speed. In tasks related to target detection and recognition, the clustering algorithm is mainly used for feature extraction and clustering, and it achieves the segmentation of the background and target location. However, existing research results [63], [66], [67], [69] have indicated that the clustering algorithm usually needs to be used together with other algorithms to achieve the classification and grasping of different targets.

C. BAYESIAN ALGORITHM

The Bayesian algorithm plays an important role in manipulator grasping planning. The naive Bayesian model originated from classical mathematical theory; it has a solid mathematical foundation and stable classification efficiency, performs well in small-scale datasets, and can handle multiclassification tasks. Budiharto [70] proposed a fast object detection algorithm based on stereo vision and used the Bayesian algorithm to reduce camera noise and achieve robust tracking. Wang *et al.* [71] proposed an online estimation method of a robot vision servo system based on a traceless particle filter and the Jacobian matrix. First, the definition of the total Jacobian matrix is given, and the estimation of the total Jacobian matrix is transformed into a Bayesian filtering framework. Then, the paper proposes to estimate the Jacobian matrix by a traceless particle filter and use the traceless Kalman filter equation to propagate and update each particle. Bekiroglu *et al.* [72] proposed a probabilistic framework for grasp modeling and stability assessment, which integrates supervised learning and unsupervised learning, and Bayesian networks are used to model the conditional relationship between tasks and multiple sensory flows (vision, ontological sensation and tactile). The obtained model can not only predict the success rate of grasping but also provide insight into the dependency between the related variables and features of object grabbing.

The Bayesian algorithm is widely used in noise reduction, servo control and grasping probability prediction in the research of target detection and recognition and robot grasping, which is mainly due to its solid mathematical foundation and its ability to address multiclassification tasks.

D. PRINCIPAL COMPONENT ANALYSIS (PCA)

In addition to the above algorithms, PCA also has applications in the field of vision and robotics. PCA finds the principal axis direction, which is used to effectively represent the common characteristics of the same type of samples. Song *et al.* [73] developed a general framework to estimate the ability of grasping from the 2D data of an object, which includes the identification of the similarity of the local features of the object and the generation of the object grabbing strategy based on the experience obtained from

the prelearning. Zhang *et al.* [74] proposed a shared control wheelchair manipulator, which can automatically detect a water cup based on vision and help the disabled achieve the task of drinking water. In this scheme, a CNN and PCA are used to separately identify and estimate the attitude and direction of the object. Mattar [75] proposed a learning mechanism for stable grasping and control of a manipulator. Based on a PCA neural network and the Widrow-Hoff method to learn a large number of patterns of prosthetic behavior, good grasp control of the prosthetic is realized.

PCA is an unsupervised learning method without parameter limitations, but it is seldom used in the image processing field. To achieve ideal robot grasping operation, PCA is commonly used with a CNN.

Machine learning algorithms have a long history of development and have made outstanding achievements in their respective fields. According to the algorithm principle and research (Table 1), it is found that target detection recognition and image processing are not the strong points of machine learning. First, machine learning algorithms require an arduous amount of feature engineering, which greatly increases the difficulty and cost of image processing. Second, machine learning requires a variety of algorithms to work together or with CNNs to achieve complete recognition, positioning and grasping, which increases the difficulty of model building and training. Finally, with the explosive growth of data in the era of big data, the disadvantages of traditional machine learning have become increasingly prominent.

TABLE 1. Comparison of machine learning application scenarios.

Algorithms	Supervised/ Unsupervised	Detect ion	Recog nition	Contr ol	Classific ation
Support Vector Machine (SVM)	Supervised	×	×	√	√
Clustering Algorithm	Unsupervised	√	×	×	×
Bayesian Algorithm	Supervised	×	×	√	√
Principal Component Analysis (PCA)	Unsupervised	×	×	√	×

III. CONVOLUTIONAL NEURAL NETWORK (CNN)

The CNN is one of the most representative neural networks in the field of deep learning and has made many breakthroughs in the field of image analysis and processing. Based on the standard image annotation set, ImageNet, the CNN has many achievements, including image feature extraction and classification, scene and target recognition, and so on. Compared with the traditional image processing algorithm, the CNN has the advantages of no preprocessing requirements and high precision [76]–[80], [82], [83]. In 1998, Yann Lecun *et al.* proposed a gradient-based back-propagation algorithm (LeNet-5) for supervised training of networks [84]. Yann Lecun is known as the father of the CNN for his outstanding contributions to machine learning and computer vision. Due to the lack of large-scale training datasets and hardware, LeNet-5 is not ideal for complex problems. In 2012, the AlexNet proposed by Alex Krizhevsky *et al.* won the image classification championship on the ImageNet training set, making the CNN a key research direction in computer vision. AlexNet uses the rectified linear unit (ReLU) instead of the sigmoid as the activation function, and it achieves good results and solves the problem of gradient disappearance when the network is deep [22]. At the same time, the use of the GPU-based Compute Unified Device Architecture (CUDA) greatly accelerates the training speed of neural networks. Based on the above advantages, AlexNet has been applied in defect detection, location and visual tracking of dynamic objects [85], [86]. In 2014, the GoogLeNet network proposed by Google [87] won the ILSVRC competition, and its error rate was lower than that of VGGNet proposed in the same year. Generally, the position and size of the same object in different images are greatly varied, and an accurate convolution operation is needed to recognize this type of object. To solve the problem exemplified by large convolution kernels, which usually tend to perceive global information, while small convolution kernels mainly capture local information, the idea of GoogLeNet is to use multiple convolution kernels of different sizes in the same layer to capture information, and this structure is called inception [88]–[90]. Due to the good performance of GoogLeNet in image recognition, it has also achieved good accuracy in robot target detection [91]. VGGNet achieved second place in the classification task of the ILSVRC competition in 2014 (first place was GoogLeNet) and first place in the positioning task. At the same time, the model has good generalization ability for use with other datasets, and VGGNet has proven that a deeper network can affect the recognition effect of the network to a certain extent [92]. Because of its simple structure and strong feature extraction ability, VGGNet has a wide range of application scenarios. It is often used in the backbone of target detection (Fast-RCNN, single-shot multibox detector (SSD), etc.) to extract features [93], [94] and for target detection of robot grasping [95], [96]. The ResNet deep residual network proposed in 2015 won first place in the classification task of the ImageNet competition [97]. Because of its simple and practical structure, many target detection, segmentation and recognition algorithms are completed on the basis of ResNet50 or ResNet101 [98], [99]. The residual design mainly solves the performance degradation problem of deep networks and reduces the computation through a long jump connection. Even if the number of model layers is very deep, it can ensure normal training. The SSD algorithm proposed in 2016 is improved on the basis of VGG-16 and uses a multiscale feature map to a priori detect and set a box

for target detection [100]. The entire process of SSD requires only one step, so its most substantial advantage is that it runs faster. First, dense sampling is carried out in different positions of the image according to different scales and aspect ratios, and then the features are extracted by a CNN and then directly classified and regressed [101], [102]. However, uniform density sampling will lead to the imbalance of positive and negative samples, which makes training more difficult and leads to a reduction in model accuracy. The You Only Look Once (YOLO) algorithm proposed in 2016 is a typical one-stage method for target detection; the core idea is to transform the object detection problem into a regression problem. The model can directly predict the bounding box and category probability from the input image by using a CNN structure [103]. The execution speed is fast, and very high detection accuracy can be achieved by using a regression method. From YOLOv1 in 2016 to YOLOv3 in 2018, the YOLO algorithm has continuously absorbed the advantages of similar algorithms (such as the feature pyramid network (FPN) and the Fast-Region-based CNN (RCNN)) and achieved higher detection speed and accuracy through its own continuous improvement and progress, which is more in line with the real-time requirements of the industry for the target detection algorithm compared with other algorithms [104]. As two algorithms proposed in the same year, SSD and YOLO algorithms have made outstanding achievements in the field of image and vision, and they have good performance in target recognition, location and capture strategy generation [104]–[110]. The greatest contribution of the RetinaNet algorithm put forward by Tsung-Yi Lin *et al.* in 2018 is the proposal of focal loss to solve the problem of class imbalance [111], thus enabling the algorithm accuracy to exceed the target detection model of the classic two-stage approach. Both one-stage and two-stage detection algorithms are proposed based on an anchor mechanism (e.g., Fast-RCNN, RetinaNet, YOLO, or SSD), and these anchors are mainly used to find the location of the box; however, all of these algorithms incur excessive costs because of the anchor mechanism. This mechanism has two disadvantages. First, many anchors will be generated in the network, and most of these anchors cannot box the target; therefore, most of them are negative samples, with few positive samples. This outcome leads to the problem of unbalanced positive and negative samples and consumes an extensive amount of computation. Second, the anchor mechanism introduces a vast amount of superparameters for the complex network, which often makes the adjustment of these superparameters very complicated and increases the complexity of the network. Based on the above problems, Hei Law *et al.* proposed an anchor-free mechanism in 2019, and it used the upper left corner and the lower right corner to predict the bounding box instead of implementing an anchor [112]. Fig. 2 lists the major improvement process of the CNN algorithm from 1998 to 2019 and illustrates the core structure of various improved algorithms. The recognition accuracy and operation speed of algorithms have greatly improved by these developments. To date, various improved algorithms based on CNNs continue to emerge and are one of the main research directions in the field of vision.

A. ROBOT GRASP POINT AND GRASP STRATEGY

To solve the problem of robot grasping angle prediction, Cheng and Meng [113] proposed a two-stage cascade training process solution. First, the neural network performs 20000 iterations to obtain the ability to locate the object, and some parameters in the network are frozen. Second, the scale factor of 1.14 (superparameter) is multiplied by the $\sin(\theta)$ and $\cos(\theta)$ of the ground truth value. Through these two cascaded training processes and 500 iterations, the network can obtain strong direction prediction ability. Zunjani *et al.* [114] found that robots need to predict the ideal matrix according to the intention of the object to achieve an optimal grabbing strategy. They input the object image and intention type metadata into the full connection layer of the CNN network, which will achieve the ideal rectangular prediction. Corona *et al.* [115] designed a hierarchy model composed of three CNNs for the problem of grasping deformable objects such as textiles, which can be trained by using synthetic images and real images. Through the three steps of object recognition, the first grabbing point and the second grabbing point, accurate grabbing of the object can be achieved. Gaona and Lin [116] proposed an estimator-based particle swarm (PS) optimization algorithm by a CNN for fast and robust reasoning of robot grasping points. The cost function of PS is mainly considered from two aspects: first, the CNN divides the grabbing features into good features and poor features; and second, a magnet mechanism is designed to make particles converge to the object center. The algorithm also includes a confidence factor to reduce misjudgment between the grabbing point and the nongrabbing point. Yamazaki [117] proposed a method to detect the grabbing point from irregular-shaped knitted fabrics. Combining the grabbing point detection with the shape classifier, a CNN is used to classify the shape and extract the feature vector of the detected object shape. Using this feature, the captured points are calculated as image coordinates, and the effectiveness of this method is proven.

A reasonable grasping strategy and grasping points are the basic requirements for the robot to grasp the target based on vision, and they correspond to the nondeformable object and the deformable object, respectively. An end-to-end deep learning model is constructed based on the CNN algorithm, and the images collected by the camera are input into the model to realize the reasonable output of the grabbing strategy and grabbing points. However, at present, there are two main problems. First, the image processing effect is poor if the noise is large, so image preprocessing and noise reduction are necessary to realize the grabbing strategy. Second, it is necessary to manually design reasonable label features to make the model achieve better results in the test set and practical applications.

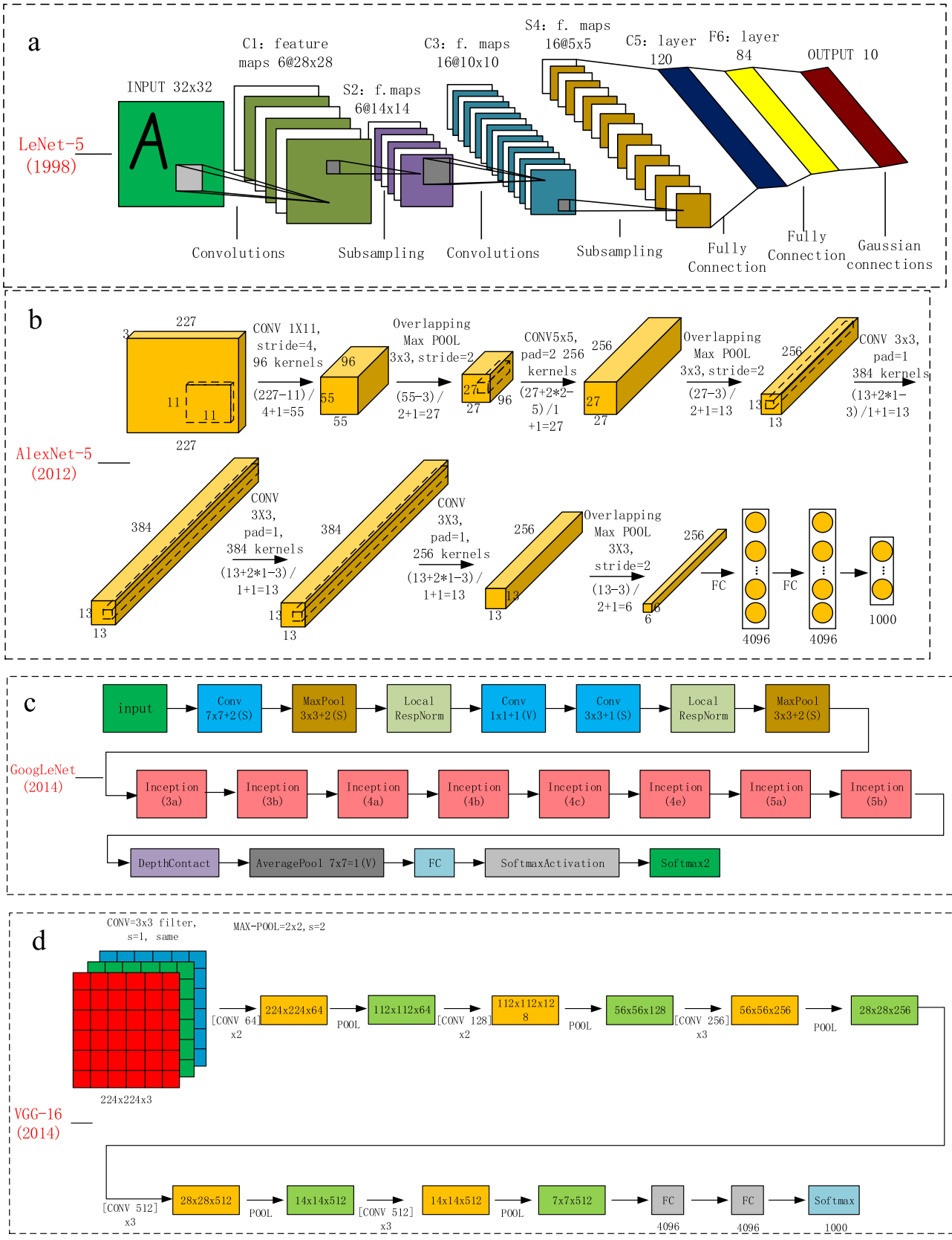
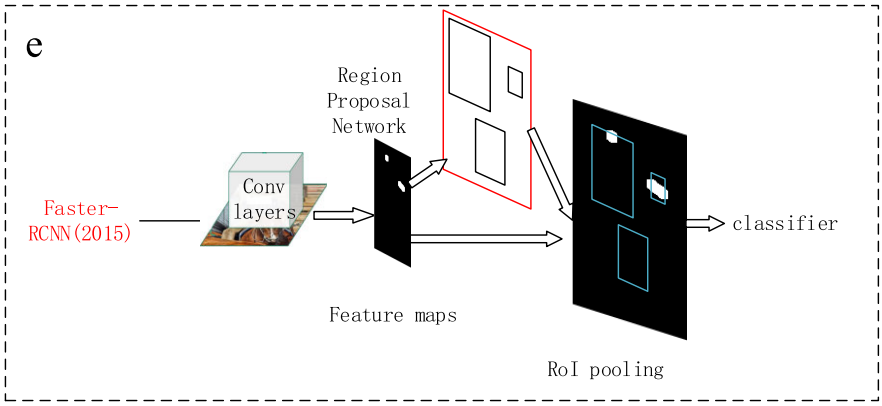


FIGURE 2. Development track of machine vision based on CNNs. (a) LeNet-5 [84]. (b) AlexNet-5 [22]. (c) GoogLeNet [87]. (d) VGG-16 [92]. (e) Faster-RCNN [94]. (f) ResNet [97]. (g) YOLO [103]. (h) SSD [100]. (i) YOLOv3 [119]. (j) RetinaNet [111]. (k) CornerNet [112].



f

Resnet (2015)

Layer name	Output size	18-layer	34-layer	50-layer	101-layer	152-layer
Conv1	112x112	7x7, 64, stride 2				
Conv2_x	56x56	3x3 max pool, stride 2				
		$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
Conv3_x	28x28	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times8$
Conv4_x	14x14	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times23$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times36$
Conv5_x	7x7	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times23$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times36$
	1x1	Average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

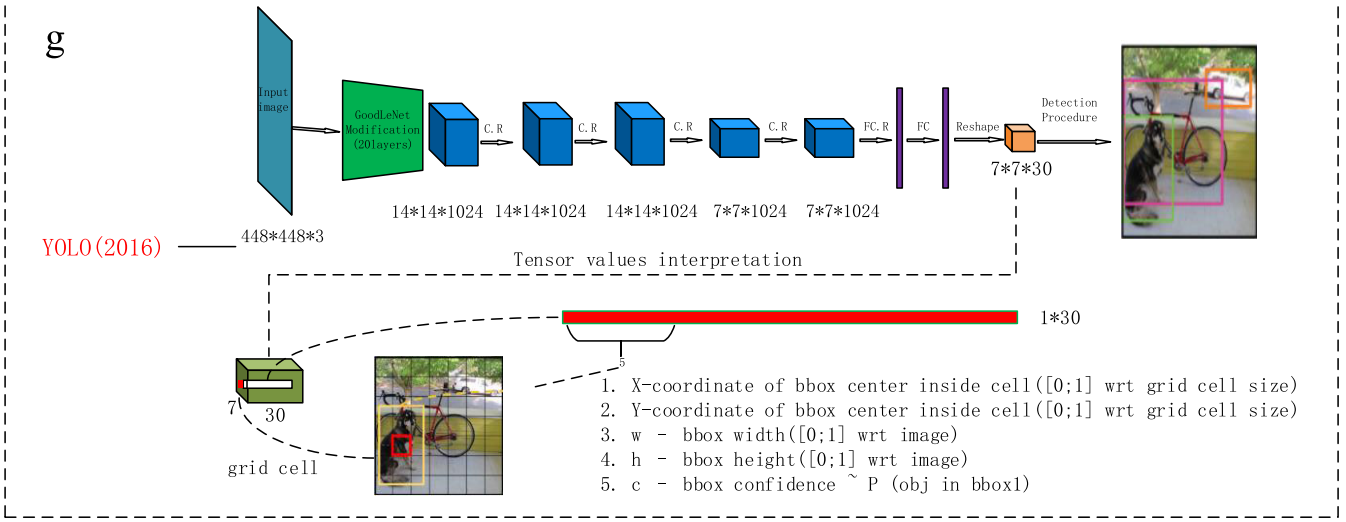


FIGURE 2. *(Continued.)* **Development track of machine vision based on CNNs. (a) LeNet-5 [84]. (b) AlexNet-5 [22]. (c) GoogLeNet [87]. (d) VGG-16 [92]. (e) Faster-RCNN [94]. (f) ResNet [97]. (g) YOLO [103]. (h) SSD [100]. (i) YOLOv3 [119]. (j) RetinaNet [111]. (k) CornerNet [112].**

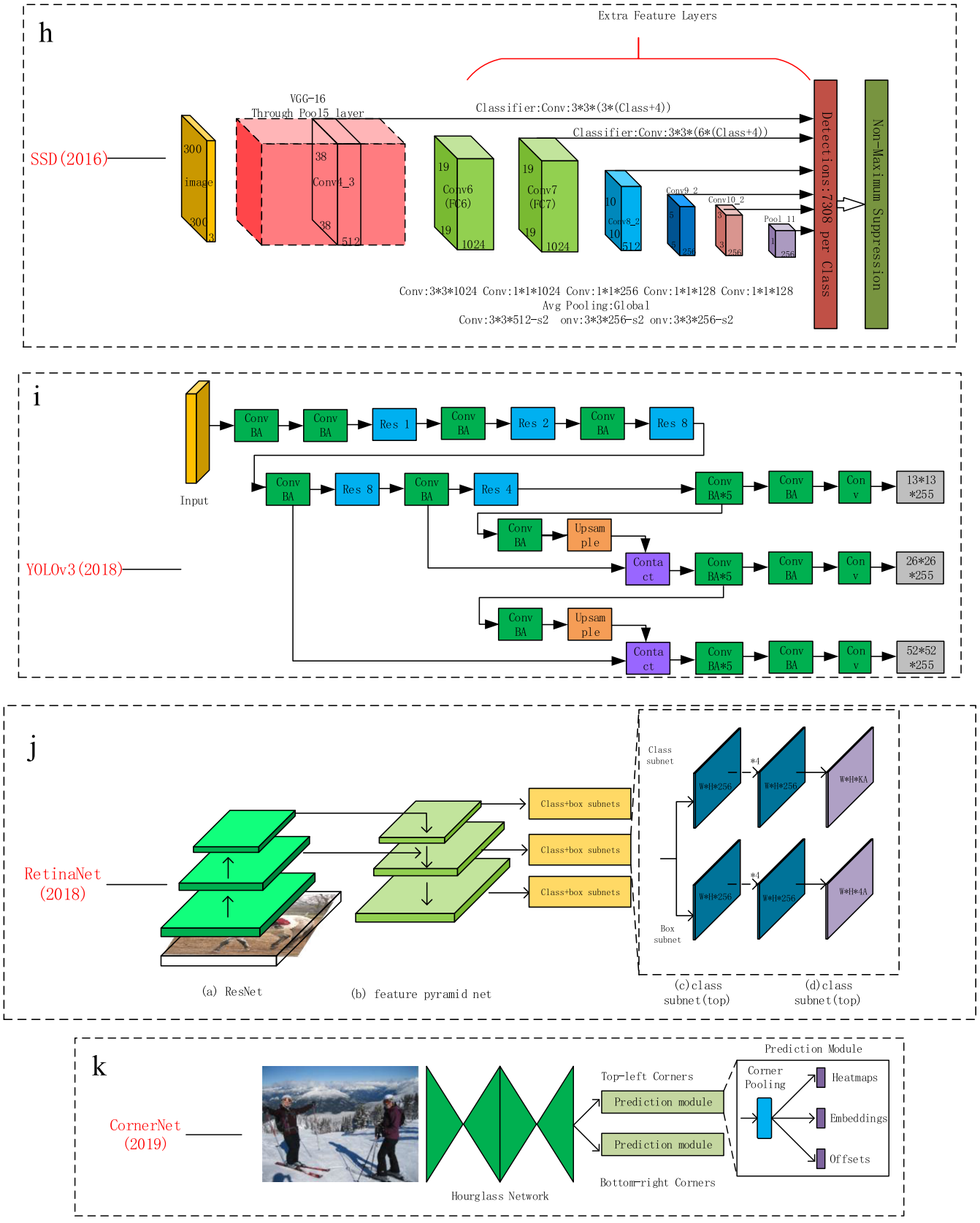


FIGURE 2. ***(Continued.)*** **Development track of machine vision based on CNNs. (a) LeNet-5 [84]. (b) AlexNet-5 [22]. (c) GoogLeNet [87]. (d) VGG-16 [92]. (e) Faster-RCNN [94]. (f) ResNet [97]. (g) YOLO [103]. (h) SSD [100]. (i) YOLOv3 [119]. (j) RetinaNet [111]. (k) CornerNet [112].**

B. MULTITASK COOPERATIVE OPERATION

Haochen *et al.* [118] established a neural network for object recognition, location and attitude detection using the CNN algorithm. Pose detection is treated as a classification problem in this model, and multiple tasks, such as recognition and location are combined at the same level to achieve good performance in printed circuit board (PCB) datasets. Chen *et al.* [120] introduced the grasp path based on CNN to predict multigrasp tasks, mapped the grasp candidate options to the grasp path and generated the mapping capture, and the deviation between them is taken as the estimation error of back-propagation. Experiments on the datasets and real scene show that this method can improve the detection accuracy and be well extended to the occluded objects.

Complex system engineering is required to realize target grabbing based on vision, which involves a series of steps, such as recognition, positioning and pose detection, that are all in the field of image processing. Therefore, building a model based on CNNs to realize the real-time processing of multiple tasks and the probability ranking of output results is an important research direction.

C. OBJECT 3D SHAPE CONSTRUCTION

Roy *et al.* [121] used CNN (VGG16) to classify the objects grasped by the manipulator into four categories, cylindrical, spherical, cubic and conical, and then generated four different grasping strategies. This method achieves 93% accuracy in real-time object recognition and grasping. Yan *et al.* [122] introduced a deep geometry-aware grasping network (DGGN), which divides learning into two steps. First, the 3D shape model and scene are generated and reconstructed by RGB-D, and then the construct of geometry representation is acquisition. Second, the results are predicted by learning the geometry perception representation within the model. Satish *et al.* [123] learned the deep strategy from the comprehensive training datasets of a point cloud and used the analysis algorithm of a random noise model to randomly sample, grab and reward the domain to explore how the distribution of comprehensive training examples affects the speed and reliability of the robot learning strategy. A comprehensive data sampling distribution is proposed in this paper, which combines the grabbing sample from the strategy action set and the guide sample from the supervisor with high robustness. This method is used to train the robot grasping strategy based on a full convolution network architecture, which evaluates millions of grasping options in four degrees of freedom (three-dimensional position and plane direction). The experimental results show that CNN based on full convolution grasp quality (FC-GQ-CNNs) has better speed and reliability. Liang *et al.* [124] proposed an end-to-end grabbing evaluation model (PointNetGPD) to solve the problem of grabbing configuration directly from the point cloud map. The model is lightweight and takes the original point cloud image as the input, which can directly process and evaluate the 3D point cloud image inside the grabber. Even if the point cloud is very sparse, it can capture the complex geometry of the contact area between the grabber and the object.

Dividing objects into several categories according to their general shapes and then generating different grabbing strategies based on the category is a good approach, but the generality is poor. It is very important to realize the 3D reconstruction of the object based on vision. The RGB-D image collected by the depth camera is input into the deep learning model to realize 3D reconstruction, which can improve the success rate and speed of the capture.

D. MOTION PATH

To solve the problem of dexterous hand grasp force when performing tasks, Sun *et al.* [125] proposed a motion reproduction system based on several motion and depth data. At the same time, CNN is used to estimate the motion instructions of the depth image, and the force data is saved to generate the label training datasets. Deng *et al.* [126] proposed a learning framework combining semantic reach-to-grasp (RTG) with trajectory generation, aiming for the successful realization of semantic reach-to-grasp in unstructured environments. First, an object detection model based on deep learning is used to detect the interested objects, and the trained network based on the Bayesian search algorithm is used to find the most successful grabbing configuration from the object segmentation image. Second, a model-based trajectory generation method is designed for the robot's arrival motion, which is inspired by the theory of the human internal model to generate the trajectory satisfying the constraints; the effectiveness of this method has been proven.

Different grasping forces are the key to grasping different objects successfully. Associating scene images with force data and using the CNN model to complete training can improve the adaptability of the robot grasping force. The combination of a CNN and the traditional machine learning algorithm can realize the sorting of several options and output the optimal value.

E. REAL-TIME MOTION

González-Díaz *et al.* [127] proposed a real-time solution to the problem of grasping action in self-centered video. First, aiming to address the problem of deciding which object will be grabbed and when to trigger the grabbing operation from a given classification, this paper determines the grabbing area based on the gaze-guided CNN focusing on an object. Second, the fixed sequence obtained is noisy because of distraction and visual fatigue, and gaze is not always reliable for the object of interest. To solve this problem, video-level annotation is used to represent the object to be grabbed, and a loss function is used in a deep CNN. To detect when a person removes an object, the prediction ability of long- and short-term memory networks is used to analyze gaze and visual dynamics. The results show that this method has better performance than other methods in real datasets. Farag *et al.* [128] proposed a real-time object detection algorithm based on a selective flexible assembly

TABLE 2. Comprehensive performance comparison of mainstream CNN models.

Year	Author	Model	Datasets	Key Points
1998	YANN LECUN et al.[84]	LeNet-5	MNIST	LeNet-5 is one of the earliest CNNs and the origin of a large number of neural network architectures. However, it does not perform well on complex issues.
2012	Alex Krizhevsky et al.[22].	AlexNet	ImageNet ILSVRC-2010	With the development of AlexNet (8-layer neural network), the CNN has become a key research direction in computer vision, and the top-5 error rate was reduced to 16.4% for the first time in the 2012 ILSVRC competition.
2014	Christian Szegedy et al.[87]	GoogLeNet	ImageNet ILSVRC-2014	GoogLeNet's greatest contribution (22-layer neural network) is to propose the Inception Architecture and cancel the full connection layer to decrease the number of parameters and thus reduce the top-5 error rate to 6.7% in the 2014 ILSVRC competition.
2014	Karen Simonyan et al.[92]	VGGNet	ImageNet ILSVRC-2014	VGGNet (19-layer neural network) ranked second in the ILSVRC classification task in 2014, with a top-5 error rate of 7.3%, but it won the first place in the positioning task.
2015	Kaiming He et al.[97]	ResNet	ImageNet	ResNet (152-layer neural network) has fewer parameters than VGGNet, but its performance and training speed are greatly improved. The ResNet algorithm won the first place in the classification task of the ILSVRC competition in 2015, and the top-5 error rate is 3.57%.
2015	Shaoqing Ren et al.[94]	Faster-RCNN	Pascal Voc2007+2012	Region proposal networks are used to generate candidate regions and train an RPN and Fast RCNN to share a convolution layer, which greatly improves the detection speed of the network. However, it takes a considerable amount of time to generate candidate regions, which also affects the detection performance.
2016	Joseph Redmon et al.[103]	YOLOv1	Pascal Voc2007	The YOLO algorithm solves the speed problem in deep learning and has strong generalization ability.
2016	Wei Liu et al.[100]	SSD	Pascal Voc2007	The SSD algorithm is much better than YOLOv1 in accuracy and speed, and SSD directly uses the convolution layer in the last layer to extract the detection results of different feature maps.
2018	Joseph Redmon[119]	YOLOv3	Pascal Voc2012	YOLOv3 mainly integrates some good schemes and achieves good results and improves the accuracy under the premise of ensuring the speed advantage, especially strengthening the ability of small object recognition.
2018	Tsung-Yi Lin et al.[111]	RetinaNet	MS COCO	To solve the problem of attention imbalance, RetinaNet proposes a new focal loss function, which adds a weight that depends on probability to adjust the cross entropy loss.

manipulator (SCARA) for robot grasping and positioning in industrial assembly lines. The motion of a SCARA robot is composed of two parts: target detection based on deep learning and position measurement based on edge detection.

Real-time performance is very important for robot grasping, and good real-time performance can guide the robot to realize the recognition, positioning and grasping of dynamic objects. Based on the CNN-AlexNet, the researchers used the transfer learning method to establish a target detection model, knowledge and statistics superimposing network (KSSNet), which achieved a 100% success rate in target detection, location and capture.

The target detection, recognition, location and grasp strategy generation involved in robot visual grasping are all in the field of image processing, and the CNN has strong performance in such a field. Therefore, the CNN is widely used in the field of visual grasping and has a good effect. As shown in Table 2, from the proposal of the first full-fledged CNN in 1998 to the RetinaNet network in 2018, deep learning has been developing rapidly, and the accuracy and speed have greatly improved. At present, CNN research is generally based on supervised learning, which needs a large number of labeled datasets for model training. However, with the continuous development of computer vision, it is increasingly difficult to obtain valuable labeled datasets, and most of the labeled data are calibrated by humans, which greatly increases the consistency, difficulty and cost of labeled data acquisition. Because of the above reasons, neural networks that do not need or rely on labeled data have become a worldwide priority research direction. These algorithms need little or no labeled data or do not need manually labeled data, which greatly reduces the need for human intervention in the model training process.

IV. DIFFERENT MACHINE VISION ALGORITHMS WITHOUT LABELED DATA

Supervised learning (especially CNN) has made remarkable achievements in the field of vision after nearly ten years of rapid development, but it has also attracted some criticism. Label data are very important for the training of supervised learning, and the label data of traditional supervised learning need to be labeled manually, which not only leads to the high cost but also appears less intelligent. With the rapid increase in artificial intelligence applications, especially machine vision, researchers hope to achieve model training without a large number of artificial annotation datasets. Unsupervised learning can complete training based on unlabeled data, so it can realize object recognition and grasping very intelligently [129]–[132]. Self-supervised learning is a special case of supervised learning that does not need a large number of manually labeled datasets to realize model training [133]–[137]. Reinforcement learning is learning an

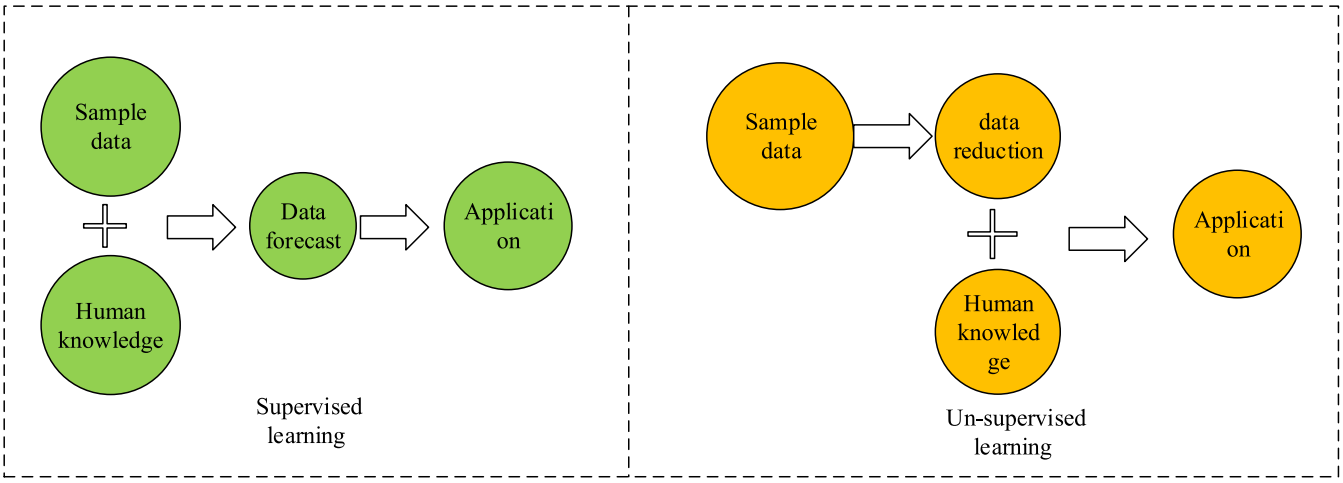


FIGURE 3. Comparison between supervised and unsupervised learning.

optimal policy, which can make the agent perform an action according to the current state in a specific environment to obtain the maximum return. Reinforcement learning was not the focus in the early stage, but with Google's successful application in Atari and Go games, this branch of machine learning has attracted much attention. With the development of deep reinforcement learning, researchers have combined it with machine vision [138]–[142] in the hope of removing the need for labeled data and artificial means to achieve intelligence.

A. UNSUPERVISED LEARNING

Unsupervised learning is one of the most difficult and important problems in machine vision and machine learning. Many researchers believe that learning from a large amount of unlabeled data can help solve problems concerning intelligence and the nature of learning. In addition, unsupervised learning has practical application value in many fields of computer vision and robot grasping because of the low cost and ease of collecting unlabeled image datasets. It is easy to see why researchers think unsupervised learning is more intelligent through Fig. 3.

Unsupervised learning can be regarded as a branch of traditional machine learning. Dimension reduction and clustering are well-known unsupervised learning methods, but traditional unsupervised learning is significant in data analysis. With the rapid development of deep learning and the difficulty of label data acquisition, the combination of deep learning and unsupervised learning has gradually become a reasonable research direction. Lenz *et al.* [143] designed a system to achieve robot grasping from RGB-D images by using deep learning. This method can label data without manual work. To quickly select the grabbing options, this paper proposes a two-step series deep learning network. The first network quickly selects several grabbing strategies with high probability, and the second network takes the output of the first network as the input and calculates the optimal grabbing strategy. Ardon *et al.* [144] proposed a method to detect and extract multiple grabbing signals through visual input. This method does not need to manually define label data but collects their distribution, location and executable grasp label data from 1269 objects to obtain their relationship with input. The model not only learns to grasp the object but also has better generalization ability in different environments based on these datasets. Detry *et al.* [129] designed a new method of object recognition and grabbing based on the reduced dimension and clustering algorithm and let the model learn from a group of grabbing examples to improve the generalization ability. Unsupervised learning has the advantage of object classification based on multimodal information because it does not require label data [130]–[132]. However, due to the inherent defects of vision and the development of sensor technology, it has become a hot direction to integrate the information of vision, tactile feedback and hearing to help the robot achieve accurate recognition and grasping of the object.

Because unsupervised learning does not need labeled data, it has good generalization and can extend some features of known objects to similar objects to achieve the grasping of unknown objects. Alternately, as a pretraining method, unsupervised learning plays an important role in the success of deep neural networks.

B. SELF-SUPERVISED LEARNING

Self-supervised learning mainly uses pretext tasks to mine its own supervision information from large-scale unlabeled datasets, and the training of the neural network is based on constructed supervision information to learn valuable representations of downstream tasks. As shown in Fig. 4, the assessment of self-supervised learning ability is mainly completed through a pretraining-fine-tuning mode. First, the network is trained by pretext from a large number of unlabeled datasets (automatic construction of supervision information in the data), and the pretraining model is

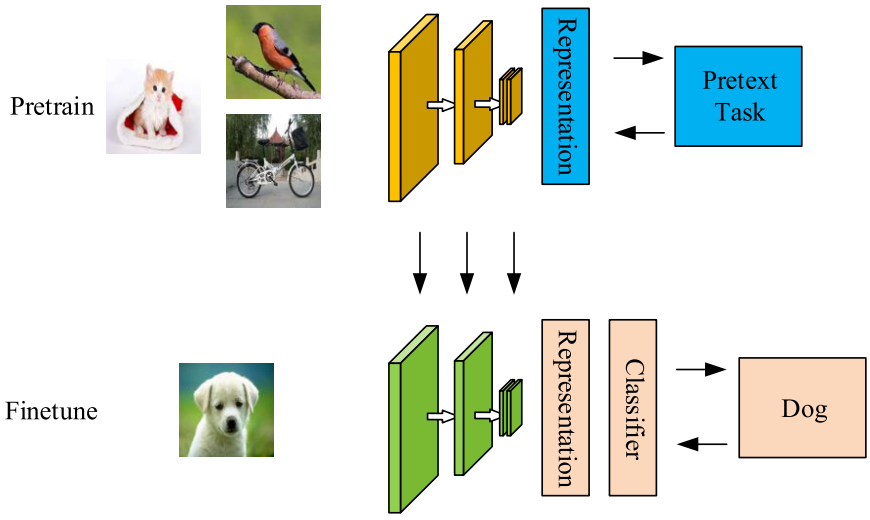


FIGURE 4. Process of self-supervised learning.

obtained. Then, for the new downstream tasks, the algorithm adopts a method similar to supervised learning, which can obtain parameters through transfer learning and then fine-tune them. Thus, the ability of self-supervised learning is mainly reflected by the performance of downstream tasks.

Nguyen *et al.* [133] adopted a self-supervised learning method in which the training datasets are automatically marked by the model. In this paper, a continuous level neural network is proposed to reduce the runtime of the grabbing task by eliminating the nonextractable samples from the reasoning process, and the network can estimate 18 grabbing postures and classify 4 objects at the same time. The experimental results show that the accuracy of the network is 94.8% for grasping posture estimation and 100% for object classification within 0.65 seconds. Murali *et al.* [134] proposed a new method to accelerate the self-supervised learning process and mapped visual information to a high-level and high-dimensional movement space to realize the training strategy of the model. Florence *et al.* [135] used self-supervised correspondence to improve the generalization ability and sample efficiency of visually driven strategy learning. Yang *et al.* [137] proposed a critical policy form to design a deep learning method for a new problem named "grasping the invisible," where a robot is tasked with grasping an initially invisible object via a sequence of nonprehensile (e.g., pushing) and prehensile (e.g., grasping) actions. In this paper, the Bayesian algorithm and classifier model are combined, the self-supervised method is used to train the motion critic and the classifier in the interaction between robot and environment, and a good success rate is achieved in the experiment.

Self-supervised learning is a type of unsupervised learning that realizes the supervised training through the automatic generation of labels. Self-supervised learning not only achieves high accuracy and speed in object recognition classification and grasping attitude estimation but also has good generalization performance.

C. REINFORCEMENT LEARNING

Reinforcement learning has achieved good results in many decision-making fields, especially in the game field, which has reached or even surpassed the human level. However, it is not widely used in the field of machine vision, which may be because vision does not seem to directly correspond to a decision-making environment or interpretable action steps similar to that seen in games. Even so, because reinforcement learning does not need label data and works similar to human beings, it has aroused researchers' enthusiasm to apply it to the visual field. Fig. 5 lists several mainstream reinforcement learning algorithms and their core structures. From the initial Q-learning to the recently popular deep reinforcement learning, it shows that reinforcement learning is developing rapidly. The training process of reinforcement learning with little or no human intervention has fascinated many researchers. As early as 2014, the Google DeepMind team applied deep reinforcement learning to the attention mechanism [145]. In 2018, Yu *et al.* [146] applied deep reinforcement learning to image repair and achieved good results. James *et al.* [147] proposed a new benchmark and learning environment for challenging robotic learning: RLBench, which is designed to accelerate progress in the field of visually guided manipulation. The above research lays a foundation for the application of deep reinforcement learning in machine vision to guide robots in recognizing and grasping objects.

The model-free deep reinforcement learning proposed by Zeng *et al.* [148] found that it was feasible for robots to learn some cooperative grasping strategies. By training two complete convolution neural networks, the first is from vision mapping to action, and the other is used for robot grasping. These two networks are jointly trained in the Q-learning

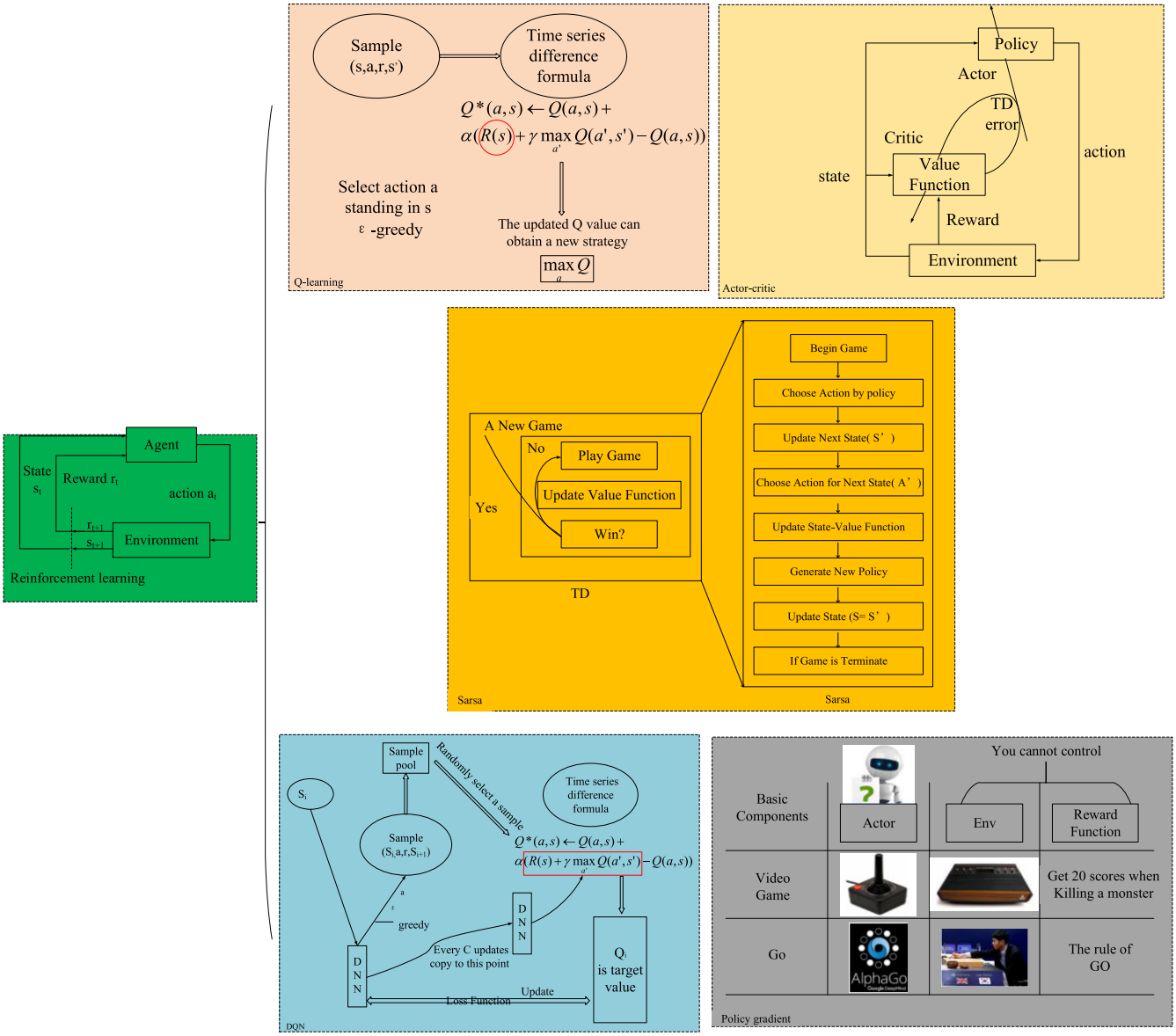


FIGURE 5. Mainstream reinforcement learning algorithms.

framework, and self-supervised training is completely carried out by a trial-and-error method. In the trial-and-error method, the successful completion of the action can be rewarded, and the learning strategy can promote action in this way. Wang *et al.* [149] proposed a method combining Q-learning and a visual servo to solve the grasping problem of wheeled mobile robots and realized the robust grasping of robots. Gu *et al.* [150] proposed a new deep reinforcement learning algorithm based on deep Q-functions nonstrategy training that can adapt to complex 3D operation tasks.

Breyer *et al.* [151] proposed an object grabbing algorithm based on reinforcement learning. In this paper, the image collected by a depth camera is mapped to the closed-loop control strategy of motion command, and several different methods are compared to ensure the rationality of the algorithm. Katyal *et al.* [152] used deep reinforcement learning to make a robot immune to the changes of manipulator or environment and achieve robustness to changes of the environment without clear prior knowledge and fine kinematics knowledge of the human arm structure and without careful hand-eye calibration. Ghadirzadeh *et al.* [153], to solve the inherent delay in motion perception processes, proposed a data-based deep predictive policy training (DPPT) framework, which maps the observed images to a series of motion activation. The system consists of three subnetworks, namely, the perception, strategy and behavior superlayer, and each task is trained by strategy search reinforcement learning. Nguyen *et al.* [154] compared the performance of proprioceptive/kinesthetic input and original visual input in the framework of deep reinforcement learning and found that the

TABLE 3. Analysis of advantages and disadvantages of unlabeled data algorithms.

Algorithm	Nature	Advantages	Disadvantages	Main Application Field
Unsupervised Learning	Statistical calculation	No labeled data are required.	Inefficiency	Robot[163-166] Computer vision[167-171] Data analysis[23, 172, 173]
Self-supervised Learning	Automatic generation of label data	There is no need to label manually.	It is difficult to obtain appropriate pretext tasks.	Robot[136, 174-178] Computer vision[135, 136, 178-181] Natural language processing[182, 183]
Reinforcement Learning	Learning from delayed reward	The task is transformed into a Markov decision problem and without labeled data.	The algorithm is difficult to converge and needs to set up the reward function and network structure.	Robot[147, 151, 156, 184] Computer vision[185-189] Game[190-192] Autonomous driving[185, 186]

former greatly improved the performance of the agent compared with the latter. Beltran-Hernandez *et al.* [155] proposed a reinforcement learning model based on a strategy search algorithm, which shows good robustness in the generalization from a simple shape object to a complex one. Li *et al.* [156] put forward a type of reinforcement learning strategy for the operation and grasping of a mobile manipulator to solve the problem of human-like mobile robot learning complex grasping action in a human environment. This strategy reduces the complexity of visual feedback and can deal with the changing operation dynamics and uncertain external interference. Miljković *et al.* [157] proposed a robot intelligent visual servo controller based on reinforcement learning, developed two different time difference algorithms (Q-learning and SARSA) and combined them with a neural network, and then tested them in different visual control scenes. Compared with the traditional image-based visual servo system, the algorithm proposed in this paper has better performance for low-cost visual system manipulators.

Bousmalis *et al.* [158] studied how to extend the random simulation environment and region adaptive method to the training grabbing system to grab new objects from the original monocular RGB image. By using only unlabeled real-world data and the grasp generative adversarial network (GraspGAN) algorithm in this paper, the grabbing performance is similar to that obtained with 939,777 labeled real-world samples. James *et al.* [159] proposed a method called random to canonical adaptation networks (RCANs) to solve the problem of difficult acquisition of real label data in the field of robotics, which can achieve real-world effects by using nonreal-world data. The paper trained a visual-based closed-loop grabbing reinforcement learning agent in simulation and then transferred it to the real world, achieving very good performance and proving the effectiveness of this sim-to-real method. Hellman *et al.* [160] proposed a contextual multiarmed bandit (C-MAB) reinforcement learning algorithm that integrates vision and tactile feedback to realize the closure function of a transparent and easily deformable zipper bag. Platt [161] took tactile feedback as the main information source and combined part of the visual information to achieve better performance in the experiment of grasping plane objects. Merzic *et al.* [162] used model-free deep reinforcement learning to combine vision and tactile feedback to generate a control strategy. The results show that tactile feedback can significantly improve the grasping robustness of objects with attitude uncertainty and complex features.

Traditional reinforcement learning has the limitation of a small action space and sample space, and it is usually used in a discrete situation. However, being more complex and closer to the actual situation of the task often yields a large state space and continuous action space. When the input data are images or sound, it often has a high dimension, and the traditional reinforcement learning has difficulty addressing it. The deep reinforcement learning combines the deep learning and reinforcement learning to make the two complementary and achieve better performance.

As shown in Table 3, the three types of algorithms do not need manually labeled data, which has great advantages over the traditional CNN algorithm. The three algorithms not only have achieved outstanding results in their respective fields but also achieved good performance in the fields of vision and robotics. The clustering algorithm in unsupervised learning is widely used in the field of vision. Through the fusion of the clustering algorithm and deep learning, it can realize the accurate recognition and classification of the objects and the recognition of the robot's running posture and trajectory, but it also has the disadvantage of low efficiency. The data are easy to obtain, but the labeling cost is high, so researchers hope that supervised learning can train a model with good generalization performance by using few labeled datasets. However, if a good feature expression can be obtained, it will be conducive to the fine-tuning of downstream tasks and multitask training, which is also the core idea of self-supervised

learning. Self-supervised learning takes unlabeled datasets as input, automatically constructs labels through the structure or characteristics of the data itself, and then carries out training similar to supervised learning. Based on the above advantages, self-supervised learning has achieved good training effects and high-precision target recognition and positioning in the field of vision, but it has the problem of label rationality. The principle of reinforcement learning makes it not dominant in the field of vision and target detection and recognition. The integration of reinforcement learning and deep learning is the mainstream research direction and has achieved good performance in many decision-making fields. The visual perception model based on deep reinforcement learning can predict all possible actions in the current state when only the original image is input. Therefore, deep reinforcement learning has some research achievements in the action conditional video prediction task. In addition, the deep reinforcement learning based on the strategy gradient (e.g., trust region policy optimization (TRPO), generalized advantage estimation (GAE), stochastic value gradient (SVG), and asynchronous advantage actor-critic (A3C)) realizes the behavior control of the robot and is verified in the actual application scenario. The low sampling efficiency of reinforcement learning makes training difficult, and a reasonable reward function and network structure need to be designed to achieve better results.

V. FUSION OF VISUAL AND TACTILE FEEDBACK

After years of development, object recognition and location based on machine vision has achieved great success, which lays a solid foundation for research on robot grasping. At present, representative object detection algorithms (e.g., Faster-RCNN [94], SSD [100], and YOLOv3 [119]) can quickly identify and locate objects, but relying on precise location alone cannot make the manipulator achieve stable grasping in complex environments. From the view of people's own experience in grasping objects, a series of attributes, such as the hardness and quality of objects, are needed to ensure the success of grasping. In addition, the accuracy of machine vision is greatly affected by the surrounding environment. When the robot is applied in a variable light source environment, such as life scenes, the robustness of machine vision is low [193]–[197], and it is difficult to achieve stable grasping only by machine vision when the object can easily deform [198]. To solve these problems, researchers in the field of robotics and vision consider adding additional tactile sensors to the robot to achieve more stable grasping. The research direction is mainly divided into single tactile object perception [199]–[203] and vision-tactile fusion [204]–[207] object recognition and grabbing.

A. TACTILE FEEDBACK

For human beings, tactile feedback is the second most important signal receptor after vision, which plays an important role in life. With the development of tactile sensor technology [208]–[212], researchers hope that robots can also have the same tactile perception ability as humans and further realize intelligence. It is a good idea to apply tactile technology to the robot alone, which can avoid the fusion of different signals and improve the processing speed of the system. As shown in Fig. 6, Sundaram *et al.* [213] proposed a low-cost and high-robustness tactile glove, which weaves the array pressure sensor on the surface of the flexible glove and then wears it on the hands of the experimenter to collect the tactile data of different objects. By touching different objects, different pressure point cloud images are obtained and introduced into the neural network for training to realize object recognition and weight estimation without vision.

Rasouli *et al.* [199] developed a neural morphological system for tactile pattern recognition, aiming to address the problem of low efficiency and capability of artificial tactile sensors. The system achieved 92% classification accuracy in a texture recognition task and proved that there is a tradeoff between response time and classification accuracy. Ward-Cherrier *et al.* [200] studied the development of the gripping platform Gr2, which demonstrated the reorientation of the grasped object through active tactile manipulation and used a new tactile sensor for tactile manipulation. The active tactile manipulation proposed in this study is modeless and can be used to study the operation principle of a dexterous hand. Bimbo *et al.* [201] proposed a method to locate the grabbed object in the robot's hand, which includes calculating the covariance of the pressure data of the tactile sensor and the eigen basis vector from the main axis. Liu *et al.* [202] regarded tactile data as a time series, used a dynamic time warping method to evaluate its difference, and proposed a joint kernel sparse coding model to solve the representation and classification of tactile data. Bhattacharjee *et al.* [203] used the first two seconds of force, heat, and motion sensing data collected by a robot in a real environment to solve the impact of the surrounding environment on tactile perception when the robot works in a human environment (such as a home), and data-driven approaches to the problems of various tactile perception performances (neighbor, SVM, hidden Markov model, and long short-term memory) have been characterized. The results show the value of multimodel tactile perception and data-driven methods for short-term contact tactile perception.

The research history of machine tactile feedback is relatively short and clearly lags behind machine vision. This lag is mainly due to the backward hardware performance of tactile sensors and the confusion of sensor types. Alternately, the lack of research content and methods of tactile technology also causes the lag of tactile research. With the rapid development of intelligent robots, tactile feedback has gradually attracted the attention of researchers, and there are many fruitful research achievements. At present, the research of tactile technology mainly focuses on three areas: 1. Hardware improvement of tactile sensors is needed. Through the improvement of hardware, the sensitivity of the sensor can be improved, and multiple types of data can be collected at the same time (e.g., temperature, pressure, friction, etc.). 2. Based on the sense of tactile feedback, the precise

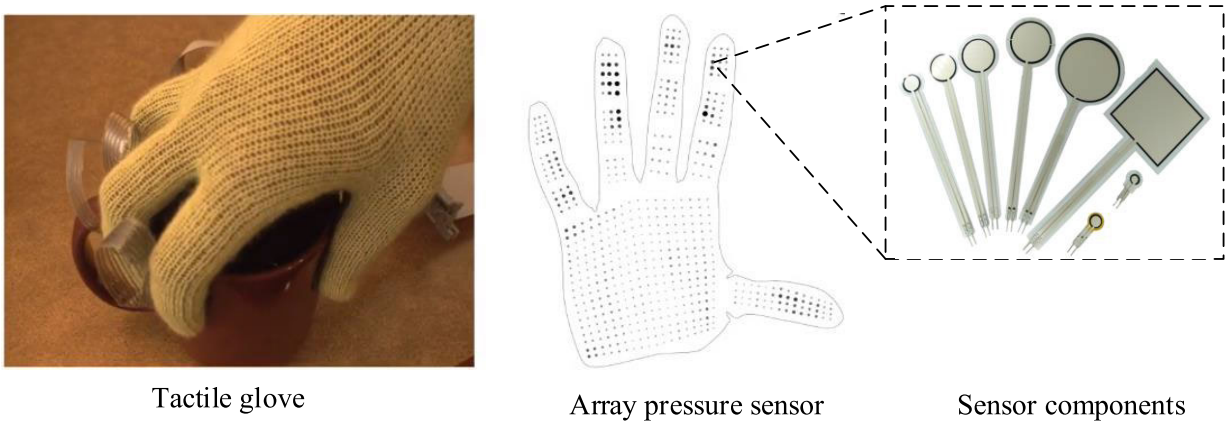


FIGURE 6. Tactile sensor schematic.

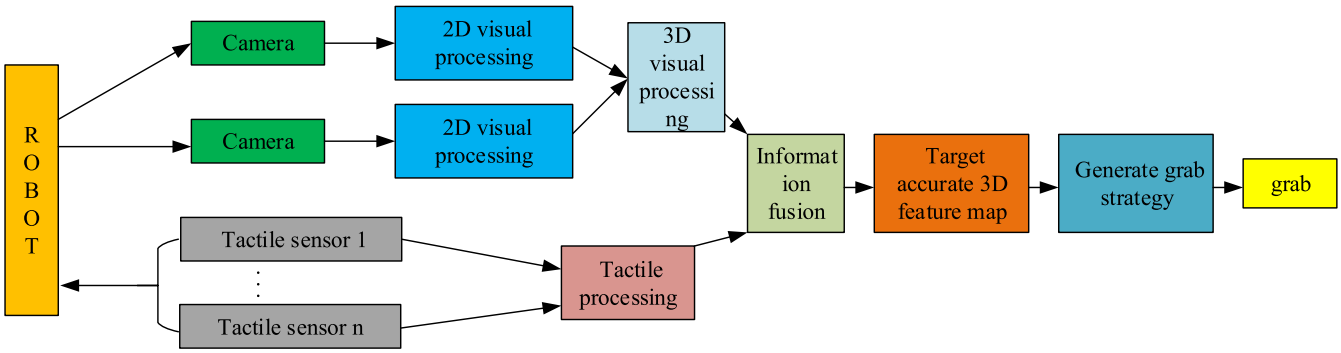


FIGURE 7. Framework of visual-tactile fusion for object recognition.

extraction of object features can be achieved, and then the model-free stable operation of objects (e.g., grasping, classification, recognition, attitude estimation, etc.) can be achieved to improve the generalization of the tactile feedback. 3. The tactile feedback and deep learning are combined to realize the acquisition and training of tactile datasets, and then the deep neural network (DNN) realizes the weight perception, grasp and classification of objects. Tactile feedback is second only to vision in information perception, but its research and application are much worse than those of vision, mainly due to the poor universality and reliability of tactile feedback. The application of tactile technology to multisensor sensing systems to realize complementary information perception is a reasonable future research direction.

B. FUSION OF VISION AND TACTILE FEEDBACK

The integration of vision and tactile feedback helps the robot to achieve better grasping, which is also more in line with human expectations for the robot. However, the research time of robot tactile feedback is relatively short, and many types of sensors and multisensor data fusion approaches are involved, which leads to the difficulty of tactile research, which is scattered and unsystematic. As shown in Fig. 7, the object recognition and grabbing system based on visual-tactile fusion is generally divided into four steps. First, 2D vision processing technology is used to determine the object position and boundary area, and then 3D vision is used to determine the object's center of mass as the starting point of tactile detection. Second, tactile exploration is carried out for features and positions (e.g., pits, holes or occluded areas) that are hard to determine by vision to further determine the object surface features. Third, the information collected by vision and tactile feedback is fused to generate accurate 3D point cloud images. Fourth, an appropriate grasping strategy is generated to guide the robotic arm to complete the object grasping based on the visual centroid and tactile features.

Calandra *et al.* [204] studied how robots learn to use tactile information for iterative operations to effectively adjust their grasping strategy. In this paper, an end-to-end action condition model is proposed to learn the grasping strategy from the original visual-tactile data. Guo *et al.* [205] proposed a method of vision-tactile combination based on deep learning for robot grasping detection, and experiments show that tactile data is helpful for deep learning to learn better object characteristics of robot grasping detection tasks. Li *et al.* [206] designed a sliding detection algorithm using the GelSight tactile sensor and the camera installed on the side of the gripper without knowing the physical

parameters of the object in advance. Using the image sequences collected by two sensors, a DNN is trained to classify the grabbed objects and evaluate the stability of the grabbing process. Garg *et al.* [207] proposed an adaptive grasping method based on tactile and visual feedback. This method combines model-based partially observable Markov decision process (POMDP) planning with simulation learning, which has strong robustness under uncertainty, strong generalization ability and fast execution ability for multiple objects. Wang *et al.* [214] proposed a new method to solve the problems of imprecise visual modeling and low tactile efficiency. Through the combination of vision and tactile feedback, as well as learning the prior knowledge of common object shapes from a large shape database, this method can effectively perceive the accurate 3D information of the object. Hogan *et al.* [215] proposed a regrasp control strategy using a tactile sensor to adjust the local grasping action. In this paper, the local transformation of the actual search tactile value is used to determine the regrasp action to improve the quality of the grasp. The success rate of vision-tactile fusion is 70% higher than that of vision alone. Sun *et al.* [216] put forward two different tactile sequence models according to the advantages of vision and tactile feedback, proposed an object shape modeling method based on the direction description histogram features, and then considered the accuracy of the grasping point and the rapid planning of hand kinematics to achieve the grasping operation.

Through the research results of the above papers, it is found that the fusion of vision and tactile feedback improves the robustness and success rate of robot grasping, indicating that the introduction of tactile feedback provides a new direction for robot grasping research. The grasping of deformable objects has always been a difficult problem, and the operation needs to accurately estimate the real-time state of the objects. At present, the main research direction is machine vision, but the vision is very sensitive to occlusion, which is inevitable when the robot moves. Compared with vision, tactile feedback has strong robustness, so the addition of tactile feedback can solve this problem well. Sanchez *et al.* [198] proposed a modular pipeline that can track the shape of deformable objects online by coupling the tactile sensor with the deformation model and achieve robust grasping through the combination of vision and tactile feedback. Jain *et al.* [217] proposed a simulation-based learning method that uses a simulated five-fingered dexterous hand to train the deep visual motion strategy of various operation tasks and found that using tactile sensitive information can make the task with a highly occluded object exhibit faster learning speed and better asymptotic performance. Yu *et al.* [194] proposed a framework that fuses vision and tactile feedback to estimate the attitude and contact state of objects relative to the environment in real-time, aiming to address the application of inserting objects picked up by a suction cup into a small space. The fusion algorithm based on iSAM (an online estimation technology) is adopted in the framework to realize the fusion of robot motion measurement, geometric contact between object and container, and visual tracking. Finally, a data-driven method is proposed to deduce the contact information to achieve better grasping and placement. Santina *et al.* [218] proposed a data-driven autonomous grasping mechanism of a humanoid soft hand to improve the grasping performance. The nail of the humanoid soft hand is equipped with an inertial measurement device to detect the contact with objects. In this paper, a classifier is obtained by a deep neural network, which takes the visual information of the grasped object as input and predicts the grabbing action. Hang *et al.* [219] proposed a unified framework for grasping planning and hand grasping adaptation based on visual, tactile and proprioception feedback. The main purpose of the framework is to solve the problems of object deformation, sliding and external interference to achieve grasping.

TABLE 4. Comparative analysis of vision and tactile feedback.

	Advantages	Disadvantages	Future studies
vision	Efficient Noncontact Strong generality	Low precision Poor robustness	Acquisition of high-resolution images to achieve fast image processing and then improve the accuracy and robustness of operation.
Tactile	High precision No model	Poor versatility inefficiency	Develop new tactile sensors to realize accurate and fast perception of various types of information.
Visual-tactile fusion	Good versatility Rich functions and applications. High accuracy	It is difficult to process multivariate data. Lack of unified framework and evaluation indicators.	Based on the end-to-end deep learning model, the multivariate data processing is realized, and the general research framework is developed to promote better vision and tactile development.

As shown in Table 4, visual and tactile feedback are the basic ways for a human or robot to perceive the environment or target, and they are the key research fields of scholars across the globe. Because of their different principles and data structures, they both have advantages and disadvantages in perception and recognition, so combining them is a reasonable choice. The combination of vision and tactile feedback realizes complementary advantages, which can achieve more accurate object recognition, real-time state estimation, grasp force adjustment, 3D object modeling, grasping pose detection and other functions, but the process of multivariate data analysis is difficult. At present, the mainstream research direction of visual-tactile fusion is to realize the direct input of visual-tactile data and the output of results via end-to-end deep learning. However, some problems remain, such as the lack of a general research framework, confusion over methods and challenges related to unified evaluations.

VI. DISCUSSION AND FUTURE DIRECTIONS

The ultimate goal of researchers is to create machine vision and robots that have the same visual recognition and grasping

ability as human beings; this is an important step that must be achieved so that robots can be more widely applied—from industry to daily life. Although there has been great progress in object recognition, location, grasping speed and accuracy, there is still a vast gap that must be crossed, with human beings in the face of unstructured life scenes, which is an important reason why robots cannot be applied in daily life at present. Based on the development status of machine vision and the analogy analysis with human vision, the following thoughts are put forward regarding the future development of robot grasping.

1. Vision is still the mainstream technology. Due to the noncontact and high-efficiency characteristics of vision, it has great advantages. With the development of camera technology, the collection of environmental and object information will be more accurate and robust, which will greatly enhance the development of machine vision.

2. Tactile feedback will become an important part of robot grasping systems. Due to the inherent defects of vision, it is difficult to generate an appropriate grabbing strategy in complex environments according to the characteristics of objects collected by vision. Hence, the combination of vision and tactile feedback will be an important future development direction so that accurate recognition and positioning of the object and stable grasping can be achieved.

3. CNNs will still develop rapidly over a short period of time, but they may be replaced in the future. The CNN model evolves from a giant to a lightweight network step by step and achieves continuously higher accuracy in the process. However, it needs a massive amount of labeled data for training, which is time consuming. Real artificial intelligence (AI) needs the ability to complete few-shot learning.

4. Reinforcement learning and unsupervised learning will develop rapidly. Due to their low dependence on label data, the training process is relatively intelligent, which meets people's expectations of AI.

VII. CONCLUSION

Machine vision and robotics are two research directions that serve as inspiration for researchers all over the world. People hope to combine these two streams of research to create robots that have the same target recognition and grabbing ability as humans, which could lead to the partial realization of futuristic scenes in movies or science fiction. In this paper, the mainstream machine vision technology applied in robots is reviewed in detail, including traditional machine learning; CNNs, which have achieved good accomplishments in recent years; and reinforcement learning, unsupervised learning and self-supervised learning, which preclude labeled data limitations. In view of the limitations of vision, this paper also summarizes the development of tactile feedback in detail. This survey provides a detailed reference for the evaluation of current research on robot grasping based on machine vision and tactile feedback. Future research directions of machine vision and robot grasping are also considered.

REFERENCES

[1] L. Bozhkov and P. Georgieva, "Overview of deep learning architectures for EEG-based brain imaging," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*. Rio de Janeiro, Brazil: IEEE, Jul. 2018, pp. 1–7.

[2] X. Shen, H.-S. Kim, S. Komatsu, A. Markman, and B. Javidi, "Spatial-temporal human gesture recognition under degraded conditions using three-dimensional integral imaging: An overview," in *Proc. 17th Workshop Inf. Opt. (WIO)*. Québec, QC, Canada: IEEE, Jul. 2018, pp. 13938–13951.

[3] B. Gite, K. Nikhal, and F. Palnak, "Evaluating facial expressions in real time," in *Proc. Intell. Syst. Conf. (IntelliSys)*. London, U.K.: IEEE, Sep. 2017, pp. 849–855.

[4] P. Panchal, V. C. Raman, and S. Mantri, "Plant diseases detection and classification using machine learning models," in *Proc. 4th Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solution (CSITSS)*. Bengaluru, India: IEEE, Dec. 2019, pp. 1–6.

[5] M. Gao, J. Jiang, G. Zou, V. John, and Z. Liu, "RGB-D-Based object recognition using multimodal convolutional neural networks: A survey," *IEEE Access*, vol. 7, pp. 43110–43136, 2019.

[6] H. Wang, H. Du, Y. Zhao, and J. Yan, "A comprehensive overview of person re-identification approaches," *IEEE Access*, vol. 8, pp. 45556–45583, 2020.

[7] M. E. Celebi, N. Codella, and A. Halpern, "Dermoscopy image analysis: Overview and future directions," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 2, pp. 474–478, Mar. 2019.

[8] H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1153–1159, May 2016.

[9] D. Zhao, Y. Chen, and L. Lv, "Deep reinforcement learning with visual attention for vehicle classification," *IEEE Trans. Cognit. Develop. Syst.*, vol. 9, no. 4, pp. 356–367, Dec. 2017.

[10] W. Zhang, K. Song, X. Rong, and Y. Li, "Coarse-to-fine UAV target tracking with deep reinforcement learning," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 4, pp. 1522–1530, Oct. 2019.

[11] N. Hajj and M. Awad, "On biologically inspired stochastic reinforcement deep learning: A case study on visual surveillance," *IEEE Access*, vol. 7, pp. 108431–108437, 2019.

[12] H. Yuan, D. Li, and J. Wu, "Efficient learning of grasp selection for five-finger dexterous hand," in *Proc. IEEE 7th Annu. Int. Conf. CYBER Technol. Autom., Control, Intell. Syst. (CYBER)*. Honolulu, HI, USA: IEEE, Jul. 2017, pp. 1101–1106.

[13] J. Yang, S. Li, Z. Gao, Z. Wang, and W. Liu, "Real-time recognition method for 0.8 cm darning needles and KR22 bearings based on convolution neural networks and data increase," *Appl. Sci.*, vol. 8, no. 1857, pp. 1–18, 2018.

[14] J. Yang, S. Li, Z. Wang, and G. Yang, "Real-time tiny part defect detection system in manufacturing using deep learning," *IEEE Access*, vol. 7, pp. 89278–89291, 2019.

[15] A. Wang, M. Chu, M. Sha, and L. Liu, "A new process industry fault diagnosis algorithm based on ensemble improved binary-tree SVM," *Chin. J. Electron.*, vol. 24, no. 2, pp. 258–262, Apr. 2015.

[16] J. Li, N. Allinson, D. Tao, and X. Li, "Multitraining support vector machine for image retrieval," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3597–3601, Nov. 2006.

[17] E. Pasolli, F. Melgani, and Y. Bazi, "Support vector machine active learning through significance space construction," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 431–435, May 2011.

[18] D. Singh, D. Roy, and C. K. Mohan, "DiP-SVM : Distribution preserving kernel support vector machine for big data," *IEEE Trans. Big Data*, vol. 3, no. 1, pp. 79–90, Mar. 2017.

[19] J. Ruan, H. Jiang, X. Li, Y. Shi, F. T. S. Chan, and W. Rao, "A granular GA-SVM predictor for big data in agricultural cyber-physical systems," *IEEE Trans. Ind. Informat.*, vol. 15, no. 12, pp. 6510–6521, Dec. 2019.

[20] X. Hu, P. Niu, J. Wang, and X. Zhang, "A dynamic rectified linear activation units," *IEEE Access*, vol. 7, pp. 180409–180416, 2019.

[21] B. Zhang, M. Zhu, M. Yu, D. Pu, and G. Feng, "Extreme residual connected convolution-based collaborative filtering for document context-aware rating prediction," *IEEE Access*, vol. 8, pp. 53604–53613, 2020.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[23] L. Xiang, G. Zhao, Q. Li, W. Hao, and F. Li, "TUMK-ELM: A fast unsupervised heterogeneous data learning approach," *IEEE Access*, vol. 6, pp. 35305–35315, 2018.
[24] M. Usama, J. Qadir, A. Raza, H. Arif, K.-L.-A. Yau, Y. Elkhatib, A. Hussain, and A. Al-Fuqaha, "Unsupervised machine learning for networking: Techniques, applications and research challenges," *IEEE Access*, vol. 7, pp. 65579–65615, 2019.
[25] J.-Y. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 862–875, Apr. 2015.
[26] C. Liu, L. Song, J. Zhang, K. Chen, and J. Xu, "Self-supervised learning for specified latent representation," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 1, pp. 47–59, Jan. 2020.
[27] A. Zhao, J. Dong, and H. Zhou, "Self-supervised learning from multi-sensor data for sleep recognition," *IEEE Access*, vol. 8, pp. 93907–93921, 2020.
[28] W. Abdullah Al and I. D. Yun, "Partial policy-based reinforcement learning for anatomical landmark localization in 3D medical images," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 1245–1255, Apr. 2020.
[29] H. Liu, Y. Yu, F. Sun, and J. Gu, "Visual–tactile fusion for object recognition," *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 2, pp. 996–1008, Apr. 2017.
[30] X. Li, H. Liu, J. Zhou, and F. Sun, "Learning cross-modal visual-tactile representation using ensembled generative adversarial networks," *Cognit. Comput. Syst.*, vol. 1, no. 2, pp. 40–44, Jul. 2019.
[31] P. Falco, S. Lu, C. Natale, S. Pirozzi, and D. Lee, "A transfer learning approach to cross-modal object recognition: From visual observation to robotic haptic exploration," *IEEE Trans. Robot.*, vol. 35, no. 4, pp. 987–998, Aug. 2019.
[32] F. D. Ledezma and S. Haddadin, "FOP networks for learning humanoid body schema and dynamics," in *Proc. IEEE-RAS 18th Int. Conf. Humanoid Robots (Humanoids)*. Beijing, China: IEEE, Nov. 2018, pp. 1–9.
[33] M. C. Capolei, N. A. Andersen, H. H. Lund, E. Falotico, and S. Tolu, "A cerebellar internal models control architecture for online sensorimotor adaptation of a humanoid robot acting in a dynamic environment," *IEEE Robot. Autom. Lett.*, vol. 5, no. 1, pp. 80–87, Jan. 2020.
[34] F. Keyrouz, "A novel robotic sound localization and separation using non-causal filtering and Bayesian fusion," in *Proc. IEEE 26th Int. Workshop Mach. Learn. Signal Process. (MLSP)*. Vietri sul Mare, Italy: IEEE, Sep. 2016, pp. 1–6.
[35] E. Sauser and A. Billard, "Biologically inspired multimodal integration: Interferences in a human-robot interaction game," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* Beijing, China: IEEE, Oct. 2006, pp. 5619–5624.
[36] M. Toussaint and C. Goerick, "Probabilistic inference for structured planning in robotics," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* San Diego, CA, USA: IEEE, Oct. 2007, pp. 3068–3073.
[37] Q. Zhang, L. T. Yang, and Z. Chen, "Deep computation model for unsupervised feature learning on big data," *IEEE Trans. Services Comput.*, vol. 9, no. 1, pp. 161–171, Feb. 2016.
[38] W. Wang and M. Zhang, "Tensor deep learning model for heterogeneous data fusion in Internet of Things," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 4, no. 1, pp. 32–41, Feb. 2020.
[39] Y. Lei, F. Jia, J. Lin, S. Xing, and S. X. Ding, "An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data," *IEEE Trans. Ind. Electron.*, vol. 63, no. 5, pp. 3137–3147, May 2016.
[40] G. A. Susto, A. Schirru, S. Pampuri, and S. McLoone, "Supervised aggregative feature extraction for big data time series regression," *IEEE Trans. Ind. Informat.*, vol. 12, no. 3, pp. 1243–1252, Jun. 2016.
[41] N. Yu, Z. Li, and Z. Yu, "Survey on encoding schemes for genomic data representation and feature learning—From signal processing to machine learning," *Big Data Mining Anal.*, vol. 1, no. 3, pp. 191–210, 2018.
[42] F. Ye, Z. Zhang, K. Chakrabarty, and X. Gu, "Board-level functional fault diagnosis using multikernel support vector machines and incremental learning," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 33, no. 2, pp. 279–290, Feb. 2014.
[43] D. Elizondo, "The linear separability problem: Some testing methods," *IEEE Trans. Neural Netw.*, vol. 17, no. 2, pp. 330–344, Mar. 2006.
[44] A. J. Stimpson and M. L. Cummings, "Assessing intervention timing in computer-based education using machine learning algorithms," *IEEE Access*, vol. 2, pp. 78–87, 2014.
[45] M. Butcher and A. Karimi, "Linear parameter-varying iterative learning control with application to a linear motor system," *IEEE/ASME Trans. Mechatronics*, vol. 15, no. 3, pp. 412–420, Jun. 2010.
[46] J.-G. Hsieh, Y.-L. Lin, and J.-H. Jeng, "Preliminary study on Wilcoxon learning machines," *IEEE Trans. Neural Netw.*, vol. 19, no. 2, pp. 201–211, Feb. 2008.
[47] J. Song, F. Dong, J. Zhao, H. Wang, Z. He, and L. Wang, "An efficient multiobjective design optimization method for a PMSLM based on an extreme learning machine," *IEEE Trans. Ind. Electron.*, vol. 66, no. 2, pp. 1001–1011, Feb. 2019.
[48] N. D. Vanli, M. O. Sayin, I. Delibalta, and S. S. Kozat, "Sequential nonlinear learning for distributed multiagent systems via extreme learning machines," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 546–558, Mar. 2017.
[49] M. H. C. Law and A. K. Jain, "Incremental nonlinear dimensionality reduction by manifold learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 377–391, Mar. 2006.
[50] H. Liu, Z. Liu, S. Liu, Y. Liu, J. Bin, F. Shi, and H. Dong, "A nonlinear regression application via machine learning techniques for geomagnetic data reconstruction processing," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 128–140, Jan. 2019.
[51] G. Chen, J. Du, L. Sun, W. Zhang, K. Xu, X. Chen, G. T. Reed, and Z. He, "Nonlinear distortion mitigation by machine learning of SVM classification for PAM-4 and PAM-8 modulated optical interconnection," *J. Lightw. Technol.*, vol. 36, no. 3, pp. 650–657, Feb. 1, 2018.
[52] K. Gao, W. Guo, X. Yu, B. Liu, A. Yu, and X. Wei, "Deep induction network for small samples classification of hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3462–3477, 2020.
[53] D. Zhang, W. Ding, C. Liu, H. Wang, and B. Zhang, "Modulated autocorrelation convolution networks for automatic modulation classification based on small sample set," *IEEE Access*, vol. 8, pp. 27097–27105, 2020.
[54] Q. Zhou and X. He, "Broad learning model based on enhanced features learning," *IEEE Access*, vol. 7, pp. 42536–42550, 2019.
[55] J. Xu, Y. Y. Tang, B. Zou, Z. Xu, L. Li, Y. Lu, and B. Zhang, "The generalization ability of SVM classification based on Markov sampling," *IEEE Trans. Cybern.*, vol. 45, no. 6, pp. 1169–1179, Jun. 2015.
[56] C. Lu, A. Devos, J. A. K. Suykens, C. Arus, and S. Van Huffel, "Bagging linear sparse Bayesian learning models for variable selection in cancer diagnosis," *IEEE Trans. Inf. Technol. Biomed.*, vol. 11, no. 3, pp. 338–347, May 2007.
[57] A. Luo, F. An, X. Zhang, and H. J. Mattausch, "A hardware-efficient recognition accelerator using Haar-like feature and SVM classifier," *IEEE Access*, vol. 7, pp. 14472–14487, 2019.
[58] R. Trinchero, P. Manfredi, I. S. Stievano, and F. G. Canavero, "Machine learning for the performance assessment of high-speed links," *IEEE Trans. Electromagn. Compat.*, vol. 60, no. 6, pp. 1627–1634, Dec. 2018.
[59] A. J. Siddiqui, A. Mammeri, and A. Boukerche, "Real-time vehicle make and model recognition based on a bag of SURF features," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 11, pp. 3205–3219, Nov. 2016.
[60] M. C. Ergene and A. Durdu, "Robotic hand grasping of objects classified by using support vector machine and bag of visual words," in *Proc. Int. Artif. Intell. Data Process. Symp. (IDAP)*. Malatya, Turkey: IEEE, Sep. 2017, pp. 1–5.
[61] Y. Hu, Z. Li, G. Li, P. Yuan, C. Yang, and R. Song, "Development of sensory-motor fusion-based manipulation and grasping control for a robotic hand-eye system," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 47, no. 7, pp. 1169–1180, Jul. 2017.
[62] C. M. o. Valente, A. Schammass, A. F. R. Araujo, and G. A. P. Caurin, "Intelligent Grasping Using Neural Modules," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.* Tokyo, Japan: IEEE, Oct. 1999, pp. 780–785.
[63] M. Hannat, N. Zrira, Y. Raoui, and E. H. Bouyakhf, "A fast object recognition and categorization technique for robot grasping using the visual bag of words," in *Proc. 5th Int. Conf. Multimedia Comput. Syst. (ICMCS)*. Marrakech, Morocco: IEEE, Sep. 2016, pp. 173–178.
[64] K. Harada, T. Tsuji, K. Nagata, N. Yamanobe, H. Onda, T. Yoshimi, and Y. Kawai, "Object placement planner for robotic pick and place tasks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* Vilamoura, Portugal: IEEE, Oct. 2012, pp. 980–985.
[65] N. K. Verma, A. Mustafa, and A. Salour, "Stereo-vision based object grasping using robotic manipulator," in *Proc. 11th Int. Conf. Ind. Inf. Syst. (ICIIS)*. Roorkee, India: IEEE, Dec. 2016, pp. 95–100.

[66] J. Zhang and L. Shen, "Clustering and recognition for automated tracking and grasping of moving objects," in *Proc. IEEE Workshop Electron., Comput. Appl.* Ottawa, ON, Canada: IEEE, May 2014, pp. 222–229.

[67] R. Kouskouridas, A. Amanatiadis, and A. Gasteratos, "Guiding a robotic gripper by visual feedback for object manipulation tasks," in *Proc. IEEE Int. Conf. Mechatronics*. Istanbul, Turkey: IEEE, Apr. 2011, pp. 433–438.

[68] G. Wiesmann, S. Schraml, M. Litzenberger, A. N. Belbachir, M. Hofstatter, and C. Bartolozzi, "Event-driven embodied system for feature extraction and object recognition in robotic applications," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*. Providence, RI, USA: IEEE, Jun. 2012, pp. 76–82.

[69] O. Skotheim, M. Lind, P. Ystgaard, and S. A. Fjerdingen, "A flexible 3D object localization system for industrial part handling," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* Vilamoura, Portugal: IEEE, Oct. 2012, pp. 3326–3333.

[70] W. Budiharto, "Robust vision-based detection and grasping object for manipulator using SIFT keypoint detector," in *Proc. Int. Conf. Adv. Mech. Syst.* Kumamoto, Japan: IEEE, Aug. 2014, pp. 448–452.

[71] F. Wang, F. Sun, J. Zhang, B. Lin, and X. Li, "Unscented particle filter for online total image Jacobian matrix estimation in robot visual servoing," *IEEE Access*, vol. 7, pp. 92020–92029, 2019.

[72] Y. Bekiroglu, D. Song, L. Wang, and D. Kragic, "A probabilistic framework for task-oriented grasp stability assessment," in *Proc. IEEE Int. Conf. Robot. Autom.* Karlsruhe, Germany: IEEE, May 2013, pp. 3040–3047.

[73] H. O. Song, M. Fritz, D. Goehring, and T. Darrell, "Learning to detect visual grasp affordance," *IEEE Trans. Autom. Sci. Eng.*, vol. 13, no. 2, pp. 798–809, Apr. 2016.

[74] Z. Zhang, S. Mao, K. Chen, L. Xiao, B. Liao, C. Li, and P. Zhang, "CNN and PCA based visual system of a wheelchair manipulator robot for automatic drinking," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*. Kuala Lumpur, Malaysia: IEEE, Dec. 2018, pp. 1280–1286.

[75] E. Mattar, "PCA Learning for Non-brain Waves-Controlled Robotic Hand (Prosthesis): Grasp Stabilization and Control," in *Proc. UKSim-AMSS 16th Int. Conf. Comput. Modeling Simulation*. Cambridge, U.K.: IEEE, Mar. 2014, pp. 211–216.

[76] T. Ishii, R. Nakamura, H. Nakada, Y. Mochizuki, and H. Ishikawa, "Surface object recognition with CNN and SVM in Landsat 8 images," in *Proc. 14th IAPR Int. Conf. Mach. Vis. Appl. (MVA)*. Miraikan, Japan: IEEE Press, May 2015, pp. 341–344.

[77] Y. Shin and I. Balasingham, "Comparison of hand-craft feature based SVM and CNN based deep learning framework for automatic polyp classification," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*. Seogwipo, South Korea: IEEE, Jul. 2017, pp. 3277–3280.

[78] A. Wibisono, M. S. Saputri, P. Mursanto, J. Rachmad, Alberto, A. T. W. Yudasubrata, F. Rizki, and E. Anderson, "Deep learning and classic machine learning approach for automatic bone age assessment," in *Proc. 4th Asia–Pacific Conf. Intell. Robot Syst. (ACIRS)*. Nagoya, Japan: IEEE, Jul. 2019, pp. 235–240.

[79] P. Wang, L. Li, Y. Jin, and G. Wang, "Detection of unwanted traffic congestion based on existing surveillance system using in freeway via a CNN-architecture trafficnet," in *Proc. 13th IEEE Conf. Ind. Electron. Appl. (ICIEA)*. Wuhan, China: IEEE, May 2018, pp. 1134–1139.

[80] Y. Wang, C. Wang, L. Luo, and Z. Zhou, "Image classification based on transfer learning of convolutional neural network," in *Proc. Chin. Control Conf. (CCC)*. Guangzhou, China: IEEE, Jul. 2019, pp. 7506–7510.

[81] S. Sudha, K. B. Jayanthi, C. Rajasekaran, and T. Sunder, "Segmentation of RoI in medical images using CNN-a comparative study," in *Proc. TENCON-IEEE Region 10th Conf. (TENCON)*. Kochi, India: IEEE, Oct. 2019, pp. 767–771.

[82] B. Jiang, J. He, S. Yang, H. Fu, T. Li, H. Song, and D. He, "Fusion of machine vision technology and AlexNet-CNNs deep learning network for the detection of postharvest apple pesticide residues," *Artif. Intell. Agricult.*, vol. 1, pp. 1–8, Mar. 2019.

[83] A. Ibrahim, A. Dalbah, A. Abualsaud, U. Tariq, and A. El-Hag, "Application of machine learning to evaluate insulator surface erosion," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 2, pp. 314–316, Feb. 2020.

[84] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[85] M. Zhou, Z. Pan, Y. Liu, Q. Zhang, Y. Cai, and H. Pan, "Leak detection and location based on ISLMD and CNN in a pipeline," *IEEE Access*, vol. 7, pp. 30457–30464, 2019.

[86] L. Xu, L. Wang, Y. Zhang, and S. Cheng, "Visual tracking based on siamese network of fused score map," *IEEE Access*, vol. 7, pp. 151389–151398, 2019.

[87] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Boston, MA, USA: IEEE, Jun. 2015, pp. 1–9.

[88] Q. Gao, J. Liu, Z. Ju, and X. Zhang, "Dual-hand detection for human–robot interaction by a parallel network based on hand detection and body pose estimation," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9663–9672, Dec. 2019.

[89] S. Yang, G. Lin, Q. Jiang, and W. Lin, "A dilated inception network for visual saliency prediction," *IEEE Trans. Multimedia*, vol. 22, no. 8, pp. 2163–2176, Aug. 2020.

[90] X. Jin, L. Wu, X. Li, X. Zhang, J. Chi, S. Peng, S. Ge, G. Zhao, and S. Li, "ILGNet: Inception modules with connected local and global features for efficient image aesthetic quality classification using domain adaptation," *IET Comput. Vis.*, vol. 13, no. 2, pp. 206–212, Mar. 2019.

[91] W. Dongyu, H. Fuwen, T. Mikolajczyk, and H. Yunhua, "Object detection for soft robotic manipulation based on RGB-D sensors," in *Proc. WRC Symp. Adv. Robot. Autom. (WRC SARA)*. Beijing, China: IEEE, Aug. 2018, pp. 52–58.

[92] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015, pp. 1–14.

[93] W. Guan, T. Wang, J. Qi, L. Zhang, and H. Lu, "Edge-aware convolution neural network based salient object detection," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 114–118, Jan. 2019.

[94] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[95] Z. Zhao, T. Cai, F. Chang, and X. Cheng, "Real-time surgical instrument detection in robot-assisted surgery using a convolutional neural network cascade," *Healthcare Technol. Lett.*, vol. 6, no. 6, pp. 275–279, Dec. 2019.

[96] L. Liu, X. Tang, J. Xie, X. Gao, W. Zhao, F. Mo, and G. Zhang, "Deep-learning and depth-map based approach for detection and 3D localization of small traffic signs," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2096–2111, 2020.

[97] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 1–12.

[98] Y. Liao, P. Xiong, W. Min, W. Min, and J. Lu, "Dynamic sign language recognition based on video sequence with BLSTM-3D residual networks," *IEEE Access*, vol. 7, pp. 38044–38054, 2019.

[99] X. Ou, P. Yan, Y. Zhang, B. Tu, G. Zhang, J. Wu, and W. Li, "Moving object detection method via ResNet-18 with Encoder–Decoder structure in complex scenes," *IEEE Access*, vol. 7, pp. 108152–108160, 2019.

[100] W. Liu, "SSD: Single Shot MultiBox Detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, 2016, pp. 1–17.

[101] X. Li, C. Liu, S. Dai, H. Lian, and G. Ding, "Scale specified single shot multibox detector," *IET Comput. Vis.*, vol. 14, no. 2, pp. 59–64, Mar. 2020.

[102] L. Chen, Z. Zhang, and L. Peng, "Fast single shot multibox detector and its application on vehicle counting system," *IET Intell. Transp. Syst.*, vol. 12, no. 10, pp. 1406–1413, Dec. 2018.

[103] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 779–788.

[104] Y. Yu, K. Zhang, H. Liu, L. Yang, and D. Zhang, "Real-time visual localization of the picking points for a ridge-planting strawberry harvesting robot," *IEEE Access*, vol. 8, pp. 116556–116568, 2020.

[105] L. Yang, M. Li, X. Song, Z. Xiong, C. Hou, and B. Qu, "Vehicle speed measurement based on binocular stereovision system," *IEEE Access*, vol. 7, pp. 106628–106641, 2019.

[106] T. Kitayama, H. Lu, Y. Li, and H. Kim "Detection of grasping position from video images based on SSD," in *Proc. 18th Int. Conf. Control, Autom. Syst. (ICCAS)*. Daegwallyeong, South Korea, Oct. 2018, pp. 1472–1475.

[107] Y. Chao, X. Chen, and N. Xiao, "Deep learning-based grasp-detection method for a five-fingered industrial robot hand," *IET Comput. Vis.*, vol. 13, no. 1, pp. 61–70, Feb. 2019.

[108] G. Wu, W. Chen, H. Cheng, W. Zuo, D. Zhang, and J. You, "Multi-object grasping detection with hierarchical feature fusion," *IEEE Access*, vol. 7, pp. 43884–43894, 2019.
[109] K. Choi, J. K. Suhr, and H. G. Jung, "Map-matching-based cascade landmark detection and vehicle localization," *IEEE Access*, vol. 7, no. 1, pp. 127874–127894, 2019.
[110] Y. Xu, L. Wang, A. Yang, and L. Chen, "GraspCNN: Real-time grasp detection using a new oriented diameter circle representation," *IEEE Access*, vol. 7, pp. 159322–159331, 2019.
[111] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.
[112] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," *Int. J. Comput. Vis.*, vol. 128, no. 3, pp. 642–656, Mar. 2020.
[113] H. Cheng and M. Q.-H. Meng, "A grasp pose detection scheme with an end-to-end CNN regression approach," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*. Kuala Lumpur, Malaysia: IEEE: Malaysia, Dec. 2018, pp. 544–549.
[114] F. H. Zunjani, S. Sen, H. Shekhar, A. Powale, D. Godnaik, and G. C. Nandi, "Intent-based object grasping by a robot using deep learning," in *Proc. IEEE 8th Int. Advance Comput. Conf. (IACC)*. Greater Noida, India: IEEE, Dec. 2018, pp. 246–251.
[115] E. Corona, G. Alenya, A. Gabas, and C. Torras, "Active garment recognition and target grasping point detection using deep learning," *Pattern Recognit.*, vol. 74, pp. 629–641, Feb. 2018.
[116] A. Gaona and H.-I. Lin, "Robotic grasping estimation by evolutionary deep networks," in *Proc. Int. Autom. Control Conf. (CACS)*. Taoyuan, Taiwan: IEEE, Nov. 2018, pp. 1–7.
[117] K. Yamazaki, "Selection of grasp points of cloth product on a table based on shape classification feature," in *Proc. IEEE Int. Conf. Inf. Autom. (ICIA)*. Macau, China: IEEE, Jul. 2017, pp. 136–141.
[118] L. Haochen, Z. Bin, S. Xiaoyong, and Z. Yongting, "CNN-based model for pose detection of industrial PCB," in *Proc. 10th Int. Conf. Intell. Comput. Technol. Autom. (ICICTA)*. Changsha, China: IEEE, Oct. 2017, pp. 390–393.
[119] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767
[120] L. Chen, P. Huang, and Z. Meng, "Convolutional multi-grasp detection using grasp path for RGBD images," *Robot. Auto. Syst.*, vol. 113, pp. 94–103, Mar. 2019.
[121] R. Roy, A. Kumar, M. Mahadevappa, and C. S. Kumar, "Deep learning based object shape identification from EOG controlled vision system," in *Proc. IEEE Sensors*. New Delhi, India: IEEE, Oct. 2018, pp. 1–4.
[122] X. Yan, J. Hsu, M. Khansari, Y. Bai, A. Pathak, A. Gupta, J. Davidson, and H. Lee, "Learning 6-DOF grasping interaction via deep geometry-aware 3D representations," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*. Brisbane, QLD, Australia: IEEE, May 2018, pp. 3766–3773.
[123] V. Satish, J. Mahler, and K. Goldberg, "On-policy dataset synthesis for learning robot grasping policies using fully convolutional deep networks," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1357–1364, Apr. 2019.
[124] H. Liang, X. Ma, S. Li, M. Gorner, S. Tang, B. Fang, F. Sun, and J. Zhang, "PointNetGPD: Detecting grasp configurations from point sets," in *Proc. Int. Conf. Robot. Autom. (ICRA)*. Montreal, QC, Canada: IEEE, May 2019, pp. 3629–3635.
[125] X. Sun, T. Nozaki, T. Murakami, and K. Ohnishi, "Grasping point estimation based on stored motion and depth data in motion reproduction system," in *Proc. IEEE Int. Conf. Mechatronics (ICM)*. Ilmenau, Germany: IEEE, Mar. 2019, pp. 471–476.
[126] Z. Deng, X. Zheng, L. Zhang, and J. Zhanga, "A learning framework for semantic reach-to-grasp tasks integrating machine learning and optimization," *Robot. Auton. Syst.*, vol. 108, pp. 140–152, Oct. 2018.
[127] I. González-Díaz, J. Benois-Pineau, J.-P. Domenger, D. Cattaert, and A. de Rugy, "Perceptually-guided deep neural networks for ego-action prediction: Object grasping," *Pattern Recognit.*, vol. 88, pp. 223–235, Apr. 2019.
[128] M. Farag, A. N. A. Ghafar, and M. H. Alsibai, "Real-time robotic grasping and localization using deep learning-based object detection technique," in *Proc. IEEE Int. Conf. Autom. Control Intell. Syst. (I2CACIS)*. Selangor, Malaysia: IEEE, Jun. 2019, pp. 139–144.
[129] R. Detry, C. H. Ek, M. Madry, J. Piater, and D. Kragic, "Generalizing grasps across partly similar objects," in *Proc. IEEE Int. Conf. Robot. Autom.* Saint Paul, MN, USA: IEEE, May 2012, pp. 3791–3797.
[130] T. Nakamura, T. Nagai, and N. Iwahashi, "Multimodal categorization by hierarchical Dirichlet process," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* San Francisco, CA, USA: IEEE, Sep. 2011, pp. 1520–1525.
[131] T. Nakamura, T. Nagai, and N. Iwahashi, "Multimodal object categorization by a robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* San Diego, CA, USA: IEEE, Oct. 2007, pp. 2415–2420.
[132] T. Nagai and N. Iwahashi, "Object categorization using multimodal information," in *Proc. TENCON-IEEE Region 10th Conf.* Hong Kong: IEEE, 2006, pp. 1–4.
[133] V.-T. Nguyen, C. Lin, C.-H.-G. Li, S.-M. Guo, and J.-J.-J. Lien, "Visual-guided robot arm using self-supervised deep convolutional neural networks," in *Proc. IEEE 15th Int. Conf. Autom. Sci. Eng. (CASE)*. Vancouver, BC, Canada: IEEE, Aug. 2019, pp. 1415–1420.
[134] A. Murali, L. Pinto, D. Gandhi, and A. Gupta, "CASSL: Curriculum accelerated self-supervised learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*. Brisbane, QLD, Australia: IEEE, May 2018, pp. 6453–6460.
[135] P. Florence, L. Manuelli, and R. Tedrake, "Self-supervised correspondence in visuomotor policy learning," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 492–499, Apr. 2020.
[136] M. Yan, Y. Zhu, N. Jin, and J. Bohg, "Self-supervised learning of state estimation for manipulating deformable linear objects," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 2372–2379, Apr. 2020.
[137] Y. Yang, H. Liang, and C. Choi, "A deep learning approach to grasping the invisible," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 2232–2239, Apr. 2020.
[138] G. Zhang, H. Li, and Odbal, "Research on fuzzy enhanced learning model of multienhanced signal learning automata," *IEEE Trans. Ind. Informat.*, vol. 15, no. 11, pp. 5980–5987, Nov. 2019.
[139] S. Jeong, M. Lee, H. Arie, and J. Tani, "Developmental learning of integrating visual attention shifts and bimanual object grasping and manipulation tasks," in *Proc. IEEE 9th Int. Conf. Develop. Learn.* Ann Arbor, MI, USA: IEEE, Aug. 2010, pp. 165–170.
[140] W. Yuan, K. Hang, D. Kragic, M. Y. Wang, and J. A. Stork, "End-to-end nonprehensile rearrangement with deep reinforcement learning and simulation-to-reality transfer," *Robot. Auto. Syst.*, vol. 119, pp. 119–134, Sep. 2019.
[141] K. Terada, H. Takeda, and T. Nishida, "An acquisition of the relation between vision and action using self-organizing map and reinforcement learning," in *Proc. 2nd Int. Conf. Knowl.-Based Intell. Electron. Syst.* Adelaide, SA, Australia: IEEE, Apr. 1998, pp. 429–434.
[142] T. Lampe and M. Riedmiller, "Acquiring visual servoing reaching and grasping skills using neural reinforcement learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*. Dallas, TX, USA: IEEE, Aug. 2013.
[143] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Robot. Res.*, vols. 4–5, no. 34, pp. 705–724, 2015.
[144] P. Ardon, E. Pairet, R. P. A. Petrick, S. Ramamoorthy, and K. S. Lohan, "Learning grasp affordance reasoning through semantic relations," *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 4571–4578, Oct. 2019.
[145] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014. vol. 3, no. 6, pp. 1–9.
[146] K. Yu, C. Dong, L. Lin, and C. C. Loy, "Crafting a toolchain for image restoration by deep reinforcement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 2443–2452.
[147] S. James, Z. Ma, D. Rovick Arrojo, and A. J. Davison, "RLBench: The robot learning benchmark & learning environment," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 3019–3026, Apr. 2020.
[148] A. Zeng, S. Song, S. Welker, J. Lee, A. Rodriguez, and T. Funkhouser, "Learning synergies between pushing and grasping with self-supervised deep reinforcement learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*. Madrid, Spain: IEEE, Oct. 2018, pp. 4238–4245.
[149] Y. Wang, H. Lang, and C. W. de Silva, "A hybrid visual servo controller for robust grasping by wheeled mobile robots," *IEEE/ASME Trans. Mechatronics*, vol. 15, no. 5, pp. 757–769, Oct. 2010.
[150] S. Gu, E. Holly, T. Lillicrap, and S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*. Singapore: IEEE, May 2017, pp. 3389–3396.
[151] M. Breyer, F. Furrer, T. Novkovic, R. Siegwart, and J. Nieto, "Comparing task simplifications to learn closed-loop object picking using deep reinforcement learning," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1549–1556, Apr. 2019.

[152] K. Katyal, I.-J. Wang, and P. Burlina, "Leveraging deep reinforcement learning for reaching robotic tasks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*. Honolulu, HI, USA: IEEE, Jul. 2017, pp. 490–491.

[153] A. Ghadirzadeh, A. Maki, D. Kragic, and M. Bjorkman, "Deep predictive policy training using reinforcement learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*. Vancouver, BC, Canada: IEEE, Sep. 2017, pp. 2351–2358.

[154] K. N. Nguyen, J. Yoo, and Y. Choe, "Speeding up affordance learning for tool use, using proprioceptive and kinesthetic inputs," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*. Budapest, Hungary: IEEE, Jul. 2019, pp. 1–8.

[155] C. C. Beltran-Hernandez, D. Petit, I. G. Ramirez-Alpizar, and K. Harada, "Learning to grasp with primitive shaped object policies," in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*. Paris, France: IEEE, Jan. 2019, pp. 468–473.

[156] Z. Li, T. Zhao, F. Chen, Y. Hu, C.-Y. Su, and T. Fukuda, "Reinforcement learning of manipulation and grasping using dynamical movement primitives for a humanoidlike mobile manipulator," *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 1, pp. 121–131, Feb. 2018.

[157] Z. Miljković, M. Mitić, M. Lazarević, and B. Babić, "Neural network reinforcement learning for visual control of robot manipulators," *Expert Syst. Appl.*, vol. 40, no. 5, pp. 1721–1736, Apr. 2013.

[158] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige, S. Levine, and V. Vanhoucke, "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*. Brisbane, QLD, Australia: IEEE, May 2018, pp. 4243–4250.

[159] S. James, P. Wohlhart, M. Kalakrishnan, D. Kalashnikov, A. Irpan, J. Ibarz, S. Levine, R. Hadsell, and K. Bousmalis, "Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 12619–12629.

[160] R. B. Hellman, C. Tekin, M. van der Schaar, and V. J. Santos, "Functional contour-following via haptic perception and reinforcement learning," *IEEE Trans. Haptics*, vol. 11, no. 1, pp. 61–72, Jan. 2018.

[161] R. Platt, "Learning grasp strategies composed of contact relative motions," in *Proc. 7th IEEE-RAS Int. Conf. Humanoid Robots*. Pittsburgh, PA, USA: IEEE, Nov. 2007, pp. 49–56.

[162] H. Merzic, M. Bogdanovic, D. Kappler, L. Righetti, and J. Bohg, "Leveraging contact forces for learning to grasp," in *Proc. Int. Conf. Robot. Autom. (ICRA)*. Montreal, QC, Canada: IEEE, May 2019, pp. 3615–3621.

[163] Y. Xing, F. Shen, and J. Zhao, "Perception evolution network based on cognition deepening model—Adapting to the emergence of new sensory receptor," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 3, pp. 607–620, Mar. 2016.

[164] S. Lowry and M. J. Milford, "Supervised and unsupervised linear learning techniques for visual place recognition in changing environments," *IEEE Trans. Robot.*, vol. 32, no. 3, pp. 600–613, Jun. 2016.

[165] T. Nguyen, S. W. Chen, S. S. Shivakumar, C. J. Taylor, and V. Kumar, "Unsupervised deep homography: A fast and robust homography estimation model," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2346–2353, Jul. 2018.

[166] F. Despinoy, D. Bouget, G. Forestier, C. Penet, N. Zemiti, P. Poignet, and P. Jannin, "Unsupervised trajectory segmentation for surgical gesture recognition in robotic training," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 6, pp. 1280–1291, Jun. 2016.

[167] T. Feng and D. Gu, "SGANVO: Unsupervised deep visual odometry and depth estimation with stacked generative adversarial networks," *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 4431–4437, Oct. 2019.

[168] Y. Su, W. Li, W. Nie, D. Song, and A.-A. Liu, "Unsupervised feature learning with graph embedding for view-based 3D model retrieval," *IEEE Access*, vol. 7, pp. 95285–95296, 2019.

[169] Y. Li, C. Tao, Y. Tan, K. Shang, and J. Tian "Unsupervised multilayer feature learning for satellite image scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, pp. 157–161, Feb. 2016.

[170] H. Qiao, Y. Li, F. Li, X. Xi, and W. Wu, "Biologically inspired model for visual cognition achieving unsupervised episodic and semantic feature learning," *IEEE Trans. Cybern.*, vol. 46, no. 10, pp. 2335–2347, Oct. 2016.

[171] S. Duffner and C. Garcia, "Visual focus of attention estimation with unsupervised incremental learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 12, pp. 2264–2272, Dec. 2016.

[172] B. C. Kwon, B. Eysenbach, J. Verma, K. Ng, C. De Filippi, W. F. Stewart, and A. Perer, "Clustervision: Visual supervision of unsupervised clustering," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 1, pp. 142–151, Jan. 2018.

[173] X. Li, H. Zhang, R. Zhang, and F. Nie, "Discriminative and uncorrelated feature selection with constrained spectral analysis in unsupervised learning," *IEEE Trans. Image Process.*, vol. 29, pp. 2139–2149, 2020.

[174] M. A. Lee, Y. Zhu, P. Zachares, M. Tan, K. Srinivasan, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Learning multimodal representations for contact-rich tasks," *IEEE Trans. Robot.*, vol. 36, no. 3, pp. 582–596, Jun. 2020.

[175] L. Wellhausen, A. Dosovitskiy, R. Ranftl, K. Walas, C. Cadena, and M. Hutter, "Where should i walk? Predicting terrain properties from images via self-supervised learning," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1509–1516, Apr. 2019.

[176] X. Shu, C. Liu, T. Li, C. Wang, and C. Chi, "A self-supervised learning manipulator grasping approach based on instance segmentation," *IEEE Access*, vol. 6, pp. 65055–65064, 2018.

[177] T. Mar, V. Tikhanoff and L. Natale, "What can i do with this tool self-supervised learning of tool affordances from their 3-D geometry," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 10, no. 3, pp. 595–610, Sep. 2018.

[178] T. Schmidt, R. Newcombe, and D. Fox, "Self-supervised visual descriptor learning for dense correspondence," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 420–427, Apr. 2017.

[179] J. Yuan and Y. Wu, "Mining visual collocation patterns via self-supervised subspace learning," *IEEE Trans. Syst., Man, Cybern. B. Cybern.*, vol. 42, no. 2, pp. 334–346, Apr. 2012.

[180] K. Yun, J. Park, and J. Cho, "Robust human pose estimation for rotation via self-supervised learning," *IEEE Access*, vol. 8, pp. 32502–32517, 2020.

[181] Y. Cong, J. Liu, J. Yuan, and J. Luo, "Self-supervised online metric learning with low rank constraint for scene categorization," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3179–3191, Aug. 2013.

[182] K. Stefanov, J. Beskow, and G. Salvi, "Self-supervised vision-based detection of the active speaker as support for socially aware language acquisition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 12, no. 2, pp. 250–259, Jun. 2020.

[183] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, and M. Velimirovic, "SPICE: Self-supervised pitch estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1118–1128, 2020.

[184] M. A. Moussa, "Combining expert neural networks using reinforcement feedback for learning primitive grasping behavior," *IEEE Trans. Neural Netw.*, vol. 15, no. 3, pp. 629–638, May 2004.

[185] H. Shi, G. Sun, Y. Wang, and K.-S. Hwang, "Adaptive image-based visual servoing with temporary loss of the visual signal," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 1956–1965, Apr. 2019.

[186] R. S. Sharma, R. R. Nair, P. Agrawal, L. Behera, and V. K. Subramanian, "Robust hybrid visual servoing using reinforcement learning and finite-time adaptive FOSMC," *IEEE Syst. J.*, vol. 13, no. 3, pp. 3467–3478, Sep. 2019.

[187] S. Yun, J. Choi, Y. Yoo, K. Yun, and J. Y. Choi, "Action-driven visual object tracking with deep reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2239–2252, Jun. 2018.

[188] Y. Xie, J. Xiao, K. Huang, J. Thiyagalingam, and Y. Zhao, "Correlation filter selection for visual tracking using reinforcement learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 192–204, Jan. 2020.

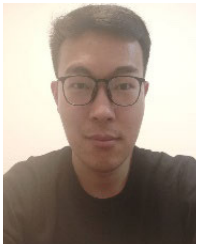
[189] Y. Wang, L. Zhang, L. Wang, and Z. Wang, "Multitask learning for object localization with deep reinforcement learning," *IEEE Trans. Cognit. Develop. Syst.*, vol. 11, no. 4, pp. 573–580, Dec. 2019.

[190] Z. Ni and S. Paul, "A multistage game in smart grid security: A reinforcement learning solution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2684–2695, Sep. 2019.

[191] Y. Zeng, K. Xu, L. Qin, and Q. Yin, "A semi-Markov decision model with inverse reinforcement learning for recognizing the destination of a maneuvering agent in real time strategy games," *IEEE Access*, vol. 8, pp. 15392–15409, 2020.

[192] M. S. Emigh, E. G. Kriminger, A. J. Brockmeier, J. C. Principe, and P. M. Pardalos, "Reinforcement learning in video games using nearest neighbor interpolation and metric learning," *IEEE Trans. Comput. Intell. AI Games*, vol. 8, no. 1, pp. 56–66, Mar. 2016.

[193] R. Li, R. Platt, W. Yuan, A. T. Pas, N. Roscup, M. A. Srinivasan, and E. Adelson, "Localization and manipulation of small parts using gelsight tactile sensing," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* Chicago, IL, USA: IEEE, Sep. 2014, pp. 3988–3993.
[194] K.-T. Yu and A. Rodriguez, "Realtime state estimation with tactile and visual Sensing. Application to planar manipulation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*. Brisbane, QLD, Australia: IEEE, May 2018, pp. 7778–7783.
[195] J. Bimbo, S. Rodriguez-Jimenez, H. Liu, X. Song, N. Burrus, L. D. Senerivatne, M. Abderrahim, and K. Althoefer, "Object pose estimation and tracking by fusing visual and tactile information," in *Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst. (MFI)*. Hamburg, Germany: IEEE, Sep. 2012, pp. 65–70.
[196] C. Schuetz, J. Pfaff, F. Sygulla, D. Rixen, and H. Ulbrich, "Motion planning for redundant manipulators in uncertain environments based on tactile feedback," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*. Hamburg, Germany: IEEE, Sep. 2015, pp. 6387–6394.
[197] J. Zhang, C. Song, Y. Hu, and B. Yu, "Improving robustness of robotic grasping by fusing multi-sensor," in *Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst. (MFI)*. Hamburg, Germany: IEEE, Sep. 2012, pp. 126–131.
[198] J. Sanchez, C. M. Mateo, J. A. Corrales, B.-C. Bouzgarrou, and Y. Mezouar, "Online shape estimation based on tactile sensing and deformation modeling for robot manipulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*. Madrid, Spain: IEEE, Oct. 2018, pp. 504–511.
[199] M. Rasouli, Y. Chen, A. Basu, S. L. Kukreja, and N. V. Thakor, "An extreme learning machine-based neuromorphic tactile sensing system for texture recognition," *IEEE Trans. Biomed. Circuits Syst.*, vol. 12, no. 2, pp. 313–325, Apr. 2018.
[200] B. Ward-Cherrier, N. Rojas, and N. F. Lepora, "Model-free precise in-hand manipulation with a 3D-printed tactile gripper," *IEEE Robot. Autom. Lett.*, vol. 2, no. 4, pp. 2056–2063, Oct. 2017.
[201] J. Bimbo, S. Luo, K. Althoefer, and H. Liu, "In-hand object pose estimation using covariance-based tactile to geometry matching," *IEEE Robot. Autom. Lett.*, vol. 1, no. 1, pp. 570–577, Jan. 2016.
[202] H. Liu, D. Guo, and F. Sun, "Object recognition using tactile measurements: Kernel sparse coding methods," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 3, pp. 656–665, Mar. 2016.
[203] T. Bhattacharjee, H. M. Clever, J. Wade, and C. C. Kemp, "Multimodal tactile perception of objects in a real home," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2523–2530, Jul. 2018.
[204] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, "More than a feeling: Learning to grasp and regrasp using vision and touch," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3300–3307, Oct. 2018.
[205] D. Guo, F. Sun, H. Liu, T. Kong, B. Fang, and N. Xi, "A hybrid deep architecture for robotic grasp detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*. Singapore: IEEE, May 2017, pp. 1609–1614.
[206] J. Li, S. Dong, and E. Adelson, "Slip detection with combined tactile and visual information," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*. Brisbane, QLD, Australia: IEEE, May 2018, pp. 7772–7777.
[207] N. P. Garg, D. Hsu, and W. S. Lee, "Learning to grasp under uncertainty using POMDPs," in *Proc. Int. Conf. Robot. Autom. (ICRA)*. Montreal, QC, Canada: IEEE, May 2019, pp. 2751–2757.
[208] B. Ward-Cherrier, L. Cramphorn, and N. F. Lepora, "Tactile manipulation with a TacThumb integrated on the open-hand m2 gripper," *IEEE Robot. Autom. Lett.*, vol. 1, no. 1, pp. 169–175, Jan. 2016.
[209] C. Li, D. Yan, and J. Shen, "A convex tactile sensor for isotropic tissue elastic modulus estimation based on the plane contact model," *IEEE Sensors J.*, vol. 19, no. 15, pp. 6251–6259, Aug. 2019.
[210] N. Pestell, J. Lloyd, J. Rossiter, and N. F. Lepora, "Dual-modal tactile perception and exploration," *IEEE Robot. Autom. Lett.*, vol. 3, no. 2, pp. 1033–1040, Apr. 2018.
[211] W. Zheng, B. Wang, H. Liu, X. Wang, Y. Li, and C. Zhang, "Bio-inspired magnetostrictive tactile sensor for surface material recognition," *IEEE Trans. Magn.*, vol. 55, no. 7, pp. 1–7, Jul. 2019.
[212] M. N. Saadatzi, J. R. Baptist, Z. Yang, and D. O. Popa, "Modeling and fabrication of scalable tactile sensor arrays for flexible robot skins," *IEEE Sensors J.*, vol. 19, no. 17, pp. 7632–7643, Sep. 2019.
[213] S. Sundaram, P. Kellnhofer, Y. Li, J.-Y. Zhu, A. Torralba, and W. Matusik, "Learning the signatures of the human grasp using a scalable tactile glove," *Nature*, vol. 569, no. 7758, pp. 698–702, May 2019.
[214] S. Wang, J. Wu, X. Sun, W. Yuan, W. T. Freeman, J. B. Tenenbaum, and E. H. Adelson, "3D shape perception from monocular vision, touch, and shape priors," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*. Madrid, Spain: IEEE, Oct. 2018, pp. 1606–1613.
[215] F. R. Hogan, M. Bauza, O. Canal, E. Donlon, and A. Rodriguez, "Tactile regrasp: Grasp adjustments via simulated tactile transformations," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*. Madrid, Spain: IEEE, Oct. 2018, pp. 2963–2970.
[216] F. Sun, C. Liu, W. Huang, and J. Zhang, "Object classification and grasp planning using visual and tactile sensing," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 46, no. 7, pp. 969–979, Jul. 2016.
[217] D. Jain, A. Li, S. Singhal, A. Rajeswaran, V. Kumar, and E. Todorov, "Learning deep visuomotor policies for dexterous hand manipulation," in *Proc. Int. Conf. Robot. Autom. (ICRA)*. Montreal, QC, Canada: IEEE, May 2019, pp. 3636–3643.
[218] C. D. Santina, V. Arapi, G. Averta, F. Damiani, G. Fiore, A. Settimi, M. G. Catalano, D. Bacciu, A. Bicchi, and M. Bianchi, "Learning from humans how to grasp: A data-driven architecture for autonomous grasping with anthropomorphic soft hands," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1533–1540, Apr. 2019.
[219] K. Hang, M. Li, J. A. Stork, Y. Bekiroglu, F. T. Pokorny, A. Billard, and D. Kragic, "Hierarchical fingertip space: A unified framework for grasp planning and in-hand grasp adaptation," *IEEE Trans. Robot.*, vol. 32, no. 4, pp. 960–972, Aug. 2016.



QIANG BAI received the B.Sc. degree from Zaozhuang University, in 2015, the double master's degree from Yuan Ze University, and the M.Sc. degree from Guizhou University and Yuan Ze University, in 2018. He is currently pursuing the Ph.D. degree with the School of Mechanical Engineering, Guizhou University, Guiyang, China. From September 2016 to August 2017, he was Joint Educated with Yuan Ze University. His research interests include machine learning, robot, grasp, vision, and location.

SHAOBO LI was a Professor with the School of Mechanical Engineering, Guizhou University (GZU), China. From 2007 to 2015, he was the Vice Director of the Key Laboratory of Advanced Manufacturing Technology, Ministry of Education, GZU. Since 2015, he has been the Dean of the School of Mechanical Engineering, GZU. His research has been supported by the National Science Foundation of China (NSFC) and the National High-Tech Research and Development Program (863 Program). His main research interests include intelligence manufacturing and big data.

JING YANG (Member, IEEE) received the B.Sc. degree from Anyang Normal University, in 2015, and the Ph.D. degree from the School of Mechanical Engineering. He is currently a Lecturer with Guizhou University. From August 2018 to September 2019, he was awarded a scholarship by the China Scholarship Council (CSC) under the State Scholarship Fund to pursue his study with Oklahoma State University, as a Joint Ph.D. Student with the Institute for Mechatronic Engineering, where he joined the Guoliang Fan's Group, as a Professor. He has published over ten papers in reputed journals/conferences. His main research interests include machine vision, deep learning, and smart manufacturing applications. He has also served as a Reviewer for several journals, such as IEEE Access and the IEEE Transactions on Semiconductor Manufacturing.

QISONG SONG received the B.S. degree in mechanical engineering from the Harbin University of Science and Technology, Harbin, China, in 2018. He is currently pursuing the M.S. degree in mechanical engineering with Guizhou University, Guiyang, China. His research interest includes mobile robot path planning.

ZHIANG LI received the B.S. degree in mechanical engineering from Guizhou University, in 2018, where he is currently pursuing the master's degree. His research interests include trajectory planning and machine vision of manipulator.

XINGXING ZHANG received the B.S. degree in mechanical engineering from Nanjing Normal University, in 2018. She is currently pursuing the master's degree with the School of Mechanical Engineering, Guizhou University. Her research interests include robots, tactile sensing, and so on.