machines

MDPI

Review

Object Detection and Distance Measurement in Teleoperation

Ailing Zhang [1], Meng Chu [2], Zixin Chen [3], Fuqiang Zhou [2] and Shuo Gao [2,*]

[1] School of Computer Science and Engineering, Beihang University, Beijing 100191, China; 19375203@buaa.edu.cn
[2] School of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing 100191, China; trueman@buaa.edu.cn (M.C.); zfq@buaa.edu.cn (F.Z.)
[3] School of Mechanical Engineering and Automation, Beihang University, Beijing 100191, China; 19375202@buaa.edu.cn
* Correspondence: shuo_gao@buaa.edu.cn

Abstract: In recent years, teleoperation has experienced rapid development. Numerous teleoperation applications in diverse areas have been reported. Among all teleoperation-related components, computer vision (CV) is treated as one of the must-have technologies, because it allows users to observe remote scenarios. In addition, CV can further help the user to identify and track the desired targets from complex scenes. It has been proven that efficient CV methods can significantly improve the operation accuracy and relieve user's physical and mental fatigue. Therefore, furthering understanding about CV techniques and reviewing the latest research outcomes is necessary for teleoperation designers. In this context, this review article was composed.

Keywords: teleoperation; computer vision; object detection; distance measurement

1. Introduction

Teleoperation, which allows users to manipulate remote objects through sensing, actuating and communication technologies, has been applied in diverse application scenarios such as space exploration and remote surgery. A typical teleoperation system can mainly be divided into a master part and a slave part. The former is at the operator's side, functioning to receive the operator's commands by employing intention detection techniques, such as body motion monitoring and brain activity detection. The latter is with the remote object, for conducting desired operations issued by the user, normally by robotic arms. To allow users to acclimatize to the remote situation, which is essential for performing precise operations, different types of sensors can be installed at the slave side. Among all sensing techniques, computer vision is treated as the most important, because visual perception is the key functionality of human beings, i.e., computer vision techniques can provide users direct knowledge of the remote situation.

Computer vision is itself a very comprehensive subject. In terms of teleoperation, two CV techniques are of significance: object detection and distance measurement. These can help the user to find and locate desired objects easily, and they are essential to avoid mental/physical fatigue issues and enhance the operation accuracy. In 2014, the deep learning method demonstrated its advantage with the emergence of the family of R-CNN, which shows benefits in terms of robustness compared with traditional methods. Distance measurement methods are based on monocular vision, stereo vision, time-of-flight (ToF), and structured-light-based mechanisms. The integration of the two novel methods is of significance in both teleoperation and automatic manipulation. Figure 1 shows how object detection and distance measurement visually assist the operator in teleoperation. When the target object is localized and classified, and the distance information is learned, the operator can precisely move the arm to the expected position. Similarly, in automatic manipulation, this visual information is essential in making decisions, which could be applied in many broad fields requiring high implementation accuracy, such as remote surgery.

check for updates

Citation: Zhang, A.; Chu, M.; Chen, Z.; Zhou, F.; Gao, S. Object Detection and Distance Measurement in Teleoperation. *Machines* **2022**, *10*, 402. https://doi.org/10.3390/machines10050402

Received: 11 April 2022
Accepted: 17 May 2022
Published: 21 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

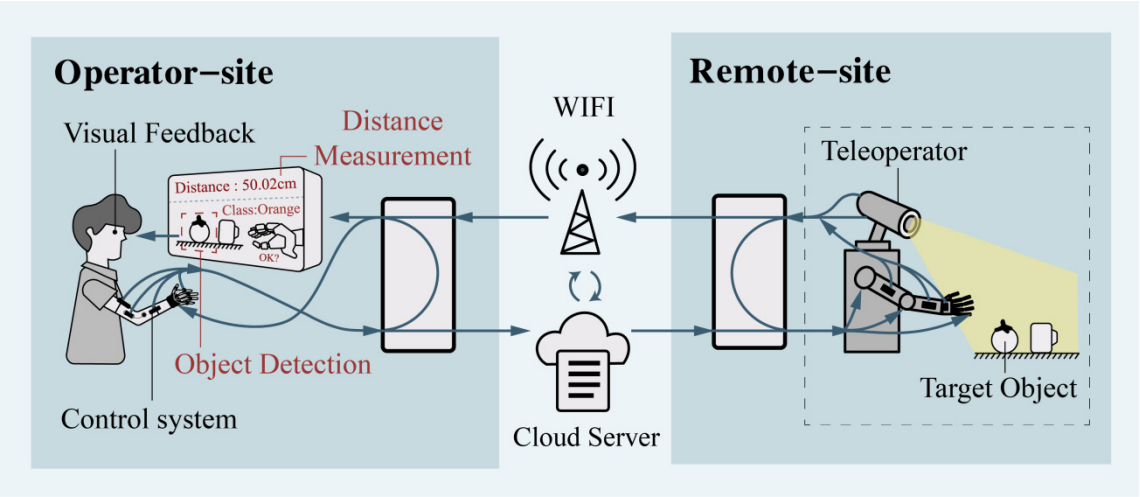


Figure 1. A general schematic showing how object detection and distance measurement assist the operator visually in teleoperation.

The development of object detection and distance measurement has been unprecedentedly fast. Each year, numerous new methods are proposed and implemented in teleoperation, and they advance the field considerably but also boost the entry-level threshold for young scientists and engineers in the meantime. Although there has been some research using object detection and distance measurement methods to assist teleoperation, little work has been carried out to summarize and organize them, making it difficult for people who need to use these techniques to obtain relevant information easily and quickly. To help beginners learn the basic and mainstream methods in object detection and distance measurement, and develop research for specific application scenarios, we composed this review article.

We searched Google Scholar for studies published from 2000 to 2021 on the application of object detection and distance measurement methods based on computer vision in teleoperation and automatic manipulation. Keywords used include "Object detection", "distance measurement", "computer vision", "teleoperation", "automatic manipulation", and their combinations. Computer vision technology has only begun to flourish in recent years; therefore, the combination of object detection and distance measurement methods, which are based on computer vision, with teleoperation technology is still in the early stage. There are few related research papers; thus, we summarized and organized the research papers according to the different application scenarios as much as possible. In addition, some applications in automatic manipulation have been included to show how object detection and distance measurement methods operate together in remote control. A total of 22 studies which applied object detection and distance measurement methods were found. The application scenarios could be divided into five application scenarios: medical surgeries, nuclear decommissioning, space and undersea exploration, various industries, and daily service. All these studies indicate that object detection and distance measurement methods play an important auxiliary role in teleoperation. It also shows that the cross-direction cooperation between teleoperation and automatic manipulation and teleoperation has great potential. Table 1 shows the number of papers published, the number of papers cited in this article, and the methodology used for selection. In addition, in order to enable beginners to quickly learn object detection and distance measurements, and put them into use for their own remote control systems, 45 papers covering the mainstream algorithms of object detection and 36 papers on vision-based distance measurements, which include four mainstream methods (monocular vision, stereo vision, ToF, and structured light), are cited in this article.

Table 1. The number of papers published, the number of papers cited in this article, and the methodology used for selection.

	Overall	Selected	Methodology
Object detection technology	6000–8000	45	Representative mainstream algorithm with strong universality and high performance, which is suitable for beginners to gain a quick overview of the field
Distance measurement technology	More than 10,000	36	
Applications	22	22	All studies that apply object detection and distance measurement to remote control from 2000 to 2021

The rest of this paper is organized as follows. Section 2 reviews the features of traditional objection detection and deep-learning-based object detection methods. This section also includes some paradigms of two-shot detectors and one-shot detectors as well as the generative model and discriminable model of target tracking. Section 3 covers four usual target tracking methods. Section 4 offers applications of these methods in the field of teleoperation in medical surgeries, nuclear decommissioning, space and undersea exploration, various other industries, and daily service. The conclusions and potential trends for future development are presented in Section 5.

2. Object Detection

Object detection is a method used to locate and classify the target object in a given image. As a result, the target object will be located by a bounding box with a category label, which can provide the operators with good operating experience in teleoperation.

In this section, we summarize the structure of object detection and review the mainstream paradigms.

2.1. Object Detection Structure

Object detection has developed rapidly as the algorithms continue to be optimized and updated. In early eras, traditional object detection algorithms were mainly based on hand-crafted feature extraction, which performed well on specific datasets, and were simple and quick. However, traditional detection algorithms have made slow progress in developing high performance. In 2012, Krizhevsky et al. [1] applied deep convolutional neural networks (DCNNs) to successfully classify images. Then, in 2013, Sermanet et al. [2] developed one of the earliest deep-learning-based detectors, OverFeat. Subsequently, object detection experienced a new era of deep learning, and its performance greatly exceeded that of traditional detection algorithms.

From the perspective of top-level structures, object detection can be divided into three steps: proposal generation, feature extraction, and classification.

The proposal generation stage mainly involves the search of possible ROIs (regions of interest) of the target from images containing a large amount of background information. The most direct method is to scan the whole image with different ratios and scales using a sliding window [3–5]. However, due to the need to scan the image line by line, this method is time-consuming, and selective search is later produced through the relationship between pixels [6].

In the feature extraction stage, representative features need to be extracted from the ROIs to facilitate subsequent classification and regression tasks. Many typical traditional target detection methods elaborate feature descriptors to describe features, such as SIFT (scale-invariant feature transformation) [7] and HOG (gradient histogram) [8].

After the first two steps, the targets are contained by a large number of overlapping bounding boxes. There is a need to filter and reorganize to determine each target single box. Then, the target region is classified by a region classifier. Typical classifiers include the support vector machine (SVM) [9] and deformable component-based model (DPM) [10]. Support vector machines are used because of their good performance with a small range of training data. DPM is a flexible model which deals with severe deformation by com-

bining the deformation cost with parts of the object. In addition, Bagging [11], cascading learning [12], and AdaBoost [13] are also widely used.

In recent years, interest in deep learning methods has surged because they have been shown to outperform previous state-of-the-art techniques. First of all, because traditional methods require manual analyses of features of images, many professionals are required, whereas deep learning networks can automatically extract and filter features. In addition, deep learning has advantages in the recognition ability and adaptability of the algorithm. Therefore, in the sections below, we review deep-learning-based detection.

2.2. Detection Paradigms

There are two paradigms of object detection: one-stage and two-stage. Two-stage techniques are performed with two different networks: one is for generating region proposals, and the other is for classification. One-stage techniques can export bounding boxes and classification labels performed with only one network. One-stage detectors are much faster and more desirable for real-time object detection applications, but exhibit a relatively poor performance compared with two-stage detectors. We present the milestones of deep-learning-based detection methods in Figure 2.

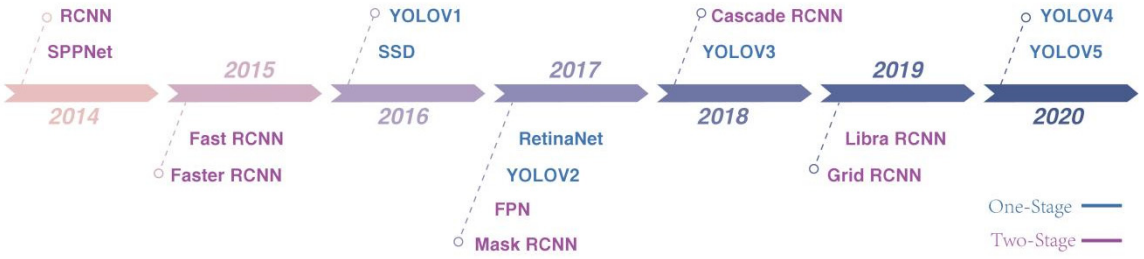


Figure 2. Deep-learning-based detection milestones.

2.2.1. Two-Stage Detectors

Two-stage detectors are divided into two stages. First, proposal regions are generated. Then, feature vectors of the generated proposals are extracted for deep convolutional neural networks to predict the target category.

In 2014, two-stage detectors came into focus with the introduction of R-CNNs [14]. Subsequently, methods such as SPP [15], Fast R-CNN [16] and Faster R-CNN [17] have constantly been derived to further promote the development of two-stage detectors. In 2014, The Faster R-CNN was proposed as a regional proposal network with a milestone significance, which can improve the efficiency of detectors and enable the end-to-end training of detectors. Since then, various approaches have emerged to enhance the performance of Faster R-CNN from different perspectives. For example, FPN [18] implements the processing of scale variance through pyramid prediction. Cascade R-CNN [19] extended the Faster R-CNN to multilevel detectors through a cascade architecture. Mask R-CNN [20] added a Mask branch to refine detection results through multi-task learning. In addition, the Libra R-CNN [21] and Grid R-CNN [22] were derived from Faster R-CNN. Figure 3 shows the principles of R-CNN, Fast R-CNN, Faster R-CNN, and Mask R-CNN.

- RCNN.

R-CNN is a groundbreaking two-stage detector. Its pipeline can mainly be divided into three steps. First, the input image is selectively searched, and the bottom-up region proposal is extracted. Then, the proposal is cropped and resized, and a large convolutional network is used to compute the characteristics of each proposal. Finally, a class-specific linear support vector machine is used to classify each region and a bounding box regressor is used to localize the target object more tightly. An average accuracy of 53.7% (mAP) was achieved on PASCAL VOC 2010.

- Fast RCNN.

Similar to R-CNN, Fast R-CNN still uses selective search to generate proposal regions, but the difference is that all regional suggestions are extracted separately and then classified by a support vector machine classifier. In addition, feature extraction, region classification, and bounding box regression can all be optimized end-to-end, without requiring additional cache space to store features. Compared with R-CNN, it has better accuracy and training speed.

- Faster RCNN.

Faster R-CNN is the third iteration of R-CNN. By adding a regional proposal network (RPN), it tries to eliminate the dependence on selective search, so that the model can fully realize end-to-end training. Faster R-CNN could make predictions at five frames per second on a GPU and achieved state-of-the-art results on many common benchmark datasets such as PASCAL VOC 2007, 2012 and MSCOCO. Currently, detectors based on Faster R-CNN have a large number of variants and are widely used.

- Mask RCNN.

Mask R-CNN is also improved on the basis of Faster R-CNN. Similar to the existing branches for classification and bounding box regression, Mask R-CNN adds a branch to predict segmentation masks by pixel to pixel. In addition, Mask R-CNN adopts a simple layer, named "RoI Align", to retain the corresponding relationship in pixel space, thus eliminating the negative impact of the bias introduced by Faster R-CNN in feature calculation on pixel-to-pixel Mask prediction. This small change greatly improves the accuracy of the mask.

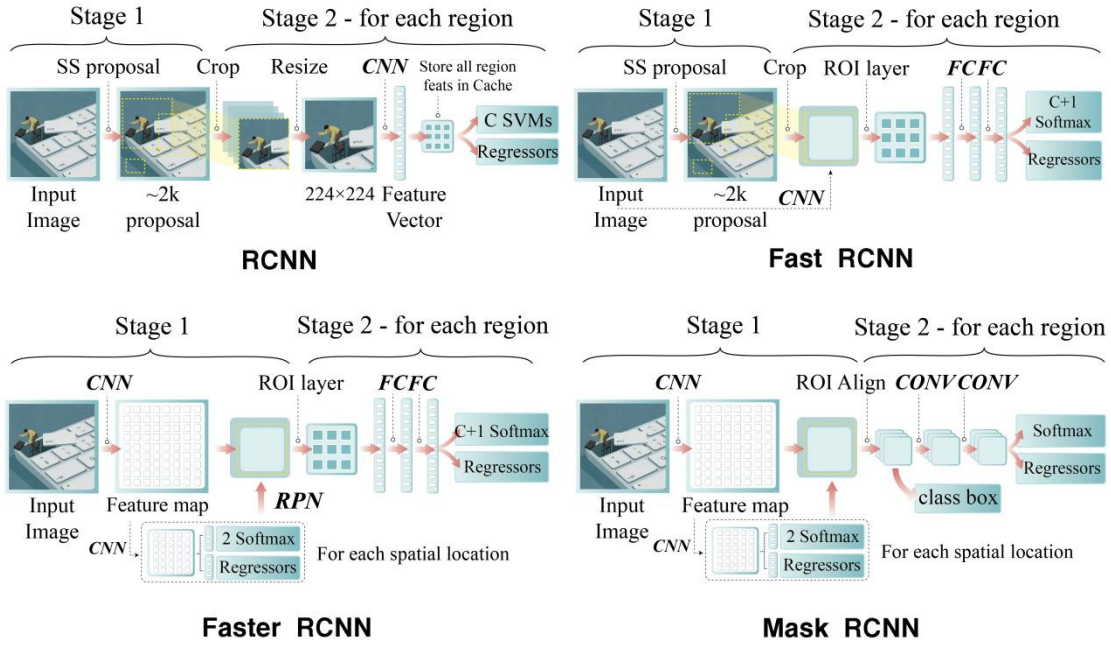


Figure 3. Overview of the principle of two-stage detectors such as R-CNN, Fast R-CNN, Faster R-CNN, and Mask R-CNN.

2.2.2. One-Stage Detectors

Different from two-stage detectors, which splits the detection pipeline into two phases, proposal generation and region classification, one-stage detectors do not have an independent stage for proposal generation. They typically consider all positions on the image as potential objects and try to classify each region of interest as either a background or a target object.

One-stage detectors mainly included OverFeat, SSD variants [23,24], YOLO variants [25–28], and RetinaNet [29]. Figure 4 shows the principles of OverFeat, SSD, YOLO, and RetinaNet.

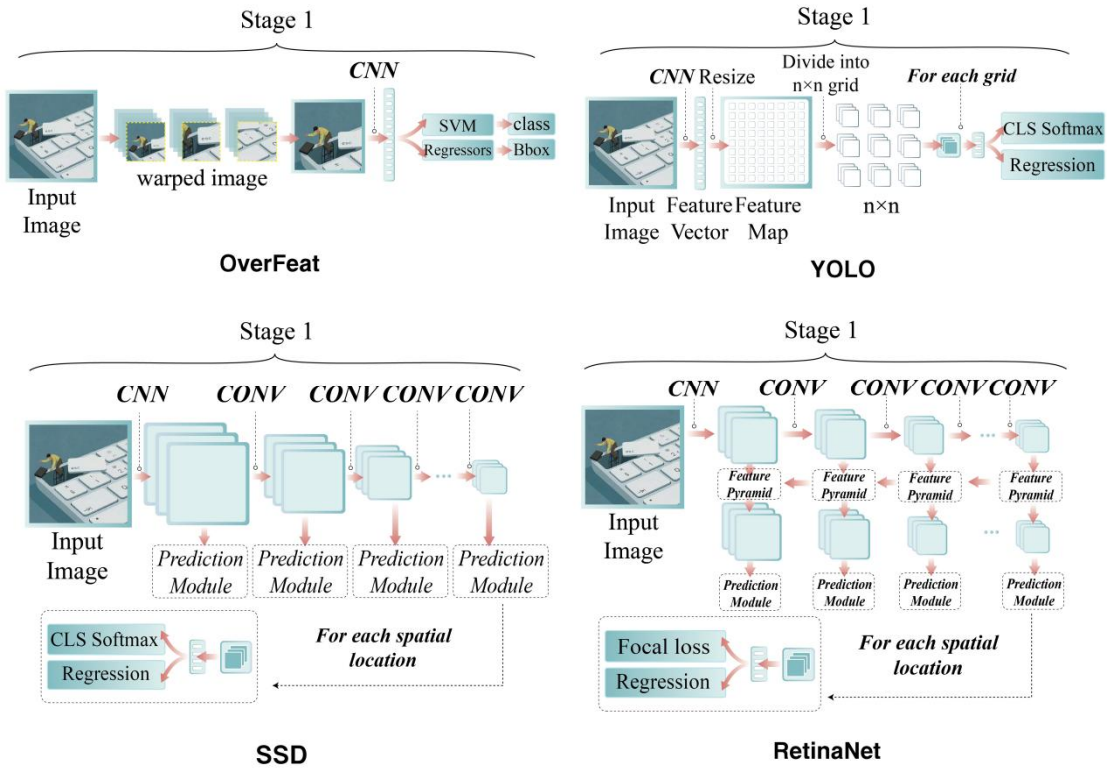


Figure 4. Overview of the principle of one-stage detectors such as OverFeat, YOLO, SSD, and RetinaNet.

- OverFeat.

OverFeat is one of the early successful deep-learning-based one-stage detectors. The main innovation of the OverFeat method is the integration of multi-scale, sliding window, offset pool, and identification, localization, and detection methods based on AlexNet. By using the convolutional layer to share the overlapping region, it has a significant speed advantage compared with R-CNN. However, the training of the classifier and regressor is separated and cannot be optimized together.

- YOLO.

YOLO is a real-time detector that can directly predict boundary boxes and class probabilities from a complete image in a single evaluation. The omitted the proposal generation step can be optimized end-to-end. When the image was inputted into the detector, feature vectors are extracted and the feature vectors are resized into feature maps. Then, the feature map is divided into fixed number of grid cells. Additionally, the class prediction and bounding box used for localization are made for each cell.

- SSD.

SSD is another single-stage detector that addresses YOLO's limitations. It also divides the image into grid cells, but unlike the fixed grid cell prediction adopted by YOLO, a set of anchor points with multiple scales and aspect ratios are generated in each grid cell to discretize the output space of the bounding box.

- RetinaNet.

RetinaNet used feature pyramid network to obtain feature pyramid and used two subnetworks for each feature layer to perform class prediction and bounding-box regression. Before the proposal of RetinaNet, the two-stage approach based on R-CNN applied a

classifier and achieved the highest accuracy of the target detector. In contrast, single-stage detectors may be faster and simpler, but have consistently lagged behind two-stage detectors in accuracy. However, RetinaNet is able to achieve the speed of previous single-stage detectors while exceeding the accuracy of the most advanced two-stage detectors with both speed and accuracy.

2.2.3. Detector Structure

The backbone, as a part of feature extraction in a detector, greatly affects the final performance of object detection. An appropriate backbone can be selected according to the accuracy and efficiency requirements of specific tasks.

When high accuracy is required, compact backbone can be selected, such as VGG [30], ResNet [31], ResNeXt [32], or DenseNet [33]. They usually run on the GPU. When speed and efficiency are pursued, lightweight backbone can be selected. Examples include MobileNet [34], ShuffleNet [35], SqueezeNet [36], and Xception [37], which usually run on the CPU.

2.3. Datasets

There are four datasets commonly used in object detection: Pascal VOC [38], MSCOCO [39], ImageNet [40], and Open Images [41].

- Pascal VOC.

Pascal VOC is a mid-scale dataset. Pascal VOC2007 and VOC2012 are the most widely used, which both contains 20 object categories. They are all split into three subsets (training, validation, and testing). The former has 2501, 2510, and 5011 images, respectively, and the latter has 5717, 5823 and 10,991 images, respectively.

- MSCOCO.

MSCOCO is the most widely available dataset for object detection, which contains 0.2 million images and 80 object categories with annotations concluding bounding boxes and per instance segmentation masks.

- ImageNet.

ImageNet is the largest database for image recognition, which contains more than 14 million images covering more than 20,000 categories. More than one million images were labeled with locations and specific categories.

- Open Images.

Open Images is the largest contemporary dataset, which contains 1.9 million images, 600 categories, and 15.4 million bounding-box annotations. Most of the object position annotations are manually labeled by professionals, ensuring their consistency and accuracy.

Table 2 shows how the detectors introduced in Sections 2.2.1 and 2.2.2 performed on Pascal VOC or COCO dataset.

2.4. Others

To better complete the teleoperation, the techniques of 3D object detection and target tracking can be used. Three-dimensional object detection can achieve a more interactive interface when the teleoperator is in control of the robotic arm. Meanwhile, target tracking can provide online tracking so that the target object can be precisely controlled under remote control.

2.4.1. Three-Dimensional Object Detection

In recent years, 3D object recognition has undergone rapid progress, with advances in 3D deep learning and strong application demands. In contrast to the 2D object objection we introduced above, it is no longer confined to a plane. It can detect not only the object category, but also length, width, height, rotation angle, and other information in

three-dimensional space as well. At present, 3D target detection is in a period of rapid development. It mainly uses a monocular camera, binocular camera, and multi-line lidar to carry out 3D target detection. Lidar can achieve the highest accuracy, but it costs the most. However, with the continuous industrial development of lidar, the cost is constantly reduced. Moreover, there are also some technical solutions for the comprehensive use of a monocular camera and lidar with fewer rays.

Despite the many studies on 3D object detection that are currently being conducted, there are still many problems in its practical application. The first problem is the dynamic environment around the object shelter, truncation, and robustness of the problem. The second problem is that the existing methods mostly depend on the object surface texture or structure, which can easily cause confusion. Finally, there is an issue with meeting the accuracy requirements and algorithm efficiency.

Table 2. Comparison of object detection methods. The mean average precision (mAP) is an index to measure the detection accuracy in object detection.

Name	Year	Author	Type	mAP (%)	Test Time on GPU (s/Image)
R-CNN [14]	2014	Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik	Two-Stage	58.5 on VOC 2007	13
SPP-Net [15]	2014	Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun	Two-Stage	59.2 on VOC 2007	0.14
Fast R-CNN [16]	2015	Ross Girshick	Two-Stage	66.9 on VOC 2007	0.32
Faster R-CNN [17]	2015	Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun	Two-Stage	73.2 on VOC 2007	0.2
YOLOv1 [25]	2016	Joseph Redmon, Santosh Divvala, and Ali Farhadi et al.	One-Stage	63.4 on VOC 2007	0.02
SSD [23]	2016	Wei Liu, Cheng-Yang Fu 1, and Alexander C. Berg et al.	One-Stage	74.3 on VOC 2007	0.017
Mask R-CNN [20]	2017	Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick	Two-Stage	37.1 on COCO	0.2
FPN [18]	2017	Tsung-Yi Lin, Bharath Hariharan, and Serge Belongie et al.	Two-Stage	33.9 on COCO	0.2
YOLOv2 [26]	2017	Joseph Redmon, Ali Farhadi	One-Stage	78.6 on VOC 2007	0.025
RetinaNet [28]	2017	Tsung-Yi Lin, Ross Girshick, and Piotr Dollár et al.	One-Stage	39.1 on COCO	0.122
Cascade R-CNN [19]	2018	Zhaowei Cai, Nuno Vasconcelos	Two-Stage	42.8 on COCO	0.115
YOLOv3 [27]	2018	Joseph Redmon, Ali Farhadi	One-Stage	57.9 on COCO	0.051
Libra R-CNN [21]	2019	Jiangmiao Pang, Huajun Feng, and Dahua Lin et al.	Two-Stage	43.0 on COCO	0.2
Grid R-CNN [22]	2019	Xin Lu, Junjie Yan, and Quanquan Li et al.	Two-Stage	43.2 on COCO	0.2
YOLOv4 [29]	2020	Alexey Bochkovskiy, Chien-Yao Wang et al.	Two-Stage	43.5 on COCO	0.015

2.4.2. Target Tracking

This section reviews the main object tracking algorithms. Early work focused on generative model tracking algorithms, such as the optical flow method, Kalman [42] filtering, particle filtering, Meanshift algorithm, and Camshift algorithm. This method first builds the target model or extracts the target features and searches for similar features in the subsequent frames. Step-by-step iterations achieve target positioning. However, this method also has obvious disadvantages, i.e., the background information of the image has not been fully utilized. Additionally, the appearance of the target itself has characteristics of randomness and diversity, so it is very limited to describe the target to be tracked by a single mathematical model. Specifically, in the case of illumination changes, motion blur, low resolution, target rotation deformation, etc., the establishment of the model will be greatly affected, thus affecting the accuracy of tracking. The establishment of the model has no effective prediction mechanism, and it cannot be solved well when target occlusion occurs.

Since 2000, people have gradually tried to use classical machine learning methods to train classifiers, such as MIL, TLD, support vector machine, structured learning, random forest, multi-instance learning, and metric learning. With the wide application of deep learning, depth features have been applied to target tracking. Subsequently, people began to consider using deep learning to establish a new framework for target tracking. In the past two years, target tracking algorithms based on deep learning have emerged in an endless flow, such as PrDiMP [43], D3S [44], etc., and achieved good results.

3. Distance Measurement

The image processing techniques explained in the previous sections can support the functionalities of identifying the object categories, localizing objects of interest, and continuously tracking target objects in real time during teleoperation. However, they have merely study special two-dimensional (2D) knowledge at the x–y axes, lacking the objects' depth information (z-axis), which is essential for the high level of precision required in teleoperate applications.

To implement depth detection, vision-based methods are broadly used. Among the available methods, monocular vision, stereo vision, time-of-flight (ToF), and structured light technologies are in widespread use.

3.1. Monocular Vision

Monocular vision requires only one camera to capture image information. The general pipeline is shown in Figure 5. The network can mainly be divided into two parts—encoding layers and decoding layers—which consist of a series of operations such as convolution, pooling, and deconvolution to deal with the image input. The network is end-to-end, which means that a dense depth map can be directly obtained after the image is put into the network.

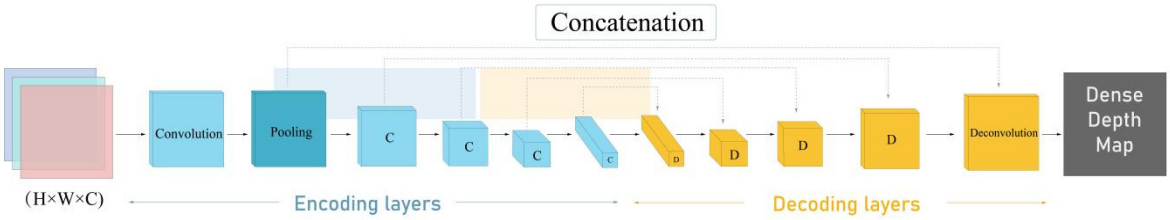


Figure 5. The general pipeline for monocular vision distance measurement.

Inspired by the fact that humans can use prior information about the world to perceive depth information from a single image. Early depth estimation algorithms realized single-image depth estimation by combining existing information, such as geometric relationships between buildings, the sky, and earth [45].

However, this prior information requires certain assumptions. When the assumptions are invalid, these methods are not reliable. In recent years, many deep learning models have been developed for monocular vision in depth estimation [46–49]. With the development of computing hardware and the availability of rich training data, a monocular depth estimation method based on convolutional neural networks (CNNs) [50–56] is proposed, which greatly improves performance. In addition, a lot of work has been carried out on CRFs [57] to achieve further improvements. Some methods [58–62] combine CNN with CRF models to generate depth maps that are more edge-friendly. In addition, depth estimations have advanced by combining closely related data [63,64], such as surface normal flow and optical flow.

3.2. Stereo Vision

Stereo-vision-based methods simulate the way the human brain interprets the distances between objects. The general pipeline is shown in Figure 6. Two RGB images from two cameras with baselines represent the 2D scenes generated by our left and right eyes. By establishing the corresponding relationship between the two images and calculating the parallax (relative displacement), the distance information of the target can be determined. There are two strategies for matching correspondence: the local correspondence method and global correspondence method [65].

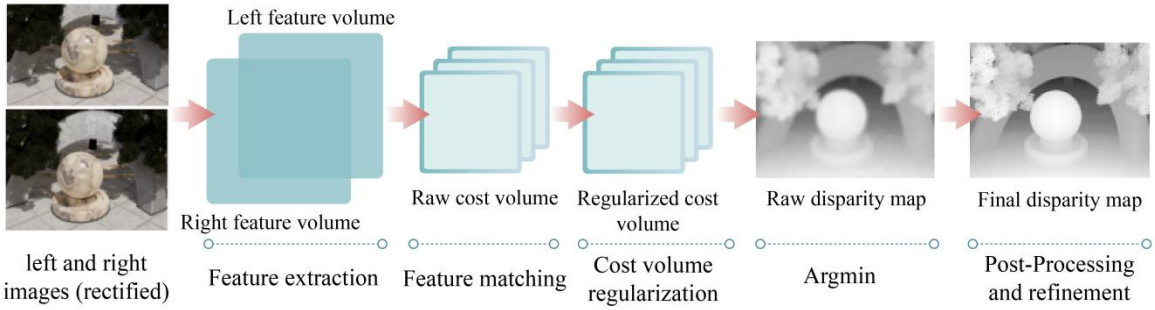


Figure 6. The general pipeline for stereo vision distance measurement.

Local methods take region block matching [66] and gradient-based optimization [67] as sub-methods, and tend to search for corresponding pixels within a limited window or through features such as grayscale and color. The local method has high efficiency due to its low computational complexity. However, it is limited when dealing with large, untextured areas and occlusion boundaries.

Global methods include belief propagation [68], dynamic programming [69], and graph cutting [70], and regard the matching of two images as an optimization problem. By using non-local features, the global method has better performance than the local method.

Although stereo vision has been widely used, there is still a series of problems. For the two strategies mentioned above, the local counterpart algorithm is effective, but not as accurate as the global counterpart algorithm, which has a high computational cost. In addition, performance suffers to some extent due to the lack of information such as the surface texture of the object.

3.3. ToF

Figure 7 shows the working principle of the ToF (time-of-flight) camera. It works by illuminating the scene with a modulated light source and observing the reflected light. The phase shift is measured between illumination and reflection and converts it into distance. The ToF system does not depend on geometric parameters. Typically, ToF ranging systems use pulse modulation [71] or continuous wave (CW) modulated light sources [72].

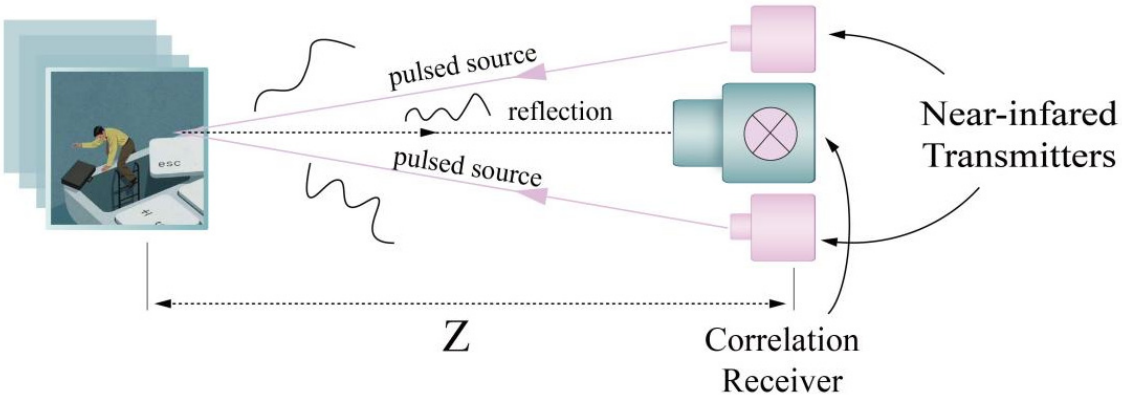


Figure 7. ToF camera operation.

Due to the simple operation of ToF processing in depth map estimation, no moving components can be used to generate dense depth maps, without artifacts caused by occlusion, scene textures, etc. ToF-based methods have strong robustness and can perform good depth measurements in real-time scenes. Other depth estimation systems such as structured light (SL) and stereo vision systems lack these capabilities; therefore, ToF cameras are better in many cases [73].

However, ToF has not become the dominant mode of deep acquisition due to its disadvantages such as low depth resolution and high noise when working in close ranges. In addition, because the components of the emitted light may undergo multiple reflections within the scene, light from different paths ends up being received by a pixel, leading to an incorrect estimation of the corresponding depth [74–76].

3.4. Structured Light

Structured light is a high-precision depth measurement method. The principle of structured light for distance measurement is shown in Figure 8. In this method, the encoding pattern is projected onto the scene, the distorted image is captured by the camera, and then the parallax between the image and the distorted image is calculated. Methods using structured light can be divided into space coding and time coding.

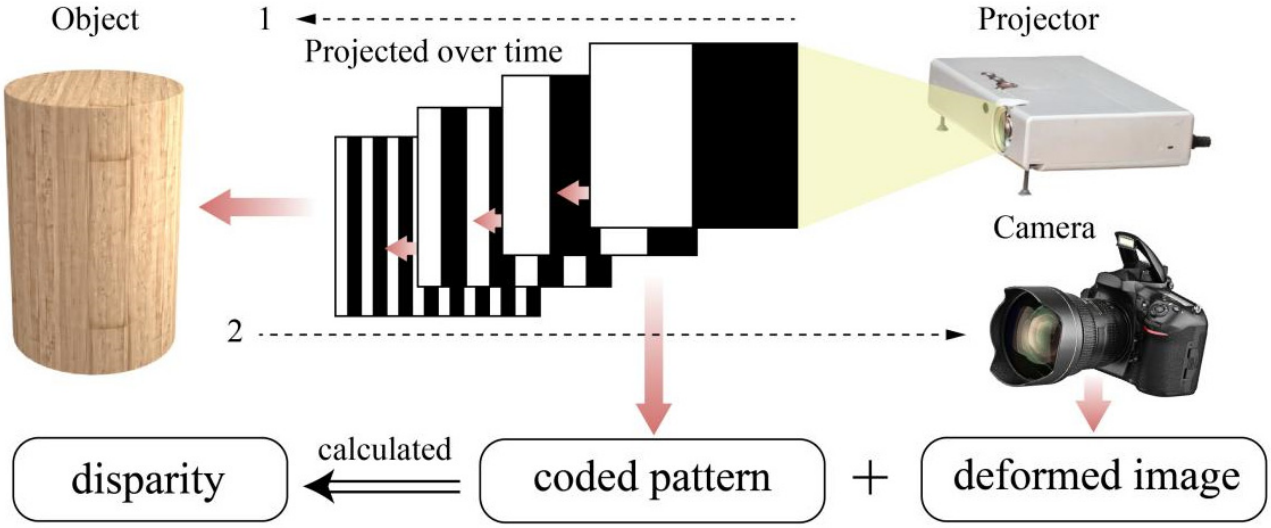


Figure 8. The principle of structured light for performing distance measurements.

Spatial coding techniques generate finely structured light patterns of various colors, shapes and intensities, encoding features with unique tags. The patterns designed by spatial coding technology are single beat patterns; therefore, the spatial coding mode is suitable for dynamic scenes.

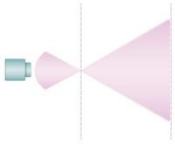
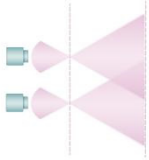
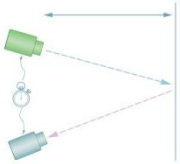
Time coding technology designs structured light patterns based on time series. Time-encoding technology is performed in multi-shot mode, which encodes a pixel in the order of several patterns projected onto the scene. Generally speaking, time coding schemes have high accuracy.

Fringe projection profilometry (FPP) is one of the most widely used structured light methods, which has strong robustness and importance. Hieu Nguyen et al. [77] studied this integration from different perspectives in order to achieve accurate 3D shape reconstruction.

Structured light can solve the problem of stereo matching and improve the performance of measuring untextured areas. In addition, because the reflected pattern is sensitive to ambient light interference, structured light has interference problems outdoors. Therefore, structured light tends to be more suitable for indoor applications.

Table 3 shows a comparison of the four distance measurement methods. Selections could be made according to specific requirements. For example, because different methods are affected differently by light conditions, if the teleoperation system works in dark conditions, structured light and ToF should be taken into consideration.

Table 3. Comparison of distance measurement methods.

Consideration	Monocular Vision	Structured Light	Stereo Vision	Time-of-Flight (ToF)
Working Principle	Focus scenes on the camera's image plane through the lens 	Measure coded optical patterns' variation through the feedback camera 	RGB image feature point matching and indirect triangulation calculation 	Measure the time delay or phase delay of the reflected light 
Response Time (frames per second)	15–400	30–60	15–50	>100
Depth Accuracy	/	0.01–1 mm	<1 mm	1–10 cm
Work Outdoor	No influence	The influence is large, especially in low power	No influence	There is influence, but little
Work in dark conditions	No	Yes	No	Yes
Range	/	Within 5 m	Within 20 m	Within 10 m

4. Applications

Object detection and distance measurement have been used in many aspects of life, including medical surgery, nuclear decommissioning, space exploration, various industries, and everyday services. In these areas, the work is either high-risk or highly repetitive. Therefore, there is an urgent need to ensure the safety of operators or free up the workforce. In addition, time and space constraints require remote operations or personnel changes. Automatic operation and semi-autonomous remote operation can solve these problems to a certain extent, and have their own advantages in different aspects. Automatic robots with vision systems can complete tasks safely, quickly, and accurately, whereas manual operations and visual inspection are time-consuming and inefficient. However, semi-autonomous remote operation is essential when human intervention is necessary for specific tasks that robots cannot currently perform.

Thus, fully autonomous manipulation, semi-autonomous teleoperation, and combinations of the two are essential. The applications are listed below. Figure 9 shows the synthesis of the application of object detection and distance measurement in teleoperation and automatic manipulation.

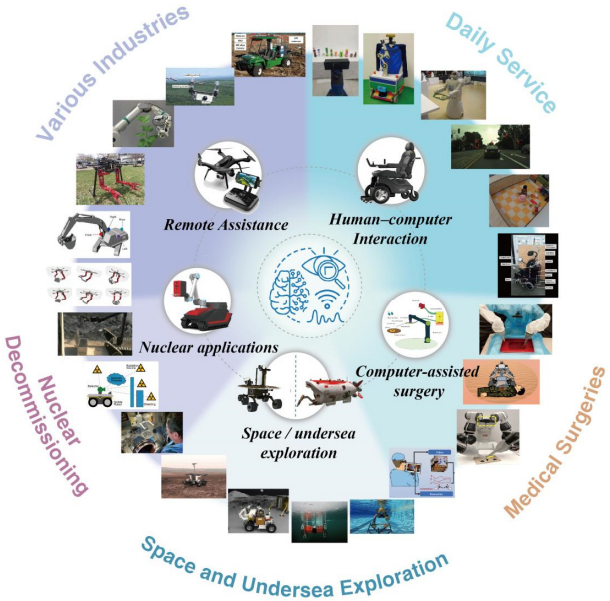


Figure 9. Synthesis of the application of object detection and distance measurement in teleoperation and automatic manipulation.

4.1. Medical Surgeries

In severe environments such as remote rural areas or battlefields, it is often difficult to obtain timely medical assistance. In order to ensure the timeliness of rescue and the safety of rescuers, the ability to fully automate the delivery of medical services by machines or enable medical personnel to remotely control machines to perform medical procedures is especially critical.

Computer-assisted surgery (CAS) [78] is referred to as image-guided surgery, surgical navigation, and 3D computer surgery, using technologies such as 3D imaging and real-time sensing surgical procedures.

One of the most well-known surgical systems in robot-assisted surgery is the Da Vinci robot, which can perform a variety of laparoscopic surgeries, including scaling the surgeon's movements with very little communication delay. Baiquan Su [79] proposed a robot system to complete the task of blood removal surgery. The system consists of a pair of dual cameras, a 6-DOF (degrees of freedom) manipulator, a suction device with a handle fixed on the manipulator and a pump connected to the suction device. In order to thoroughly clean the blood flowing from the incision, a Mask R-CNN frame was used for blood profile detection.

However, due to technical limitations, it is still difficult to achieve the full automation of various medical services, and semi-autonomous remote operation in medical assistance has become a viable predecessor. Md Masudur Rahman et al. [80] proposed a semi-autonomous robot framework to infer high-level commands from the surgeon's movements and then perform them semi-autonomously on the robot. A 3D camera mounted on the top of the robot takes continuous color (RGB) and depth images at 30 frames per second. It uses the YOLOv3 network to detect the 2D object bounding box of each RGB frame. There may be multiple items of the same class in the scenario; therefore, a target tracker is added that uses the output of the YOLOv3 network. The target tracking algorithm is an extension of SORT [81] (Bewley et al., 2016). If the object disappears from the scene and becomes visible again, its position is updated by the tracker. The task execution success rate of the system is as high as 86.6%. These results show that the semi-autonomous remote-control system with high autonomy has the potential for success.

4.2. Nuclear Decommissioning

Radiation from nuclear resources is harmful for animals and humans because their cells' structure and normal function could be destroyed by radiation. In this case, it is crucial to remove radioactive materials from these nuclear facilities. This method could help to reduce harm to the environment, and the by-product of this process may have commercial use. Changing one operating nuclear facility to a safe, non-operational type is known as decommissioning. To finish the process, it has several steps involving characterization, decontaminating, and waste management.

Nowadays, to implement characterization tasks, robotic systems always use electrical power. Teleoperated by a human user, these systems also provide visual information on environments. In order to detect the environment, various sensors are employed to help construct detailed 2D or 3D computer models to perform the characterization process. Liam Cragg and Huosheng Hu [82] designed one new architecture that could enhance the practical decommissioning ability of existing robotic characterization systems with limited functionality. Existing distributed computing, autonomous multi-robot, and Internet robotic architectures are combined to make the characterization ability stronger. The novel architecture could overcome the effects of latency and limited bandwidth of Internet communications, making the system much more efficient.

In addition, remote operation is essential for manipulation in industrial environments, such as nuclear applications. Remote operation in the nuclear industry ensures operator safety and provides a cost-effective solution for manipulating, inspecting, and maintaining nuclear power plants. It can be advantageous to perform operations that operators carry out today to reduce workloads and risks of accidents or contamination. The system

is a remote one-sided structure; therefore, visual feedback is critical. To minimize the risk of human contact, glove boxes with sealed volumes are used to perform handling. Although the operator could perform complex handling tasks with gloves, there is a significant risk of breaking the seal. Ozan Tokatli et al. [83] presented a robotic synthesizer demonstrator for handling nuclear material in a nuclear glove box. For object detection, they chose models similar to "look once only" (YOLO) because they can produce detection results at a faster rate (45 frames per second). A deep learning model was trained with the dataset, using custom annotated images representing the glove box environment, feeding the output detection results to the tracking algorithm during the whole process. Moreover, it uses a monocular vision approach which could capture depth images with an analog RGB-D camera mounted on the robot's wrist and taking it as an input to the neural network. Manuel Bandala [84] introduced the development of a vision-based semi-automatic object grasping system which could be employed in a hydraulically driven dual manipulator retraction robot. Based on the SDP model, a novel approach was applied to a hydraulic manipulator, with the function of signal calibration, system identification, and nonlinear control design for a situation where dynamic characteristics changed over time. Kui Qian et al. [85] designed a small teleoperated robot for nuclear radiation and chemical leak detection. In this system, a 5-DOF vehicle manipulator is used to perform sample collection. During the process, the system receives all the data, including sound, image, and other sensor information in real time. This information provides a basis for mission command and decision making. The control range for this system is 5000 m maximum in open areas and 1750 m maximum in urban areas. Telepresence is brought to the operator, demonstrating good performance in tests.

4.3. Space and Undersea Exploration

Robots can be employed to investigate unknown things and explore environments that are inaccessible to humans. The use of them satisfies the ambition of scientists, engineers, and explorers in numerous fields.

Due to the complexity and unpredictability of the lunar surface, teleoperation is the most important control method for the operation of a rover in space exploration. To support the teleoperation of a rover, computer vision is an important technology. There are many technical challenges, such as rapid positioning of the landing point and seamless high-resolution mapping of the landing site during the powered descent stage and lunar surface exploration. Teleoperation based on computer vision could effectively solve these problems. Wang J., Li J., and Wang S. et al. [86] presented an important study providing support for the topographical analysis of a landing site and mission planning for subsequent teleoperations. To finish the positioning of the landing point for the Yutu-2rover, they used a stereo vision method to obtain descent images. James Bird [87] introduced a simulator-based approach using computer vision to address the problem of obtaining useful images during deep-space travel. Their results showed that the simulator provides a training environment which could be used to train models of features not yet observed by humans. The research also presented an immersive and adaptive environment, allowing the function of navigation in a novel way. Moreover, a teleoperated robotic hand [88] has been developed for the European Space Agency by the German Aerospace Center (DLR) for a lunar rover prototype. Good vision of the robot contributes to this assessment ability. Good vision and force feedback produced nearly 100% correct assessments in grasp success tests.

In undersea exploration, robots could also be applied in various ways. Ocean One [89] is one kind of humanoid underwater robot which has a specific structure for underwater manipulation. There is a remote connection between a humanoid robot and a human pilot, which makes it possible to perform dexterous manipulation in deep sea. While avoiding micromanagement, whole-body controller-coordinated manipulation, posture, and a set of haptic and visual human interfaces, which could facilitate intimate interactions. In [90], an underwater vehicle employed physical underwater light transport models, with target ocean and mission parameters to obtain clearer images during ocean exploration.

In [91], Lu H. et al. proposed one computing model with YOLO to achieve the real-time underwater recognition and tracking of multi-objects. In a comparison experiment, YOLO demonstrated approximately the same performance as MedianFlow. However, YOLO had the advantage on implementing detection on each frame and did not need to specify the bounding box to initialize it, but MedianFlow needs the bounding box.

4.4. Various Industries

Aerial manipulation is also a challenging task. Often, researchers choose to use unmanned aerial vehicles to finish this difficult task. This technique fits various applications, and most of them are dangerous to human operators. With one RGB-D camera and dual manipulators, R. Pablo et al. [92] proposed a system for grasping known objects with the unmanned aerial vehicle. Three different devices tested CNN algorithms for performing tasks of object detection to determine which kind of CNN could fit the system requirements for the manipulation task and the payload limitations of the aerial manipulator. Faster RCNN, SSD300, and YOLO (tiny YOLOv2) have tested. Under the hardware of one Jetson TX1 with IrisGPU, the average computational times were 0.47 s, 0.113 s, and 0.051 s. S. Hussmann et al. [93] established a robot vision system that demonstrates a real-word problem on a small scale. When a container ship has to be loaded using a minimum storage area, it needs the user to deliver the range data of the measured objects for the robot system. A 3D ToF camera on top of the robot system was mounted. In this case, the robot system can now pick up the containers correctly and place them on the container ship.

Fruit-picking operations are labor-intensive; therefore, it is urgent to realize automatic fruit-picking to improve labor efficiency. The technique of artificial intelligence and production has developed rapidly over the years; as such, we could replace manual fruit-picking with machines. Researchers have developed a variety of robots for packing based on the premise that machines could precisely and quickly recognize the fruits for picking. S. Wan et al. [94] proposed an accurate and real-time image-based multi-class fruit detection system through R-CNN. This system could complete tasks such as facilitating innovative higher-level farm tasks such as yield mapping and robotic harvesting. T. Nan et al. [95] used a Fog Robotic system to perform grasping tasks through Mask R-CNN and other intelligent grasping systems. It could instruct robots to finish generalized human-compliant object pickup and manipulation tasks. With a teleoperated climbing robot and enhanced deep learning algorithm, B. Ramalingam [96] proposed an aircraft surface to perform object detection. Additionally, an enhanced SSD MobileNet framework was added to improve the detection accuracy (96.2%) of aircraft surface defects (with an average confidence level of 97%) and stains (with a 94% confidence level). In contrast, compared with the conventional SSD MobileNet and other defect detection algorithms, the proposed system achieved better detection accuracy because of removing most of the unwanted background data.

4.5. Daily Service

In domestic areas, remote assistance is also required in specific situations. T. Sano [97] trained Mask R-CNN with a COCO dataset, so that their robot could detect an area where children were. In addition, they used a monocular approach to estimate the distance between a toddler and the robot through a deep neural network. The accuracy of distance estimation was over 65%.

L. Lecrosnier et al. [98] developed an advanced driver assistance system (ADAS) to improve autonomy for disabled people on an innovative electric wheelchair. Based on the detection, depth estimation, localization, and tracking of objects in a wheelchair's indoor environment, namely, doors and door handles, they proposed an adaptation of the YOLOv3 object detection algorithm to solve the problem. In addition, with deep-learning-based data augmentation, they increased their dataset's diversity. Then, using an Intel RealSense camera, they showcased their depth estimation approach. Finally, they demonstrated the 3D object tracking approach based on the SORT algorithm. They

implemented different experiments in a controlled indoor environment to validate all of the developments. Using their own dataset, distance estimation and object tracking were tested. Their models were re-trained with two datasets, involving the open MCIndoor20000 dataset and their own ESIGELEC dataset. They observed 85% and 90% precision for the classification of handles and doors under recall rates of 0.29% and 0.80%, respectively, on a set of 866 images. B. Thomas et al. [99] also demonstrated an autonomous service robot for domestic environments, including diverse abilities such as manipulation with an image-based method, human–robot interaction, person detection and tracking, object detection, and classification. To complete "dirty work" such as ordering, fetching, and sending food settlements in a restaurant, Q. Yu et al. [100] designed a restaurant service robot for the customers in a robot restaurant. They employed a segmented positioning method, which could be applied to consider the positioning costs and accuracy requirements in the different stages. They chose to adopt the shape-based matching tracking method to navigate the robot to the object. The restaurant service robot could provide its real-time coordinates with ± 2 cm positioning precision whatever its initial position.

5. Challenges

5.1. Challenges in Computer Vision

Even though computer vision has experienced great progress in recent years, an accurate and robust approach in object detection and distance measurement based on computer vision remains a great challenge today. The typical challenges of computer vision in the context of teleoperation and automatic manipulation can be summarized in five points.

- Deformation.

In remote rescue, the victims may have different postures, which makes object detection very difficult. If the object detector is trained to only find people standing or running, it may not be able to spot people lying on the ground or bent over, thus overlooking the most necessary situation in which to help the injured.

- Occlusion.

Occlusion (partial occlusion/complete occlusion) may affect the calculation of background frames. However, occlusion can easily occur. For example, in a rescue scenario, when the injured are blocked by other objects such as stones around them and cannot be recognized, it will also lead to timely rescue.

- Illumination changes.

Illumination strongly affects the appearance and characteristics of objects and backgrounds, such as the surface gloss and texture clarity of the same object when the light is turned on or off indoors.

- Cluttered or textured background.

When the background is messy or textured, the object to be recognized may be mixed into the background, making it difficult to be recognized. For example, when different objects with similar textures and colors are stacked together, it is difficult to find the target object to be recognized.

5.2. Challenges in Human Experience

The human experience here refers to using feelings and experience in remote operation. First, data communication efficiency plays an important role. To achieve teleoperation, hardware needs to wirelessly send and receive visual data (e.g., image and video data) and associated information in real time. The size of visual data is often large, and it is necessary to consider the format, space, and transmission type to efficiently store and process the data. Exchanging different types of data and vast amounts of information in real time may deteriorate wireless communication performance. If communication performance deteriorates, errors including compatibility problems may occur in collecting

and transmitting information due to data interference or transmission errors between platforms. In addition, it takes a long time to process data, and there may be a time delay in the operation. Communication errors or delay problems affect safety, work efficiency, and accuracy in performing teleoperation tasks. Therefore, establishing an efficient communication channel is important for the teleoperation of excavators. Moreover, if too much information is provided visually, it may not be able to be digested during tasks and likely lowers concentration and work efficiency. In this context, research on how to optimally deliver information feedback via robust interfaces should be conducted in terms of the contents of visual presentation, information format, and the relationship with other sensory feedback.

6. Possible Future Directions and Conclusions

To address the challenges in computer vision listed above, in the face of deformation, occlusion and illumination changes, the dataset can be expanded by sampling as much as possible, or the data can be generated as close to the real sample as possible by a generating adversarial network (GNN) [101,102]. In addition, when the part of the object is blocked, the physical structure of the object can be inferred through modal perception and modal segmentation to obtain the blocked part [103,104]. For cluttered or textured backgrounds, disturbances can be evaluated by encoding the relationships between objects [105].

Computer vision mostly acts as an auxiliary technique to make operators manipulate in their element, which is of great significance in remote operation. In the foreseeable future, the use of computer vision in remote control will progress towards the following two trends.

First, the data fusion of multiple sensors will further understandings of a specific scene for operators. For a remote-control system, diverse sensors are needed to gain information from all kinds of dimensions. However, this information is relatively independent of each other. Thus, in the near future, a good direction would be to combine visual data with other data to better achieve information synchronization. For example, when attention is focused on a certain object, a user can not only localize, track, and classify it, but also accurately learn some of its other properties such as hardness and smoothness.

In addition, physical modeling in teleoperation could be a good direction. To manipulate objects in a scene, the position of grasping and the amount of force applied are also important. Therefore, physical modeling can be used to simulate the effect of different forces and different positions of grasp on the object. In this way, in the process of real-time control, it is beneficial for the operator to predict in advance and control the object more robustly.

7. Conclusions

This article demonstrates the use of computer vision in remote control, which concludes object detection and distance measurement. The former provides the teleoperated system with an ability to accurately localize and classify the target object. The latter makes the system more comfortable for operators with the distance feedback. Moreover, it also brings better performance for localization. The applications of object detection and distance measurement in teleoperation and automatic manipulation have been reviewed as well.

For object detection, two-stage detectors (R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN, etc.) and one-stage detectors (OverFeat, YOLO, SSD, RetinaNet, etc.) have been introduced. The performance of each one listed can be used as a reference for choosing and putting it into use. Four distance measurement methods (monocular, stereo, ToF, structured light) have been reviewed in detail, with principles presented for each and comparisons drawn between them. The review of the five-application scene (medical surgeries, nuclear decommissioning, space and undersea exploration, industries, and daily service) reveals the considerable prospects.

For beginners, there are some suggestions could be given. First, beginners could learn how to determine object detection and distance measurement methods for remote operation systems from existing studies in similar application scenarios, which have been

summarized and organized as much as possible in Section 4. If there is no fit or match, they need to make their own selections. For selection of the object detection method, a one-stage detector can be selected if rapid real-time speed is desired, whereas a two-stage detector is preferred if higher accuracy is desired but speed is not required as urgently. For the selection of distance measurement methods, trade-offs can be made based on the comparisons between the four methods listed in Table 3. The two technologies process the obtained video or image information and feed back to the operator concurrently. This study focused on the integration of object detection and distance measurement methods in remote control; thus, the performance could be evaluated by checking whether the visual feedback is useful for operators to perform more accurate operations with less fatigue over a longer time.

Author Contributions: Writing—original draft preparation, A.Z.; writing—review and editing, A.Z. and M.C.; visualization, Z.C.; supervision, F.Z. and S.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Natural Science Foundation under Grant 62171014 and Grant 61803017, and in part by Beihang University under Grant KG12090401 and Grant ZG216S19C8.

Institutional Review Board Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *NIPS* **2012**, *60*, 84–90. [CrossRef]
2. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. Integrated recognition, localization and detection using convolutional networks. *arXiv* **2013**, arXiv:1312.6229.
3. Vedaldi, A.; Gulshan, V.; Varma, M.; Zisserman, A. Multiple kernels for object detection. In Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV), Kyoto, Japan, 29 September–2 October 2009.
4. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Kauai, HI, USA, 8–14 December 2001.
5. Harzallah, H.; Jurie, F.; Schmid, C. Combining efficient object localization and image classification. In Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV), Kyoto, Japan, 29 September–2 October 2009.
6. Uijlings, J.R.R.; Van de Sande, K.E.A.; Gevers, T.; Smeulders, A.W.M. Selective Search for Object Recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
7. Lowe, D.G. Object recognition from local scale-invariant features. In Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV), Kerkyra, Greece, 20–27 September 1999.
8. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005.
9. Hearst, M.A.; Dumais, S.T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Their Appl.* **1998**, *13*, 18–28. [CrossRef]
10. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1627–1645. [CrossRef]
11. Opitz, D.; Maclin, R. Popular ensemble methods: An empirical study. *J. Artifi. Intel. Res.* **1999**, *11*, 169–198. [CrossRef]
12. Viola, P.; Jones, M.J. Robust Real-Time Face Detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [CrossRef]
13. Freund, Y.; Schapire, R.E. Experiments with a new boosting algorithm. In Proceedings of the 13th International Conference on International Conference on Machine Learning (ICML), Bari, Italy, 3–6 July 1996.
14. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
15. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
16. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
17. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
18. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.

19. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
20. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
21. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra R-CNN: Towards Balanced Learning for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 821–830.
22. Lu, X.; Li, B.; Yue, Y.; Li, Q.; Yan, J. Grid R-CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7363–7372.
23. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot Multibox Detector. In Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016.
24. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. DSSD: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659, 2017.
25. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
26. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
27. Redmon, J.; Farhadi, A. Yolov3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
28. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
29. Lin, T.; Goyal, P.; Girshick, R.B.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Las Vegas, NV, USA, 22–29 October 2017.
30. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
32. Xie, S.; Girshick, R.; Doll'ar, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
33. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
34. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
35. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856. Available online: https://ieeexplore.ieee.org/document/8578814/ (accessed on 16 May 2022).
36. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 mb model size. *arXiv* **2016**, arXiv:1602.07360.
37. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
38. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
39. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Doll´ar, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the 13th European Conference (ECCV), Zurich, Switzerland, 6–12 September 2014.
40. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
41. Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Duerig, T.; et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv* **2018**, arXiv:1811.00982. [CrossRef]
42. Najafzadeh, N.; Fotouhi, M.; Kasaei, S. Object tracking using Kalman filter with adaptive sampled histogram. In Proceedings of the 23rd Iranian Conference on Electrical Engineering, Tehran, Iran, 10–14 May 2015; pp. 781–786.
43. Danelljan, M.; Van, G.L.; Timofte, R. Probabilistic Regression for Visual Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) IEEE, Seattle, WA, USA, 13–19 June 2020; pp. 7183–7192.
44. Lukezic, A.; Matas, J.; Kristan, M. D3S–A Discriminative Single Shot Segmentation Tracker. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPPR), Seattle, WA, USA, 13–19 June 2020.
45. Hoiem, D.; Efros, A.A.; Hebert, M. Automatic photo pop-up. In Proceedings of the SIGGRAPH05: Special Interest Group on Computer Graphics and Interactive Techniques Conference (ACM SIGGRAPH 2005 Papers, 2005), Los Angeles, CA, USA, 31 July–4 August 2005; pp. 577–584.
46. Hao, S.; Zhou, Y.; Guo, Y. A brief survey on semantic segmentation with deep learning. *Neurocomputing* **2020**, *406*, 302–321.
47. Lai, Z.; Lu, E.; Xie, W. Mast: A memory-augmented self-supervised tracker. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 6479–6488.

48. Wang, K.; Peng, X.; Yang, J.; Lu, S.; Qiao, Y. Suppressing uncertainties for largescale facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 6897–6906.
49. Zeng, N.; Zhang, H.; Song, B.; Liu, W.; Li, Y.; Dobaie, A.M. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing* **2018**, *273*, 643–649.
50. Chakrabarti, A.; Shao, J.; Shakhnarovich, G. Depth from a single image by harmonizing overcomplete local network predictions. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS), Barcelona, Spain, 5–10 December 2016.
51. Eigen, D.; Fergus, R. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
52. Eigen, D.; Puhrsch, C.; Fergus, R. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 8–13 December 2014.
53. Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Tao, D. Deep ordinal regression network for monocular depth estimation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
54. Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; Navab, N. Deeper depth prediction with fully convolutional residual networks. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016.
55. Lee, J.-H.; Heo, M.; Kim, K.-R.; Kim, C.-S. Single-Image Depth Estimation Based on Fourier Domain Analysis. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR, Salt Lake City, UT, USA, 18–23 June 2018.
56. Roy, A.; Todorovic, S. Monocular Depth Estimation Using Neural Regression Forest. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
57. Lafferty, J.; McCallum, A.; Pereira, F.C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001), Williamstown, MA, USA, 28 June–1 July 2001; pp. 282–289.
58. Heo, M.; Lee, J.; Kim, K.R.; Kim, C.S. Monocular depth estimation using whole strip masking and reliability-based refinement. In Proceedings of the 15th European Conference (ECCV), Munich, Germany, 8–14 September 2018.
59. Li, B.; Shen, C.; Dai, Y.; Hengel, A.V.D.; He, M. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
60. Liu, F.; Shen, C.; Lin, G. Deep convolutional neural fields for depth estimation from a single image. In Proceedings of the 15th IEEE International Conference on Computer Vision, Boston, MA, USA, 7–12 June 2015.
61. Xu, D.; Ricci, E.; Ouyang, W.; Wang, X.; Sebe, N. Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
62. Xu, D.; Wang, W.; Tang, H.; Liu, H.; Sebe, N.; Ricci, E. Structured attention guided convolutional neural fields for monocular depth estimation. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
63. Qi, X.; Liao, R.; Liu, Z.; Urtasun, R.; Jia, J. Geonet: Geometric neural network for joint depth and surface normal estimation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
64. Xu, D.; Ricci, E.; Ouyang, W.; Wang, X.; Sebe, N. PAD-Net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
65. Yang, Q.; Wang, L.; Yang, R.; Stewenius, H.; Nister, D. Stereo matching with color- weighted correlation, hierarchical belief propagation, and occlusion handling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 492–504. [CrossRef]
66. Hirschmuller, H. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 328–341. [CrossRef]
67. Pham, C.C.; Jeon, J.W. Domain transformation-based efficient cost aggregation for local stereo matching. *IEEE Trans. Circuits Syst. Video Technol.* **2013**, *23*, 1119–1130. [CrossRef]
68. Yang, Q.; Wang, L.; Yang, R. Real-time global stereo matching using hierarchical belief propagation. In Proceedings of the British Machine Vision Conference, Edinburgh, UK, 4–7 September 2006; BMVA Press: Guildford, UK, 2006; pp. 101.1–101.10. [CrossRef]
69. Ohta, Y.; Kanade, T. Stereo by Intra- and Inter-Scanline Search Using Dynamic Programming. *IEEE Trans. Pattern Anal. Mach. Intell.* **1985**, *7*, 139–154. [CrossRef]
70. Boykov, Y.; Veksler, O.; Zabih, R. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 1222–1239. [CrossRef]
71. Moring, I.; Heikkinen, T.; Myllyla, R.; Kilpela, A. Acquisition of three-dimensional image data by a scanning laser rangefinder. *Opt. Eng.* **1989**, *28*, 897–905. [CrossRef]
72. Beheim, G.; Fritsch, K. Range finding using frequency-modulated laser diode. *Appl. Opt.* **1986**, *25*, 1439–1442. [CrossRef]

73. Schmidt, M. Analysis, Modeling and Dynamic Optimization of 3d Time-of-Flight Imaging Systems. Ph.D. Thesis, Ruperto-Carola University, Heidelberg, Germany, 2011. Available online: http://www.ub.uni-heidelberg.de/archiv/12297/ (accessed on 16 May 2022).
74. Gupta, M.; Nayar, S.K.; Hullin, M.B.; Martin, J. Phasor imaging: A generalization of correlation-based time-of-flight imaging. *ACM Trans. Graph. (ToG)* **2015**, *34*, 156. [CrossRef]
75. Whyte, R.; Streeter, L.; Cree, M.J.; Dorrington, A.A. Resolving multiple prop- agation paths in time of flight range cameras using direct and global separation methods. *Opt. Eng.* **2015**, *54*, 113109. [CrossRef]
76. Geng, J. Structured-light 3D surface imaging: A. tutorial. *Adv. Opt. Photonics* **2011**, *3*, 128–160. [CrossRef]
77. Nguyen, H.; Wang, Y.; Wang, Z. Single-shot 3D shape reconstruction using structured light and deep convolutional neural networks. *Sensors* **2020**, *20*, 3718. [CrossRef]
78. Lichiardopol, S. A survey on teleoperation. *Tech. Univ. Eindh. DCT Rep.* **2007**, *20*, 40–60.
79. Su, B.; Yu, S.; Li, X.; Gong, Y.; Li, H.; Ren, Z.; Xia, Y.; Wang, H.; Zhang, Y.; Yao, W.; et al. Autonomous Robot for Removing Superficial Traumatic Blood. *IEEE J. Transl. Eng. Heath Med.* **2021**, *9*, 2600109. [CrossRef]
80. Rahman, M.M.; Balakuntala, M.V.; Gonzalez, G.; Agarwal, M.; Kaur, U.; Venkatesh, V.L.; Sanchez-Tamayo, N.; Xue, Y.; Voyles, R.M.; Aggarwal, V.; et al. SARTRES: A semi-autonomous robot teleoperation environment for surgery. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **2021**, *9*, 376–383. [CrossRef]
81. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468.
82. Cragg, L.; Huosheng, H. Application of mobile agents to robust teleoperation of internet robots in nuclear decommissioning. In Proceedings of the IEEE International Conference on Industrial Technology, Maribor, Slovenia, 10–12 December 2003; Volume 2, pp. 1214–1219. [CrossRef]
83. Tokatli, O.; Das, P.; Nath, R.; Pangione, L.; Altobelli, A.; Burroughes, G.; Jonasson, E.T.; Turner, M.F.; Skilton, R. Robot-Assisted Glovebox Teleoperation for Nuclear Industry. *Robotics* **2021**, *10*, 85. [CrossRef]
84. Bandala, M.; West, C.; Monk, S.; Montazeri, A.; Taylor, C.J. Vision-Based Assisted Tele-Operation of a Dual-Arm Hydraulically Actuated Robot for Pipe Cutting and Grasping in Nuclear Environments. *Robotics* **2019**, *8*, 42. [CrossRef]
85. Qian, K.; Song, A.; Bao, J.; Zhang, H. Small teleoperated robot for nuclear radiation and chemical leak detection. *Int. J. Adv. Robot. Syst.* **2012**, *9*, 70. [CrossRef]
86. Wang, J.; Li, J.; Wang, S.; Yu, T.; Rong, Z.; He, X.; You, Y.; Zou, Q.; Wan, W.; Wang, Y.; et al. Computer vision in the teleoperation of the Yutu-2 rover. *Remote Sens. Spat. Inf. Sci. ISPRS Geospat. Week* **2020**, *3*, 595–602. [CrossRef]
87. Bird, J.; Petzold, L.; Lubin, P.; Deacon, J. Advances in deep space exploration via simulators & deep learning. *New Astron.* **2021**, *84*, 101517.
88. Lii, N.Y.; Chen, Z.; Pleintinger, B.; Borst, C.H.; Hirzinger, G.; Schiele, A. Toward understanding the effects of visual-and force-feedback on robotic hand grasping performance for space teleoperation. In Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 18–22 October 2010; pp. 3745–3752.
89. Kunz, C.; Murphy, C.; Camilli, R.; Singh, H.; Bailey, J.; Eustice, R.; Jakuba, M.; Nakamurq, K.-I.; Roman, C.; Sato, T.; et al. Deep sea underwater robotic exploration in the ice-covered arctic ocean with AUVs. In Proceedings of the 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, Nice, France, 22–26 September 2008; pp. 3654–3660.
90. Song, Y.; Sticklus, J.; Nakath, D.; Wenzlaff, E.; Koch, R.; Köser, K. Optimization of multi-led setups for underwater robotic vision systems. In Proceedings of the International Conference on Pattern Recognition, Federal Republic of (DEU), Kiel, Germany, 10–15 January 2021; Springer International Publishing: Cham, Switzerland, 2021; pp. 390–397.
91. Lu, H.; Uemura, T.; Wang, D.; Zhu, J.; Huang, Z.; Kim, H. Deep-sea organisms tracking using dehazing and deep learning. *Mobile Netw. Appl.* **2020**, *25*, 1008–1015. [CrossRef]
92. Ramon-Soria, P.; Arrue, B.C.; Ollero, A. Grasp Planning and Visual Servoing for an Outdoors Aerial Dual Manipulator. *Engineering* **2019**, *6*, 77–88. [CrossRef]
93. Hussmann, S.; Liepert, T. Robot vision system based on a 3D-ToF camera. In Proceedings of the 2007 IEEE Instrumentation & Measurement Technology Conference IMTC 2007, Warsaw, Poland, 1–3 May 2007; pp. 1–5.
94. Wan, S.; Goudos, S. Faster R-CNN for multi-class fruit detection using a robotic vision system. *Comput. Netw.* **2019**, *168*, 107036. [CrossRef]
95. Tian, N.; Chen, J.; Zhang, R.; Huang, B.; Goldberg, K.; Sojoudi, S. A fog robotic system for dynamic visual servoing. In Proceedings of the IEEE International Conference on Robotics and Automation, Montreal, QC, Canada, 20–24 May 2019; pp. 1982–1988. [CrossRef]
96. Ramalingam, B.; Manuel, V.H.; Elara, M.R.; Vengadesh, A.; Lakshmanan, A.K.; Ilyas, M.; James, T.J.Y. Visual inspection of the aircraft surface using a teleoperated reconfigurable climbing robot and enhanced deep learning technique. *Int. J. Aerosp. Eng.* **2019**, *2019*, 5137139. [CrossRef]
97. Sano, T.; Horii, T.; Abe, K.; Nagai, T. Explainable Temperament Estimation of Toddlers by a Childcare Robot. In Proceedings of the 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Naples, Italy, 31 August–4 September 2020; pp. 159–164. [CrossRef]
98. Lecrosnier, L.; Khemmar, R.; Ragot, N.; Decoux, B.; Rossi, R.; Kefi, N.; Ertaud, J.Y. Deep learning-based object detection, localisation and tracking for smart wheelchair healthcare mobility. *Int. J. Environ. Res. Public Health* **2021**, *18*, 91. [CrossRef]

99. Breuer, T.; Giorgana Macedo, G.R.; Hartanto, R.; Hochgeschwender, N.; Holz, D.; Hegger, F.; Jin, Z.; Müller, C.; Paulus, J.; Reckhau, M.; et al. Johnny: An autonomous service robot for domestic environments. *J. Intel. Robot. Syst.* **2012**, *66*, 245–272. [CrossRef]
100. Yu, Q.; Yuan, C.; Fu, Z.; Zhao, Y. An autonomous restaurant service robot with high positioning accuracy. *Ind. Robot. Int. J.* **2012**, *39*, 271–281. [CrossRef]
101. Wang, X.; Shrivastava, A.; Gupta, A. A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
102. Höfer, T.; Shamsafar, F.; Benbarka, N.; Zell, A. Object detection and Autoencoder-based 6D pose estimation for highly cluttered Bin Picking. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 704–708.
103. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. *arXiv* **2018**, arXiv:1803.01534.
104. Ghiasi, G.; Lin, T.-Y.; Le, Q.V. NAS-FPN: Learning scalable feature pyramid architecture for object detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7029–7038. Available online: https://ieeexplore.ieee.org/document/8954436/ (accessed on 16 May 2022).
105. Xie, C.; Mousavian, A.; Xiang, Y.; Fox, D. Rice: Refining instance masks in cluttered environments with graph neural networks. In Proceedings of the 5th Conference on Robot Learning, New York, NY, USA, 11 January 2022; pp. 1655–1665.