applied sciences

MDPI

Review

6D Pose Estimation of Objects: Recent Technologies and Challenges

Zaixing He [1], Wuxi Feng [1], Xinyue Zhao [1,*] and Yongfeng Lv [2]

[1] School of Mechanical Engineering, the State Key Lab of Fluid Power & Mechatronic Systems, Zhejiang University, Hangzhou 310027, China; zaixinghe@zju.edu.cn (Z.H.); feng_wx@zju.edu.cn (W.F.)
[2] Zhejiang Institute of Mechanical and Electrical Engineering, Hangzhou 310053, China; luyongfeng@zime.edu.cn
* Correspondence: zhaoxinyue@zju.edu.cn

Abstract: 6D pose estimation is a common and important task in industry. Obtaining the 6D pose of objects is the basis for many other functions such as bin picking, autopilot, etc. Therefore, many corresponding studies have been made in order to improve the accuracy and enlarge the range of application of various approaches. After several years of development, the methods of 6D pose estimation have been enriched and improved. Although some predecessors have analyzed the methods and summarized them in detailed, there have been many new breakthroughs in recent years. To understand 6D pose estimation better, this paper will make a new and more detailed review of 6D pose estimation. We divided these methods into two approaches: Learning-based approaches and non-learning-based approaches, including 2D-information-based approach and 3D-information-based approach. Additionally, we introduce the challenges that exist in 6D pose estimation. Finally, we compare the performance of different methods qualitatively and discuss the future development trends of the 6D pose estimation.

Keywords: 6D pose estimation; learning-based approach; 2D-information-based approach; 3D-information-based approach; textureless and reflective objects; foreground occlusion; background clutter

Citation: He, Z.; Feng, W.; Zhao, X.; Lv, Y. 6D Pose Estimation of Objects: Recent Technologies and Challenges. *Appl. Sci.* **2021**, *11*, 228. https://doi.org/10.3390/app11010228

Received: 11 November 2020
Accepted: 24 December 2020
Published: 29 December 2020

1. Introduction

1.1. Overview

6D pose refers to the posture of an object, specifically on the basis of a translation vector and a rotation vector. 6D pose estimation is an important step in many industrial fields highly related to another challenge—problem tracking [1]—such as bin picking [2–6], autonomous driving [7–9], augmented reality [10–12], SLAM (Simultaneous Localization and Mapping) [13–15] and so on (Figure 1). There have been an increasing number of applications of pose estimation developed in recent years. Autonomous vehicles use the technology of 6D pose estimation to recognize roads and obstacles. In the factory, the robots use the technology of 6D pose to recognize and grab objects. In the field of augmented reality, 6D pose estimation is used to measure the pose of objects in the real environment and add the virtual objects onto them in a correct pose. Some previous approaches could only detect the object and ensure its position, as is the case for GPS (Global Positioning System) [16] and radar detection [17]. These methods cannot measure the 6D pose of objects accurately. In industrial developments, higher demands are made for new application scenarios. Therefore, 6D pose estimation has become a hot topic in industry in recent years. 6D pose estimation uses a number of kinds of information to solve problems. It obtains texture information, geometric information, and color information to measure the 6D pose of objects. Due to the development of hardware in recent years, depth information is also used frequently in 6D pose estimation. However, 6D pose estimation is faced with many challenges, such as background clutter and inadequate information. Many methods have been proposed to improve the performance and enlarge the range of applications of 6D

pose estimation, and many new methods have been proposed in recent years. Additionally, there are many challenges in the 6D pose estimation field. A deeper comprehension of these challenges will also help arrive at more practical methods. To understand these methods and challenges more deeply, more detailed classification and performance evaluation need to be carried out. This review summarizes some relevant studies published in recent years and divides these methods into three categories. At the same time, we also analyze the advantages and disadvantages of these categories and challenges in 6D pose estimation.

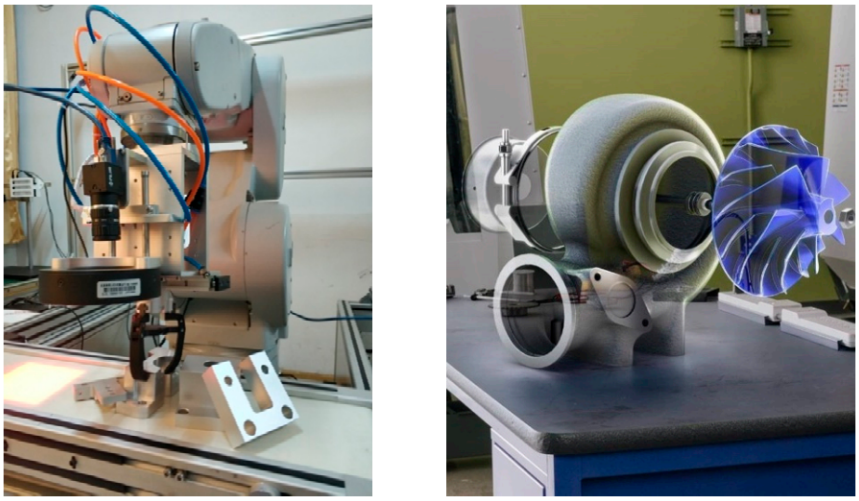

Figure 1. 6D pose estimation applied in bin picking and augmented reality.

1.2. Classification

In this paper, the 6D pose estimation approaches are divided into two categories: 1. Learning-based approach, and 2. Non-learning-based approach. The non-learning-based approach is divided into two categories: 1. 2D-information-based approach, and 2. 3D-information-based approach. The classification in this paper is mainly based on the core principle and the input information of the various methods: Learning-based approaches mainly use CNN, regression or some other methods based on deep learning to train a learning model with adequate training data and then obtain the 6D pose estimation result on the basis of these models. Approaches that do not use deep learning belong to the following two categories. 2D-information-based approaches mainly use the 2D information of the scene, such as RGB images. 3D-information-based approaches mainly use the 3D information of the scene, such as point clouds and RGB-D images. Both 2D-information-based approaches and 3D-information-based approaches convert the 6D pose estimation into image retrieval. The two types of approach both calculate the key points or key features and match the input image with the most similar image in the dataset according to the key points or key features. However, they also have some obvious differences, which will be covered in following sections.

The main purpose of learning-based approaches is to train a proper model to measure the 6D pose of an unknown situation according to the training data. Many kinds of model can be used to measure the 6D pose, such as regression models and CNN models. There are many classification methods for learning-based approaches, and those widely accepted among them are introduced in this paper. Keypoints-based approaches adopt a two-step category to measure the 6D pose, which is easier to implement than other approaches. Meanwhile, the aim of holistic approaches is to train an end-to-end network to measure the 6D pose of an object. It sees the image as a whole and tries to predict the location and orientation of the object in a single step and discretize the 6D space, converting the pose estimation task into a classification task. However, holistic approaches are more complex and time-consuming than keypoints-based approaches.

As for 2D-information-based approaches, the main purpose is to find the correlation between the input image and one of the template images through the 2D information contained in the image. Actually, 2D-information-based approaches converts the pose estimation into an image-matching problem. The matching results have a great influence on the results of the pose estimation. 2D-information-based approaches can be divided into real-image-based approaches and CAD-image-based approaches according to the kind of template used. When the approach uses real images as a template, it belongs to the real-image-based approaches. If the approach uses images generated by CAD model, it can be regarded as a CAD-image-based approach. In general, CAD-image-based approaches are more accurate than real-image-based approaches because the images generated by CAD models contain little noise. However, sometimes CAD models cannot be easily obtained, so real images are used as a template in such situations.

3D-information-based approaches also focus on the matching between the input and the dataset; however, they use the 3D information of the object, such as point clouds and RGB-D images. 3D-information-based approaches can be divided into two categories. The main idea of matching-based approaches is to match the input image and the template directly and to take the 6D pose of the matched template as the pose estimation result of the input image. Local descriptor-based approaches measure the 6D pose using the correspondence between the descriptor of input images and templates. Matching-based approaches require large storage to save enough templates to ensure the accuracy of pose estimation, and the more templates it has, the more accurate the pose estimation result will be.

1.3. Challenge

Although great progress has been made in the research of 6D pose estimation in recent years, there are still some challenging problems to be solved in practical application. When the background is messy, the viewpoint and illumination change greatly, or the scene has less texture, the accuracy and the robustness of 6D pose estimation needs to be improved. Learning-based approaches are relatively robust in these conditions. However, 2D-information-based approaches and 3D-information-based approaches do not perform well. However, for learning-based approaches, adequate training data and training time are required, and challenges related to the requirement of offline training and practicality are also problems to be solved.

1.4. Structure Layout

As shown in Figure 2, the rest of this paper can be divided into four parts: First, we introduce some research contributions in recent years in the three approaches and describe the advantages and challenges they are faced with in detail. We introduce learning-based approaches in the second section, 2D-information-based approaches and 3D-information-based approaches in the third section. Next, we focus on the challenges of 6D pose estimation. Then we compare the three approaches qualitatively. Finally, we give some views on the future development of the 6D pose estimation.

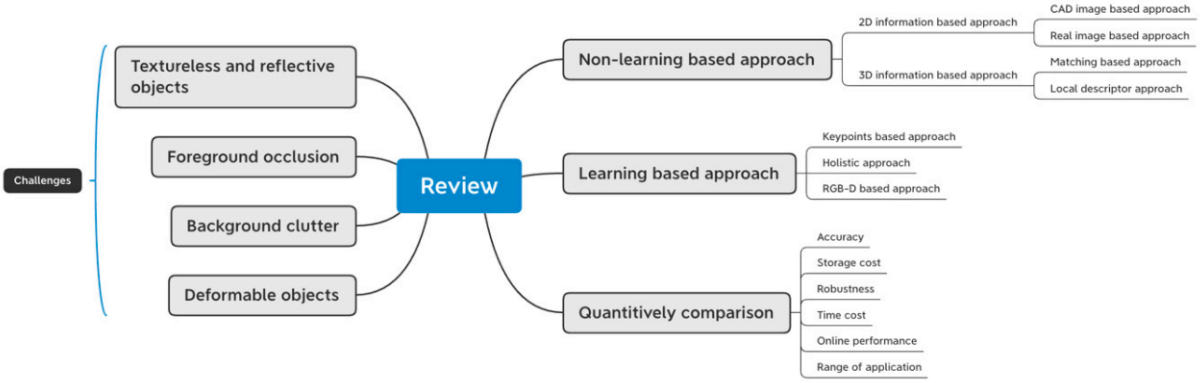

Figure 2. Structure layout of the review.

2. Learning-Based Approaches

In recent years, machine learning-based algorithms have become a hot topic due to new concepts' emergence, such as deep learning and neural networks. Many scholars have applied machine learning-based methods to 6D pose estimation and achieved good results. Posenet [18] is a monocular 6D relocalization system that trains a convolutional neural network (CNN) to regress 6D poses. The network transformed the problem of 6D pose estimation into a regression problem for which the input is a single RGB image and the output is the camera's 6D pose by using an end-to-end approach. To address the limitation of the lack of training data, a method was proposed that could generate large regression datasets of camera pose automatedly based on structure from motion. Crivellaro [19] also trained a CNN (Figure 3) to predict the 6D poses of objects that are partially visible. The key idea of the method was to predict the 3D-2D projections of feature points on each part of the object. When the test image is partially visible, the method could measure the 6D pose according to the feature points of visible part. When the test image is fully visible, the method is able to achieve more accurate results by combining all the feature points of the part. Particle Swarm Optimization (PSO)-based methods [20,21] demonstrate superior performance compared to Iterative Closest Point (ICP) algorithms. Hoang et al. [22] combined CNN with Simultaneous Localization and Mapping (SLAM), which improved the method [23] by adding a 6D object pose detector and measuring the 6D pose from different viewpoints to achieve a robust object detector system. However, there is a special problem for 6D pose estimation methods based on deep learning.

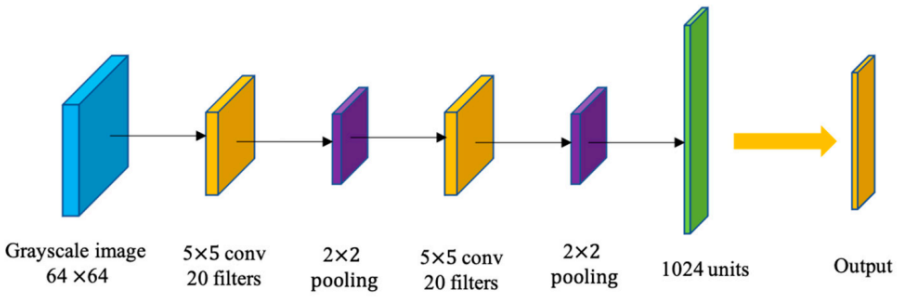

Figure 3. Architecture of the CNN predicting the projections of the control points.

Symmetrical objects' 6D poses (Figure 4) are difficult to measure correctly using normal deep learning methods [24], because the 6D pose of an object does not change from a fixed point of view when it is rotated 180 degrees. However, their actual ground truth is obviously different. For instance, if a network is trained to predict the pose using the squared loss between the ground truth poses and the predicted poses, it will converge to a model predicting the average of the possible poses for an input image, which is of course meaningless. Pitteri et al. [25] proposed an efficient method combined with Faster-RCNN, which relies on the normalization of the pose rotation. Manhardt et al. [26] proposed a method that was able to detect the rotation ambiguities and characterization of the uncertainty in the problem without further annotation or supervision. Zhang et al. [27] used the rigid transformation-invariant point-wise features of the point clouds as input features and used a hierarchical neural network that combined global point cloud information with the local patches to predict the key point coordinates.

2.1. Keypoint-Based Approaches

Keypoint-based approaches establish 2D-3D correspondences between images and then measure the pose according to these correspondences [28–30]. The procedures for keypoint-based approaches can be divided into two steps: 1. extract the 2D feature points in the input image; and 2. regress the 6D pose results using a PnP algorithm. BB8 [31] leveraged CNN to predict the 2D projections of eight vertices of the 3D bounding box of the object (Figure 5). To solve the challenges presented by textureless symmetrical objects,

BB8 restricted the range of the rotation angle of the training data and used a classifier to predict the rotation angle during the estimation step. However, when the object is partially invisible, BB8 may not obtain the correct 3D bounding box, which would have an adverse influence on PnP. To solve this problem, Hu [32] et al. proposed a method that segmented the image into several patches and made them predict both to which object they belonged and where the 2D projections were. Then, all the patches belonging to the same objects would be combined to measure the 6D pose based on PnP. Because each patch of the object is used to measure their respective local pose, this method was able to perform well when faced with occlusion. PVNet [33] predicted the direction of each pixel to each keypoint; thus, the spatial probability distribution of 2D key points can be obtained in a manner like RANSAC. According to the distribution, uncertainty-driven PnP could be used to measure the 6D pose. Predicting the direction of pixels and keypoints makes the local features more prominent. Even if one feature point is invisible, it can be positioned by means of another visible part. Jeon et al. [34] proposed a method involving learning orientation-induced primitives, rather than employing 3D bounding boxes, and calculated the rotation and translation vector in different modules. The methods mentioned above are all two-step-based; however, Hu [35] revealed the weakness of this kind of method. First, it is not an end-to-end system. Additionally, the loss function of the neural network cannot represent the accuracy of 6D pose estimation. Therefore, Hu et al. proposed a single-stage 6D pose estimation method that could directly regress the 6D pose on the basis of groups of 3D-to-2D correspondences associated with each 3D object keypoint.

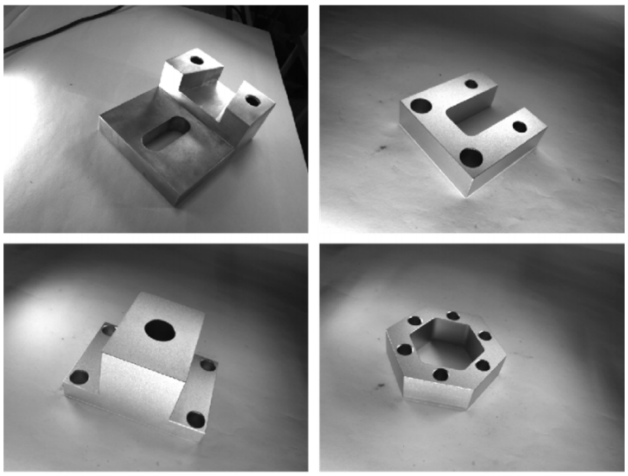

Figure 4. Symmetrical objects.

Figure 5. The red bounding boxes for the pose estimation results using BB8.

2.2. Holistic Approaches

Unlike keypoint-based approaches, holistic approaches are an end-to-end architecture that can be faster than keypoint-based approaches. Kendall et al. [36] proposed PoseNet, which firstly applies the CNN architecture to 6D pose estimation, and found that it was able to adapt well to the environment. Liu et al. [37] proposed SSD, which was the first method to associate bounding box priors with the feature maps of different spatial resolutions in the network that was able to detect objects in images using a single deep neural network. This method improved accuracy and retained a low time cost. Kehl et al. [38] proposed SSD-6D, which extended the SSD method to 6D pose estimation and allowed for easy training and handling of symmetries. Do et al. [39] proposed the deep-6DPose network, the detection and segmentation in which leverage the Region Proposal Network (RPN) [40] based on Mask R-CNN [41]. In pose estimation, it decouples the parameters into translation and rotation so that the rotation can be regressed via a Lie algebra representation. However, because the network uses the ROIs from RPN as inputs and predicts the 6D pose of the object in ROIs, the network was not able to work well when measuring the 6D pose of small or symmetrical objects. To overcome this problem, Xiang et al. [42] proposed a new network PoseCNN. This method calculated the translation vector by ensuring the center of the objects in the image and estimating the distance between the center and the camera. Then it calculated the rotation matrix by regressing to a quaternion representation. Additionally, it especially employed a novel loss function for symmetric objects. The method was able to handle occlusion and symmetric objects in cluttered scenes with RGB or RGB-D images as input.

2.3. RGB-D-Based Appraoches

Compared with the learning-based approaches mentioned above, which only use the RGB information, RGB-D-based learning approaches combine color information and depth information to measure the pose of objects, and are able to solve the problem of insufficient information in approaches that only use color information. Additionally, due to the emergence of lots of RGB-D datasets, an increasing number of studies on RGB-D-based learning approaches are being performed. In [43], Wang et al. proposed a novel method named DenseFusion that provided a two-stage method for measuring the 6D pose. In the first stage, DenseFusion uses a heterogeneous network to deal with the RGB data and point cloud data, and to save their original structure. In the second stage, a full convolutional network is used to map each pixel in RGB crop to colored feature space and uses a network based on PointNet to map each point in the point cloud to geometrical feature space. Then it merges the feature points in the colored feature space and the geometrical feature space and outputs a 6D pose estimation result. In addition, it finally refines the result by loop learning. In [44], Chen et al. proposed a 6D pose estimation framework named G2L-Net. Firstly, it extracts the coarse point cloud from RGB-D images. Then, the point cloud is added into the network to achieve 3D segmentation and object translation predictions. Finally, the fine point cloud is transferred into a local canonical coordinate to estimate initial object rotation. In [45], PVN3D was proposed, in which the method based on 2D key points was extended to 3D key points, making full use of geometric constraint information of rigid objects and improving the accuracy of the 6D estimation significantly. In [46], a method named CosyPose was proposed, which used multiple cameras to estimate 6D pose. Firstly, it estimates the 6D pose of objects in each image and then matches the individual 6D object pose hypotheses across different input images in order to jointly estimate the camera viewpoints and 6D poses of all objects in a single consistent scene.

2.4. Conclusions

In this section, we introduced learning-based approaches and classified the appraches into three categories: keypoints-based approaches, holistic approaches, and RGB-D-based approaches. Keypoints-based approaches are two-step approaches that extract 2D-3D point pairs and then use PnP to calculate the 6D pose of the object. Holistic approaches use an end-

to-end structure to measure the 6D pose, which is faster and more robust than keypoints-based approached. RGB-D-based learning approaches combine color information and depth information to achieve a more accurate and robust 6D pose result.

The comprehensive performance of learning-based approaches is better, as can be seen in the Amazon Robotic Challenge. The first Amazon Robotic Challenge was held in 2015 in the USA [47]. The competition presents a challenging problem that integrates many fields, including pose estimation. Many teams use learning-based approach and obtain robust results. In [48], the researchers presented a method for multi-class segmentation from RGB-D data. Objects were segmented by computing the possibility of each pixel belonging to each object by using a network. However, learning-based approaches still have some weaknesses. They require plenty of storage for storign training data and enough time to train the model. In conclusion, there are three challenges facing learning-based approaches. Firstly, the approaches require plenty of training data to train the model. Secondly, they require plenty of time to train the model before online use. For some complex objects that need much training data, this may take serval hours or even several days to train the model, thus restricting the possible applications of such approaches. Finally, these approaches cannot perform well when measuring poses that do not exist in the dataset.

3. Non-Learning-Based Approaches

3.1. 2D-Information-Based Approaches

Compared with 3D information, 2D information can be obtained more easily by simpler devices such as Charge-Coupled Device (CCD) cameras, Complementary Metal Oxide Semiconductor (CMOS) cameras, or even color cameras that can obtain the color information of objects. There is much 2D information that can be used to measure the 6D pose of objects, for instance, geometric information, texture information, color information, and so on. Scale-invariant feature transform (SIFT) features [49] and speeded up robust features (SURF) [50] are the early and classical features for pose estimation based on the texture of objects. SIFT and SURF features are both reliable and can achieve precise matching; however, they rely on the texture information of objects. Therefore, they cannot be used to solve the pose estimation of textureless objects. Zhang et al. [51] used geometric information, which does not rely on texture, to solve the pose estimation problem of textureless objects. E. Miyake et al. [52] combined the color information in the 6D pose estimation to improve accuracy and robustness. 2D information is extracted for the matching of the template. According to the dimensions of the template, approaches based on 2D information can be divided into two categories: 1. CAD image-based methods; and 2. real image-based methods.

3.1.1. CAD Image-Based Approaches

In contrast to real image-based approaches, CAD image-based approaches are more suitable for industrial products. The virtual images (used as templates) generated by CAD models are more accurate than real images, as the render process is not affected by illumination or blur. Moreover, CAD models can be obtained in industrial applications. Therefore, many methods based on 3D CAD model have been proposed. A hierarchical model [53] has been proposed (Figure 6), combining a coarse-to-fine search with similarity scores [54] calculated between a template and a real image or between templates. In [55], a perspective cumulated orientation feature (PCOF) was proposed based on the orientation histograms extracted from randomly generated 2D projection images using CAD models. Muñoz et al. [56] proposed the use of edge correspondences to estimate poses, with a similarity measure encoded using a pre-computed linear regression matrix. The Fine pose Parts-based Model (FPM) [57] was introduced to localize objects in an image, and to estimate their fine pose using the given CAD models. Pei et al. [58] proposed a robust method that only used one pair of vanishing points and one structural line to estimate the relative pose between image pairs. Peng et al. [59] proposed a method

which used several cameras to detect geometrical features and then combined their results to obtain the final result.

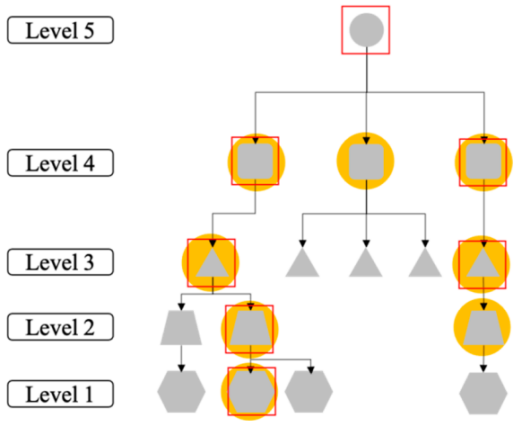

Figure 6. Object recognition using the hierarchy of views.

In [60], the Epipolar Geometry method and direct estimation method were used to estimate the 3D parameters, which were then used to construct the transformation matrix. In [61], 6D pose estimation was used in augmented reality. A local moving edge tracker was used to provide real-time tracking of points normal to the object contours. In addition, an M-estimator was used, integrated with a robust control law, to obtain good robustness. Straight lines and curves were both used in this method to complement the virtual visual servoing. In [62], 6D pose estimation was used for end effector tracking in a scanning electron microscope to aid in enabling more precise automated manipulations and measurements. Visible line features were also used to update the pose results. Kemal et al. [6] proposed a CAD model-based tracking method for visually guided microassembly. They used multiple cameras to track objects and find feature points along the edges of objects. Then, the 3D-3D for each feature point was built, and the 6D pose was calculated.

Whether real images or CAD models are being used as the template, the performance of the matching step determines the accuracy of pose estimation. Improving the efficiency and accuracy of template matching has become an important problem to ensure the results of pose estimation.

Additionally, the number of templates influences the accuracy of the pose estimation. The more templates there are, the more accurate the pose estimation will be. However, a large number of templates requires lots of storage and search time. In [63], a part-based efficient template matching method was proposed which was able to accelerate the matching step and improve the accuracy of pose estimation. Each of the templates leveraged a different forest and independently encoded similarity function.

3.1.2. Real Image-Based Approaches

Although it is possible to achieve precise results when using 3D CAD models, sometimes accurate 3D CAD models cannot be obtained. Therefore, real images are used as the template under these conditions. The histogram of gradients (HoG) [64] is a popular method that is computed on a dense grid with uniform intervals for better performance. Guo et al. [65] used multi-cooperative logos to measure 6D pose. Hinterstoisser [66] proposed a method including a novel image representation for template matching designed to be robust to small image transformations. It used the gradient orientation of the edges of objects to create templates. The method was able to be extended if 3D information was available. However, because obtaining adequate real image templates is time-consuming and challenging, and generating the images by CAD model is becoming easier, there is not much research on real image-based approaches.

3.1.3. Conclusions

In this part, we introduced 2D-information-based approaches and classified the approaches into two categories: CAD image-based approaches and real image-based approaches. The difference between them is which type of template is used. CAD image-based approaches require a CAD model of the object, but templates can be generated conveniently. Real image-based approaches use real images as templates, but if there is clutter in the real images, the information may be extracted incorrectly.

Compared with 3D-information-based approaches, 2D-information-based approaches are less robust, because these approaches use 2D information, which has less data than 3D information. Additionally, complex scenes have a bad influence on the performance of these approaches. The biggest weakness for 2D-information-based approaches is that they are not able to adapt to some special scenes, such as scenes with strong changes in illumination, large numbers of repeated structures, textureless scenes, and so on.

3.2. 3D-Information-Based Approaches

Although there are many kinds of 2D information, it is difficult to use 2D information to measure the 6D pose under some special conditions, or the method requires complex algorithms to obtain precise results. With the development of hardware, more and more devices that can record 3D scene information are appearing, such as depth cameras [67] and 3D scanners [68]. Compared with 2D information, 3D information preserves the original appearance of the object, which is more useful for measuring 6D pose [44,69]. Combined with 3D information, the method is more robust and can obtain more accurate results. Mainstream methods can be broadly divided into the following two categories.

3.2.1. Matching-Based Approaches

The aim of matching-based approaches is to search for the most similar template in the dataset and return the 6D pose of the template. Because the 3D original data is always too large, processing these data directly can be computationally expensive. Therefore, many preprocessing methods have been proposed to reduce the complexity of the task. Zhang et al. [70] proposed two methods for solving this problem. One was to use a 2D/2.5D object detector for scene point clouds. YOLO was used to segment the scene point cloud with 2D bounding boxes due to their lower time consumption. The other preprocessing method was to extract the keypoints in the template point clouds. Points with more information, such as the points on edges, were preserved and points on surfaces were removed in order to compress the point cloud. Konishi et al. [71] combined PCOF-MOD (multimodal PCOF), balanced pose tree (BPT), and optimum memory rearrangement into 6D pose estimation to optimize data storage structure and lookup speed. To improve the accuracy of matching, Park [72] et al. proposed a novel multi-task template matching (MTTM) framework that finds the nearest template of a target object from an image while predicting segmentation masks and a pose transformation between the template and a detected object in the scene using the same feature map of the object region. In [73], research was carried out on tracking and control for micro-electro-mechanical system (MEMS) microassembly. The correlation between the real-time 3D vision tracking method and the control law based on 3D vision was demonstrated, and a pose-based visual servoing approach was used to to enable a precise regulation toward zero of 3D error.

3.2.2. Local Descriptor Approaches

Approaches based on the local descriptor define and calculate a global descriptor on the model offline. The global descriptor should be invariant with respect to rotation and translation. Then, the local descriptor is calculated and matched with the global descriptor online. The iterative closest point (ICP) [74,75] algorithm is a classical one that is able to calculate the pose relation between two coordinates according to two sets of point clouds. 6D pose can be measured by the correspondence between them or as the result of voting. Guo et al. [76] proposed a global method named Super Key 4-Points Congruent Sets (SK-4PCS),

combined with invariant local features of 3D shapes, thus reducing the amount of processed data. Akizuki et al. [77] proposed a method aimed at everyday tools that does not rely on the 3D model of the measured object. It assumed that the same kinds of tools possess the same part-affordance. Therefore, using the 3D models for similar kinds of object, the method was able to measure the pose of objects belonging to this kind on the basis of the spatial relationships of part-affordance. Yu et al. [78] used the improved Oriented Fast and Rotated Brief (ORB) [79] feature and rBRIEF descriptor (a descriptor developed based on binary robust independent elementary features (BRIEF) [80]) to obtain a coarse match, and then culled mismatches to retain more correct matches. They also proposed a hybrid reprojection errors optimization model (HREOM) to improve the accuracy of the result by minimizing 3D-3D and 3D-2D reprojection errors. To measure the 6D pose of large-scale objects that are partially visible, David et al. [81] proposed a method based on semi-global descriptors. They used semi-global descriptors for scene segments and model views in combination with up-sampling and segment label merging techniques and obtained more reliable results than with other descriptors.

3.2.3. Conclusions

In this part, we introduced 3D-information-based approaches and classified the approaches into two categories: matching-based approaches and local descriptor approaches. Matching-based approaches are more computationally expensive; however, local descriptor approaches require some preprocessing oofline.

The biggest weakness for approaches based on 3D information is that they do not perform well when the object is reflective, which is due to how the approaches work. When measuring the 6D pose of reflective objects, these approaches are not able to obtain the depth information (or accurate point clouds) of the object. Additionally, another weakness is that the efficiency of this approach is relatively low, because point clouds and depth images include a lot of data, causing computational burden.

4. Comparison

In this section, we will compare these approaches in detail according to their performance in Table 1. Specifically, the accuracy, the storage cost, the robustness, the time cost, online performance and range of application will be discussed. The mentioned indicators are compared qualitatively in Table 1 according to the information in these papers, where A represents the best performance and C represents the worst performance.

Table 1. Comparison of three kinds of approaches.

		Accuracy	Storage Cost	Robustness	Time Cost	Online Performance	Range of Application
Learning-based approaches	Keypoint-based approaches	B	B	B	C	C	B
	Holistic approaches	C	B	B	B	B	B
	RGB-D-based approaches	A	C	A	C	C	C
Non-learning-based approaches	2D-information-based approaches	B	A	C	A	A	A
	3D-information-based approaches	A	B	B	B	B	C

Accuracy: The accuracy of holistic approaches is worse than that of 2D-information-based approaches and 3D-information-based approaches. This is because holistic approaches transform the problem of 6D pose estimation into a problem of classification. The accuracy of classification decides the accuracy of 6D pose estimation. However, RGB-D-based approaches use more comprehensive information. The depth information is able to provide the overall morphology of a rigid body, while the color information is able to describe the position of keypoints. Therefore, RGB-D-based approaches that combine depth information and color information are able to improve the accuracy. 2D-information-based approaches and 3D-information-based approaches convert the 6D pose estimation problem into a template matching or coordinate transformation problem. Both of them are able to measure the

6D pose more specifically. 3D-information-based approaches use more information than 2D-information-based approaches, so they perform better on accuracy.

Robustness: Robustness in this paper mainly refers to the anti-interference performance of approaches to noise and environmental changes. Plenty of data is used to train a network in learning-based approaches. The model adequately considers the information and situation in the input scene, so learning-based approaches have better robustness than the other two approaches. Additionally, using only color information may lead to missing keypoints, while depth information provides some global pose information and is complementary to color information. After a long period of training, learning models are able to distinguish object features and environmental noise. Although the other two approaches are also able to distinguish information from the environment efficiently, some wrong information may be taken into consideration under some complex situations where background clutter and foreground occlusion are present.

Storage cost: Learning-based approaches need the most storage because the training process of the learning models needs plenty of data. The other two approaches need to store templates. However, the number of templates is lower than the training data required for learning-based approaches. In particular, for coarse-to-fine methods, which just need to match a basically similar template with the input image, much fewer templates are needed. In terms of a comparison between 2D-information-based approaches and 3D-information-based approaches, the former needs less storage, because 2D information is smaller than 3D information.

Time cost and real-time performance: Because the time an approach takes determines the real-time performance of the approach, the two indicators are discussed together in this part. Learning-based approaches can be divided into two steps: offline training and online measuring. The offline step is time-consuming because it needs to train a model using plenty of data, which means that repeated calculations are necessary, although a GPU could accelerate the training process. However, in the online step, the 6D pose can be measured directly by the trained model, which costs very little time. 2D-information-based approaches and 3D-information-based approaches an also be divided into an offline step and an online step. In the offline step, many templates are generated from different angles (however, there are also some methods that do not need a lot of templates). Meanwhile, in the online step, the proper template needs to be retrieved from template dataset, which is a little time-consuming. Therefore, in the online step, the real-time performance of learning-based approaches is the best. 2D-information-based approaches cost less time than 3D-information-based approaches because the retrieval of 3D information is more time-consuming.

Range of application: Due to their principle, 3D-information-based approaches are not able to handle the problem of 6D pose estimation of reflective objects. However, such objects are common in industry, such as metal parts. Learning-based approaches need plenty of time to train a network, which may not satisfy the requirements of real-time performance.

In this section, we compared the approaches with respect to six different aspects. In general, learning-based approaches have the best robustness, but these approaches need lots of storage to save training data and plenty of time to train the model. 3D-information-based approaches achieve the most accurate results, but they cannot be used on reflective objects. The range of application of 2D-based approaches is the widest; however, 2D information is easily affected by the environment. Therefore, the three approaches all have their advantages. In different situations, different approaches are needed.

5. Challenges

5.1. Textureless and Reflective Objects

As a special case in pose estimation, the pose estimation of textureless objects is challenging. Due to the lack of reliable texture information on the surface of such objects, it is difficult to extract feature points on them. Therefore, many methods based on the surface texture information of objects cannot effectively measure the 6D pose of such objects.

However, textureless objects are common in industry. Therefore, the pose estimation of textureless objects is very important.

Zhang et al. [82] transformed sliding windows to scale-invariant RGB-D patches and applied a hash voting-based hypothesis generation scheme to compute a rough 6D pose hypothesis and then employed particle swarm optimization to improve the result, which achieved high precision and good performance on three datasets. Pan et al. [83] coarsely segmented the object in the point cloud and precisely measured the pose results in the gray image using a view-based matching method.

The methods mentioned above used a depth camera or 3D scanning device to obtain the depth image or the point cloud of objects. However, for objects with reflective surfaces, such as metal parts, depth information is hard to acquire. Therefore, some other kinds of information are used in pose estimation. Geometrical information is a reliable kind of information for textureless objects. Based on the above perspective, He et al. [84] proposed a new method (Figure 7) making full use of the geometrical information of objects. It used geometric features, such as straight lines, to generate descriptors of the objects, and proposed the GIIBOLD algorithm for matching the input image and the template image. Furthermore, accurate 2D-3D point pairs were acquired on the basis of the matched geometric features. Finally, the 6D pose was measured using the PnP-RANSAC algorithm. This method leveraged simple geometrical information, achieving fast matching and accurate measurement without the requirement for plenty of storage and time. Zhang et al. [85] first detected the object in the RGB image using a 2D bounding box and then measured the pose result in the edge image. Pan et al. [86] used multiple appearance features including color, size and aspect ratio to distinguish objects from environmental clutter and measured the 6D pose. Zhang et al. [51] proposed a novel method for measuring textureless objects on the basis of RGB images. It followed a coarse-to-fine procedure, using only the shape and contour information of the input image. Several template images with poses similar to the input image were selected to match with the input. Then the contour and shape information, specifically the ORB features, were used to establish 2D-3D correspondence and finally to calculate the 6D pose. On the basis of the studies above, without reliable texture information, there are many kinds of information can be leveraged, such as contour, color, shape, and so on. Accurate results can be obtained through the proper use of this information.

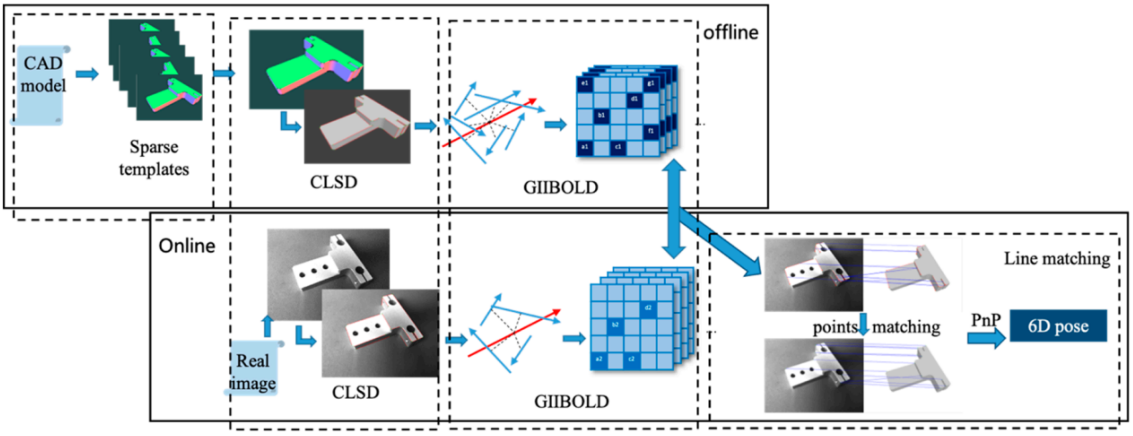

Figure 7. The overall workflow chart.

5.2. Foreground Occlusion

In complex industrial scenes, the condition of object occlusion appears frequently. Because the target is obscured by other objects, the recognition of the target's features is disturbed. In addition, due to part of the object information being missing, it is difficult to calculate accurate pose results. As a common condition, object occlusion has been studied by many researchers.

Crivellaro et al. [19] proposed a method for representing the pose of each part on the basis of the 2D reprojections of a small set of 3D control points. It was able to predict the 6D pose of the object by predicting the pose of the visible part through the reprojections of 2D control points. Even if the object was only partially visible, the method was able to calculate the 6D pose of the object. If the object had several visible parts, the method was able to combine all the information and obtain more accurate results. Distinct from the method above, Dong et al. [87] used 3D information as input and chose an end-to-end strategy. They proposed a novel network named Point-wise Pose Regression Network (PPR-Net). For each of the points in the point cloud, the network regressed a 6D pose of the object instance that the point belonged to. In the pose space, a clustering method was adopted in order to segment multiple instances from the clutter point cloud, and an instance's pose can be computed by averaging each subsidiary point's pose prediction. Essentially, the method used the information of visible parts to predict the pose of the object. The more parts of an object that could be seen, the more reliable the results obtained. Chen et al. [88] proposed a network which took the point cloud as input and regressed the point-wise unit vectors pointing to the 3D keypoints. Then the vectors were used to generate keypoint hypotheses from which 6D object pose hypotheses were computed. Tekin et al. [28] predicted the 2D image locations of the projected vertices of the object's 3D bounding box and used a PnP algorithm to estimate the object's 6D pose. Taking RGB-D images as input, Zhang et al. [89] combined holistic and local patches to measure the 6D object pose and obtained high precision and good performance under conditions of foreground occlusion and background clutter.

In conclusion, the principle of the method for dealing with object occlusion is to use the information of any visible parts to predict the 6D pose of the whole object. Generally, the number of methods to achieve this based on 3D information is greater than the number of those based on 2D information. In addition, among the methods based on 2D information, it is more appropriate to use outer information such as bounding boxes and contours. Because inner information, such as the texture, is occluded by other objects, the extraction of the information is affected to a great extent.

5.3. Background Clutter

Background clutter is also a challenge in 6D pose estimation. Because the target is surrounded by much useless information, it is difficult to measure the 6D pose directly. However, due to the complexity of practical scenarios, there are many conditions under which it is necessary to measure the 6D pose of objects in clutter.

He et al. [41] proposed Mask R-CNN, which efficiently detected objects in an image while simultaneously generating a high-quality segmentation mask for each instance. It extended Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. Mitash et al. proposed a method for measuring the 6D pose of objects in clutter. A global optimization process was employed to improve candidate poses by taking into account scene-level physical interactions between objects. Then, the combinations of candidate object poses were searched using a Monte Carlo Tree Search (MCTS) process that used the similarity between the observed depth image of the scene and the rendering of the scene given the hypothesized pose as a score, guiding the search procedure. Li et al. [90] proposed a two-step method for measuring the 6D pose. The first step was few-shot instance segmentation to segment the known objects from RGB images. Chen et al. [91] proposed a method based on point clouds that detected objects in an end-to-end manner. They introduced a point cloud-based 6D target object detection method that used segmented object point cloud patches to predict object 6D poses and identity. It used a point cloud segmentation procedure that was easier to visualize and tune in order to overcome the problem caused by background clutter.

5.4. Deformable Objects

The 6D pose estimation of deformable objects is a huge challenge in the field, because the posture of the objects is unpredictable and there are many ways for the objects to

deform. Thus, lots of conditions need to be taken into consideration, resulting in pressure on the algorithm and calculation.

Li et al. [92] proposed a novel method that was able to classify and estimate the categories and poses of deformable objects. They simulated the deformable objects and obtained the depth images from different viewpoints in different postures as a dataset. By extracting features and using deep learning, they set up a codebook for the object, which could be used in training process. This method uses learning-based approaches and sets up a two-layer framework to solve the problem of 6D pose estimation. Lots of storage space is needed, but the method would be robust in such complex situations. In [93], a predictive and model-driven approach was proposed to solve this problem. A dataset was built up by using the picked-up garments in multiple poses under gravity. The dataset was orginized in an efficient way, increasing the speed of the searching process. The proposed method constructed a fully featured 3D model of the garment in real time and used volumetric features to obtain the most similar model in the database in order to predict the object category and pose. Accurate model simulation could also be used to optimize the trajectories of the manipulation of deformable objects. Caporali et al. [94] proposed a four-step method to solve this problem of grasping clothes by using a point cloud. Firstly, the instance segmentation was performed, and then a wrinkledness measure was implemented to robustly detect the graspable regions of the cloth. Next, the identification of each individual wrinkle was accomplished by fitting a piecewise curve, and finally a pose for each detected wrinkle was estimated.

5.5. Conclusions

In this section, four challenges were discussed. For reflective objects, 3D-information-based approaches cannot accurately measure their pose. For textureless objects, only the geometrical features can be extracted for calculating the 6D pose of objects. Faced with foreground occlusion, many methods use the visible part to describe the invisible part of objects and measure their 6D pose. For background clutter, instance segmentation is used to separate the object, and then the 6D pose is measured. For deformable objects, images of the objects are captured in different poses in different views and a dataset is set up. The best matching is found in the dataset when measuring the pose of objects.

6. Conclusions

This paper divided solutions of 6D pose estimation into three kinds of approaches and made some detailed introductions to their advantages and disadvantages. Then, this paper focused on the challenges in 6D pose estimations, introducing the difficulties of these problems and summarizing some feasible solutions. Finally, some approaches were qualitatively compared with respect to several different indicators. Learning-based approaches achieved the best robustness. However, they are time-consuming and require lots of storage. 2D-information-based approaches are easy to implement and can be applied online. 3D-information-based approaches can achieve higher accuracy than 2D-information-based approaches, but they require more information (depth information) to be collected and dealt with, and they cannot measure the 6D pose of reflective objects. Generally, the methods presented have already satisfied the requirements of industrial application for the 6D pose detection of general objects. However, these methods cannot maintain their excellent performance under some challenging conditions. In future research, learning-based approaches should be further developed. On one hand, their robust performance should be retained. On the other hand, the offline training time should be decreased. These approaches can also be combined with 2D-information-based and 3D-information-based approaches to obtain more accurate results.

Author Contributions: Conceptualization, Z.H. and X.Z.; methodology, Z.H.; software, W.F.; validation, X.Z. and Y.L.; writing—original draft preparation, W.F.; writing—review and editing, Z.H. and X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (51775498, 51775497) and Zhejiang Province Public Welfare Technology Application Research Project (LGG19E050019).

Data Availability Statement: The data presented in this study is contained in the article itself.

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

References

1. Drummond, T.; Cipolla, R. Real-time visual tracking of complex structures. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 932–946. [CrossRef]
2. Peng, L.; Zhao, Y.; Qu, S.; Zhang, Y.; Weng, F. Real Time and Robust 6D Pose Estimation of RGBD Data for Robotic Bin Picking. In Proceedings of the 2019 Chinese Automation Congress (CAC), Hangzhou, China, 22–24 November 2019; pp. 5283–5288.
3. Yan, W.; Xu, Z.; Zhou, X.; Su, Q.; Li, S.; Wu, H. Fast Object Pose Estimation Using Adaptive Threshold for Bin-Picking. *IEEE Access* **2020**, *8*, 63055–63064. [CrossRef]
4. Xu, J.; Pu, S.; Zeng, G.; Zha, H. 3D pose estimation for bin-picking task using convex hull. In Proceedings of the 2012 IEEE International Conference on Mechatronics and Automation, Chengdu, China, 5–8 August 2012; pp. 1381–1385.
5. Buchholz, D. *Bin-Picking: New Approaches for a Classical Problem*; Springer: Berlin/Heidelberg, Germany, 2016.
6. Yang, J.; Zeng, G.; Wang, W.; Zuo, Y.; Yang, B.; Zhang, Y. Vehicle Pose Estimation Based on Edge Distance Using Lidar Point Clouds (Poster). In Proceedings of the 2019 22th International Conference on Information Fusion (FUSION), Ottawa, ON, Canada, 2–5 July 2019; pp. 1–6.
7. Gu, R.; Wang, G.; Hwang, J. Efficient Multi-person Hierarchical 3D Pose Estimation for Autonomous Driving. In Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 28–30 March 2019; pp. 163–168.
8. Kothari, N.; Gupta, M.; Vachhani, L.; Arya, H. Pose estimation for an autonomous vehicle using monocular vision. In Proceedings of the 2017 Indian Control Conference (ICC), Guwahati, India, 4–6 January 2017; pp. 424–431.
9. Zhang, S.; Song, C.; Radkowski, R. Setforge-Synthetic RGB-D Training Data Generation to Support CNN-Based Pose Estimation for Augmented Reality. In Proceedings of the 2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), Beijing, China, 10–18 October 2019; pp. 237–242.
10. Lu, Y.; Kourian, S.; Salvaggio, C.; Xu, C.; Lu, G. Single Image 3D Vehicle Pose Estimation for Augmented Reality. In Proceedings of the 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Ottawa, ON, Canada, 11–14 November 2019; pp. 1–5.
11. Hachiuma, R.; Saito, H. Recognition and pose estimation of primitive shapes from depth images for spatial augmented reality. In Proceedings of the 2016 IEEE 2nd Workshop on Everyday Virtual Reality (WEVR), Greenville, SC, USA, 20–20 March 2016; pp. 32–35.
12. Li, X.; Ling, H. Hybrid Camera Pose Estimation with Online Partitioning for SLAM. *IEEE Robot. Autom. Lett.* **2020**, *5*, 1453–1460. [CrossRef]
13. Ruan, X.; Wang, F.; Huang, J. Relative Pose Estimation of Visual SLAM Based on Convolutional Neural Networks. In Proceedings of the 2019 Chinese Control Conference (CCC), Guangzhou, China, 27–30 July 2019; pp. 8827–8832.
14. Xiao, Z.; Wang, X.; Wang, J.; Wu, Z. Monocular ORB SLAM based on initialization by marker pose estimation. In Proceedings of the 2017 IEEE International Conference on Information and Automation (ICIA), Macau, China, 18–20 July 2017; pp. 678–682.
15. Malyavej, V.; Torteeka, P.; Wongkharn, S.; Wiangtong, T. Pose estimation of unmanned ground vehicle based on dead-reckoning/GPS sensor fusion by unscented Kalman filter. In Proceedings of the 2009 6th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, Pattaya, Thailand, 6–9 May 2009; pp. 395–398.
16. Zhaoyang, N.; Jianyun, Z.; Zhidong, Z. Angle Estimation for Bi-static MIMO Radar Based on Tri-iterative Algorithm. In Proceedings of the 2010 First International Conference on Pervasive Computing, Signal Processing and Applications, Harbin, China, 17–19 September 2010; pp. 1264–1267.
17. Kendall, A.; Grimes, M.; Cipolla, R. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2938–2946.
18. Crivellaro, A.; Rad, M.; Verdie, Y.; Yi, K.M.; Fua, P.; Lepetit, V. A Novel Representation of Parts for Accurate 3D Object Detection and Tracking in Monocular Images. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4391–4399.
19. Zabulis, X.; Lourakis, M.; Koutlemanis, P. 3D Object Pose Refinement in Range Images. In *International conference on Computer Vision Systems*; Springer: Cham, Switzerland, 2015; pp. 263–274.
20. Zabulis, X.; Lourakis, M.I.; Koutlemanis, P. Correspondence-free pose estimation for 3D objects from noisy depth data. *Vis. Comput.* **2018**, *34*, 193–211. [CrossRef]
21. Hoang, D.; Stoyanov, T.; Lilienthal, A.J. Object-RPE: Dense 3D Reconstruction and Pose Estimation with Convolutional Neural Networks for Warehouse Robots. In Proceedings of the 2019 European Conference on Mobile Robots (ECMR), Prague, Czech Republic, 4–6 September 2019; pp. 1–6.

22. Hoang, D.-C.; Stoyanov, T.; Lilienthal, A. High-Quality Instance-Aware Semantic 3D Map Using RGB-D Camera. *arXiv* **2019**, arXiv:1903.10782.
23. Wen, Y.; Pan, H.; Yang, L.; Wang, W. Edge Enhanced Implicit Orientation Learning With Geometric Prior for 6D Pose Estimation. *IEEE Robot. Autom. Lett.* **2020**, *5*, 4931–4938. [CrossRef]
24. Pitteri, G.; Ramamonjisoa, M.; Ilic, S.; Lepetit, V. On Object Symmetries and 6D Pose Estimation from Images. In Proceedings of the 2019 International Conference on 3D Vision (3DV), Québec City, QC, Canada, 16–19 September 2019; pp. 614–622.
25. Manhardt, F.; Arroyo, D.M.; Rupprecht, C.; Busam, B.; Birdal, T.; Navab, N.; Tombari, F. Explaining the Ambiguity of Object Detection and 6D Pose From Visual Data. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6840–6849.
26. Zhang, W.; Chenkun, Q.I. Pose Estimation by Key Points Registration in Point Cloud. In Proceedings of the 2019 3rd International Symposium on Autonomous Systems (ISAS), Shanghai, China, 29–31 May 2019; pp. 65–68.
27. Tekin, B.; Sinha, S.N.; Fua, P. Real-Time Seamless Single Shot 6D Object Pose Prediction. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 292–301.
28. Li, Z.; Wang, G.; Ji, X. CDPN: Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-DoF Object Pose Estimation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 7677–7686.
29. Oberweger, M.; Rad, M.; Lepetit, V. Making Deep Heatmaps Robust to Partial Occlusions for 3D Object Pose Estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
30. Rad, M.; Lepetit, V. BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3848–3856.
31. Hu, Y.; Hugonot, J.; Fua, P.; Salzmann, M. Segmentation-Driven 6D Object Pose Estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3380–3389.
32. Peng, S.; Liu, Y.; Huang, Q.; Bao, H.; Zhou, X. PVNet: Pixel-wise Voting Network for 6DoF Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
33. Jeon, M.H.; Kim, A. PrimA6D: Rotational Primitive Reconstruction for Enhanced and Robust 6D Pose Estimation. *IEEE Robot. Autom. Lett.* **2020**, *5*, 4955–4962. [CrossRef]
34. Hu, Y.; Fua, P.; Wang, W.; Salzmann, M. Single-Stage 6D Object Pose Estimation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 2927–2936.
35. Kendall, A.; Grimes, M.; Cipolla, R. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. *Educ. Inf.* **2015**, *31*, 2938–2946.
36. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S. SSD: Single Shot MultiBox Detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016.
37. Kehl, W.; Manhardt, F.; Tombari, F.; Ilic, S.; Navab, N. SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
38. Do, T.-T.; Cai, M.; Pham, T.; Reid, I. Deep-6DPose: Recovering 6D Object Pose from a Single RGB Image. *arXiv* **2018**, arXiv:1802.10367.
39. 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.* **2015**, *39*, 1137–1149. [CrossRef]
40. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988. [CrossRef]
41. Xiang, Y.; Schmidt, T.; Narayanan, V.; Fox, D. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. *arXiv* **2017**, arXiv:1711.00199.
42. Wang, C.; Xu, D.; Zhu, Y.; Martín-Martín, R.; Lu, C.; Fei-Fei, L.; Savarese, S. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3338–3347.
43. Chen, W.; Jia, X.; Chang, H.J.; Duan, J.; Leonardis, A. G2L-Net: Global to Local Network for Real-Time 6D Pose Estimation With Embedding Vector Features. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4232–4241.
44. He, Y.; Sun, W.; Huang, H.; Liu, J.; Fan, H.; Sun, J. PVN3D: A Deep Point-Wise 3D Keypoints Voting Network for 6DoF Pose Estimation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11629–11638.
45. Labbé, Y.; Carpentier, J.; Aubry, M.; Sivic, J. CosyPose: Consistent Multi-view Multi-object 6D Pose Estimation. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 574–591.
46. Correll, N.; Bekris, K.E.; Berenson, D.; Brock, O.; Causo, A.; Hauser, K.; Okada, K.; Rodriguez, A.; Romano, J.M.; Wurman, P.R. Analysis and Observations From the First Amazon Picking Challenge. *IEEE Trans. Autom. Sci. Eng.* **2018**, *15*, 172–188. [CrossRef]
47. Jonschkowski, R.; Eppner, C.; Höfer, S.; Martín-Martín, R.; Brock, O. Probabilistic multi-class segmentation for the Amazon Picking Challenge. In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea, 9–14 October 2016; pp. 1–7.
48. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]

49. Bay, H.; Ess, A.; Tuytelaars, T.; Gool, L.V. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. [CrossRef]
50. Zhang, X.; Jiang, Z.; Zhang, H.; Wei, Q. Vision-Based Pose Estimation for Textureless Space Objects by Contour Points Matching. *IEEE Trans. Aerosp. Electron. Syst.* **2018**, *54*, 2342–2355. [CrossRef]
51. Miyake, E.; Takubo, T.; Ueno, A. 3D Pose Estimation for the Object with Knowing Color Symbol by Using Correspondence Grouping Algorithm. In Proceedings of the 2020 IEEE/SICE International Symposium on System Integration (SII), Honolulu, HI, USA, 12–15 January 2020; pp. 960–965.
52. Ulrich, M.; Wiedemann, C.; Steger, C. Combining Scale-Space and Similarity-Based Aspect Graphs for Fast 3D Object Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1902–1914. [CrossRef]
53. Steger, C. Occlusion, clutter, and illumination invariant object recognition. *Int. Arch. Photogramm. Remote Sens.* **2003**, *34*, 345–350.
54. Konishi, Y.; Hanzawa, Y.; Kawade, M.; Hashimoto, M. Fast 6D Pose Estimation from a Monocular Image Using Hierarchical Pose Trees. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; Volume 9905, pp. 398–413.
55. Muñoz, E.; Konishi, Y.; Murino, V.; Bue, A.D. Fast 6D pose estimation for texture-less objects from a single RGB image. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; pp. 5623–5630.
56. Khosla, A.; Torralba, A.; Lim, J. FPM: Fine Pose Parts-Based Model with 3D CAD Models. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014. [CrossRef]
57. Pei, L.; Liu, K.; Zou, D.; Li, T.; Wu, Q.; Zhu, Y.; Li, Y.; He, Z.; Chen, Y.; Sartori, D. IVPR: An Instant Visual Place Recognition Approach Based on Structural Lines in Manhattan World. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 4173–4187. [CrossRef]
58. Peng, J.; Xu, W.; Liang, B.; Wu, A. Virtual Stereovision Pose Measurement of Noncooperative Space Targets for a Dual-Arm Space Robot. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 76–88. [CrossRef]
59. Chaumette, F.; Hutchinson, S. Visual servo control. II. Advanced approaches [Tutorial]. *IEEE Robot. Autom. Mag.* **2007**, *14*, 109–118. [CrossRef]
60. Comport, A.I.; Marchand, E.; Pressigout, M.; Chaumette, F. Real-time markerless tracking for augmented reality: The virtual visual servoing framework. *IEEE Trans. Vis. Comput. Graph.* **2006**, *12*, 615–628. [CrossRef] [PubMed]
61. Kratochvil, B.E.; Dong, L.; Nelson, B.J. Real-time rigid-body visual tracking in a scanning electron microscope. In Proceedings of the 2007 7th IEEE Conference on Nanotechnology (IEEE NANO), Hong Kong, China, 2–5 August 2007; pp. 442–447. [CrossRef]
62. Yesin, K.B. A CAD model based tracking for visually guided three-dimensional microassembly. *Robotica* **2005**, *23*, 409–418. [CrossRef]
63. Muñoz, E.; Konishi, Y.; Beltran, C.; Murino, V.; Bue, A.D. Fast 6D pose from a single RGB image using Cascaded Forests Templates. In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea, 9–14 October 2016; pp. 4062–4069.
64. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 881, pp. 886–893.
65. Guo, J.; Wu, P.; Wang, W. A precision pose measurement technique based on multi-cooperative logo. *J. Phys. Conf. Ser.* **2020**, *1607*, 012047. [CrossRef]
66. Hinterstoisser, S.; Cagniart, C.; Ilic, S.; Sturm, P.; Navab, N.; Fua, P.; Lepetit, V. Gradient Response Maps for Real-Time Detection of Textureless Objects. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 876–888. [CrossRef] [PubMed]
67. Li, X.; Wang, H.; Yi, L.; Guibas, L.J.; Abbott, A.L.; Song, S. Category-Level Articulated Object Pose Estimation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3703–3712.
68. Wang, K.; Xie, J.; Zhang, G.; Liu, L.; Yang, J. Sequential 3D Human Pose and Shape Estimation From Point Clouds. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 7273–7282.
69. Zhang, Z.; Hu, L.; Deng, X.; Xia, S. Weakly Supervised Adversarial Learning for 3D Human Pose Estimation from Point Clouds. *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 1851–1859. [CrossRef] [PubMed]
70. Zhang, Y.; Zhang, C.; Rosenberger, M.; Notni, G. 6D Object Pose Estimation Algorithm Using Preprocessing of Segmentation and Keypoint Extraction. In Proceedings of the 2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Dubrovnik, Croatia, 25–28 May 2020; pp. 1–6.
71. Konishi, Y.; Hattori, K.; Hashimoto, M. Real-Time 6D Object Pose Estimation on CPU. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 3451–3458.
72. Park, K.; Patten, T.; Prankl, J.; Vincze, M. Multi-Task Template Matching for Object Detection, Segmentation and Pose Estimation Using Depth Images. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 7207–7213.
73. Tamadazte, B.; Marchand, E.; Dembélé, S.; Le Fort-Piat, N. CAD Model-based Tracking and 3D Visual-based Control for MEMS Microassembly. *Int. J. Robot. Res.* **2010**, *29*, 1416–1434. [CrossRef]
74. Chen, Y.; Medioni, G. Object modelling by registration of multiple range images. *Image Vis. Comput.* **1992**, *10*, 145–155. [CrossRef]
75. Besl, P.J.; McKay, N.D. A Method for Registration of 3-D Shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **1992**, *14*, 239–256. [CrossRef]

76. Guo, Z.; Chai, Z.; Liu, C.; Xiong, Z. A Fast Global Method Combined with Local Features for 6D Object Pose Estimation. In Proceedings of the 2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), Hong Kong, China, 8–12 July 2019; pp. 1–6.
77. Akizuki, S.; Aoki, Y. Pose alignment for different objects using affordance cues. In Proceedings of the 2018 International Workshop on Advanced Image Technology (IWAIT), Chiang Mai, Thailand, 7–9 January 2018; pp. 1–3.
78. Yu, H.; Fu, Q.; Yang, Z.; Tan, L.; Sun, W.; Sun, M. Robust Robot Pose Estimation for Challenging Scenes With an RGB-D Camera. *IEEE Sens. J.* **2019**, *19*, 2217–2229. [CrossRef]
79. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
80. Calonder, M.; Lepetit, V.; Strecha, C.; Fua, P. BRIEF: Binary robust independent elementary features. In Proceedings of the Proceedings of the 11th European Conference on Computer Vision: Part IV, Heraklion, Crete, Greece, 22 December 2011; pp. 778–792.
81. Nospes, D.; Safronov, K.; Gillet, S.; Brillowski, K.; Zimmermann, U.E. Recognition and 6D Pose Estimation of Large-scale Objects using 3D Semi-Global Descriptors. In Proceedings of the 2019 16th International Conference on Machine Vision Applications (MVA), Tokyo, Japan, 27–31 May 2019.
82. Zhang, H.; Cao, Q. Texture-less object detection and 6D pose estimation in RGB-D images. *Robot. Auton. Syst.* **2017**, *95*, 64–79. [CrossRef]
83. Pan, W.; Zhu, F.; Hao, Y.; Zhang, L. Fast and precise 6D pose estimation of textureless objects using the point cloud and gray image. *Appl. Opt.* **2018**, *57*, 8154–8165. [CrossRef] [PubMed]
84. He, Z.; Jiang, Z.; Zhao, X.; Zhang, S.; Wu, C. Sparse Template-Based 6-D Pose Estimation of Metal Parts Using a Monocular Camera. *IEEE Trans. Ind. Electron.* **2020**, *67*, 390–401. [CrossRef]
85. Zhang, H.; Cao, Q. Detect in RGB, Optimize in Edge: Accurate 6D Pose Estimation for Texture-less Industrial Parts. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 3486–3492.
86. Pan, W.; Zhu, F.; Hao, Y.; Zhang, L. 6D Pose Estimation Based on Multiple Appearance Features from Single Color Image. In Proceedings of the 2017 IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), Honolulu, HI, USA, 31 July–4 August 2017; pp. 406–411.
87. Dong, Z.; Liu, S.; Zhou, T.; Cheng, H.; Zeng, L.; Yu, X.; Liu, H. PPR-Net:Point-wise Pose Regression Network for Instance Segmentation and 6D Pose Estimation in Bin-picking Scenarios. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 1773–1780.
88. Chen, W.; Duan, J.; Basevi, H.; Chang, H.J.; Leonardis, A. PointPoseNet: Point Pose Network for Robust 6D Object Pose Estimation. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020; pp. 2813–2822.
89. Cao, Q.; Zhang, H. Combined Holistic and Local Patches for Recovering 6D Object Pose. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 2219–2227.
90. Li, W.; Sun, J.; Luo, Y.; Wang, P. 6D Object Pose Estimation using Few-Shot Instance Segmentation and 3D Matching. In Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China, 6–9 December 2019; pp. 1071–1077.
91. Chen, X.; Chen, Y.; You, B.; Xie, J.; Najjaran, H. Detecting 6D Poses of Target Objects From Cluttered Scenes by Learning to Align the Point Cloud Patches with the CAD Models. *IEEE Access* **2020**, *8*, 210640–210650. [CrossRef]
92. Li, Y.; Chen, C.; Allen, P.K. Recognition of deformable object category and pose. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–7 June 2014; pp. 5558–5564.
93. Li, Y.; Wang, Y.; Yue, Y.; Xu, D.; Case, M.; Chang, S.; Grinspun, E.; Allen, P.K. Model-Driven Feedforward Prediction for Manipulation of Deformable Objects. *IEEE Trans. Autom. Sci. Eng.* **2018**, *15*, 1621–1638. [CrossRef]
94. Caporali, A.; Palli, G. Pointcloud-based Identification of Optimal Grasping Poses for Cloth-like Deformable Objects. In Proceedings of the 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), Vienna, Austria, 8–11 September 2020; pp. 581–586.