sensors

MDPI

Review

A Survey of 6DoF Object Pose Estimation Methods for Different Application Scenarios

Jian Guan [1,2,3,4], Yingming Hao [1,2,3,*], Qingxiao Wu [1,2,3], Sicong Li [1,2,3] and Yingjian Fang [1,2,3,4]

[1] Key Laboratory of Opto-Electronic Information Processing, Chinese Academy of Sciences, Shenyang 110016, China
[2] Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China
[3] Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110169, China
[4] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: ymhao@sia.cn

Abstract: Recently, 6DoF object pose estimation has become increasingly important for a broad range of applications in the fields of virtual reality, augmented reality, autonomous driving, and robotic operations. This task involves extracting the target area from the input data and subsequently determining the position and orientation of the objects. In recent years, many new advances have been made in pose estimation. However, existing reviews have the problem of only summarizing category-level or instance-level methods, and not comprehensively summarizing deep learning methods. This paper will provide a comprehensive review of the latest progress in 6D pose estimation to help researchers better understanding this area. In this study, the current methods about 6DoF object pose estimation are mainly categorized into two groups: instance-level and category-level groups, based on whether it is necessary to acquire the CAD model of the object. Recent advancements about learning-based 6DoF pose estimation methods are comprehensively reviewed. The study systematically explores the innovations and applicable scenarios of various methods. It provides an overview of widely used datasets, task metrics, and diverse application scenarios. Furthermore, state-of-the-art methods are compared across publicly accessible datasets, taking into account differences in input data types. Finally, we summarize the challenges of current tasks, methods for different applications, and future development directions.

Keywords: object pose estimation; deep learning; 6DoF pose; computer vision

Citation: Guan, J.; Hao, Y.; Wu, Q.; Li, S.; Fang, Y. A Survey of 6DoF Object Pose Estimation Methods for Different Application Scenarios. *Sensors* **2024**, *24*, 1076. https://doi.org/10.3390/s24041076

Academic Editor: Anastasios Doulamis

Received: 19 December 2023
Revised: 29 January 2024
Accepted: 2 February 2024
Published: 7 February 2024

1. Introduction

Object pose estimation is a key task in the field of computer vision, whose main goal is to accurately obtain a 6DoF (6 degrees of freedom) representation of the object pose in real-life scenes. This representation plays a key role in providing comprehensive information beyond two-dimensional understanding. Specifically, it encompasses three-degree-of-freedom rotation and three-degree-of-freedom translation. The significance of this task is that it can provide the precise spatial position of objects, so 6DoF object pose estimation is increasingly important for various applications of computer vision, such as virtual reality, augmented reality [1,2], automated driving [3], and robotic operation [4]. The continuous advancement of computer vision theory and the rapid development of related fields have prompted extensive and insightful research on 6DoF object pose estimation. Existing reviews make it difficult to summarize the latest research in this field, and this article will fill this gap by summarizing different approaches from recent years.

In robotics, the ability to perform complex tasks such as precise manipulations in dynamic or unpredictable environments is crucial. Many tasks require scene understanding or object operations; 6DoF pose estimation provides comprehensive information regarding both position and orientation, enabling robots to execute tasks like recognition, localization, and grasping with heightened precision and accuracy [5]. To accomplish its grasping tasks

effectively, the robot needs to rely on information such as the position, contact, and occlusion captured by the camera. As the role of robots becomes increasingly more important for improving productivity and securing safety, 6DoF pose estimation has become a key technology. Beyond the grasping operations of robots, 6DoF pose estimation also plays a vital role in reducing environmental uncertainty. It works in two ways, including reducing the probability of robot collision and providing support for robot transportation or other activities [6,7].

Augmented reality (AR) and virtual reality (VR) technologies play an important role in the military, aerospace, education, entertainment and gaming, as well as many other fields. AR/VR systems can accurately track targets based on their poses, which is the foundation of immersive effects. One crucial aspect of this technique is 6DoF pose estimation, which enables the construction of a spatial mapping of the environment and provides information for the proper integration of AR/VR content. By estimating the precise 6DoF pose of the target, virtual objects can be rendered from different viewpoints, resulting in realistic visual effects in AR/VR experiences. In addition, precise alignment and interaction between virtual objects and the real world can be realized [8].

6DoF pose estimation plays an important role in many parts of autonomous driving tasks, including environment perception, obstacle detection, traffic condition prediction, and decision planning. It provides valuable information for trajectory planning and obstacle avoidance. With the rapid development of autonomous driving technology, the requirements for pose estimation precision are becoming increasingly high [9–11].

Ideally, the pose estimation method should be able to handle objects with different shapes and textures and show robustness to large occlusions, noise, and changes in light. Figure 1 shows two possible situations. Furthermore, it should balance accuracy and efficiency, especially in real-time navigation tasks.

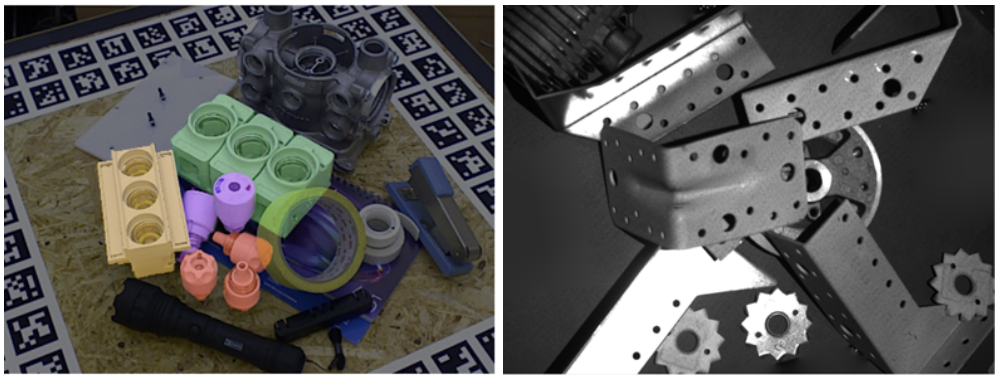

Figure 1. Different shapes (**left**) [12] and large occlusions (**right**) images [13].

In the field of computer vision, the traditional approach involves extracting target features such as points, edges, and lines directly from input images or point clouds. These features are then matched with a reference image or model to perform pose estimation. The positional relationship can be solved by measuring the coordinates of multiple points under two spatial coordinate systems and utilizing the positional relationship between marker points as constraints. However, traditional methods rely on manually designed feature extraction, and their performance may be affected by noise, data quality, or other factors [14].

Over the past few years, advances in technology have made data collection easier, and deep learning methods have performed well in many areas of computer vision. With the rapid development of deep learning technology, 6DoF pose estimation based on deep learning [15–17] has significantly improved in terms of accuracy, robustness, and adaptability to different scenes. In various daily life scenarios, as well as industrial robotics tasks, the estimation of 6DoF object poses is of great significance and can bring great convenience to production and daily life.

Figure 2 illustrates the sections of this paper where the existing methods can be coarsely categorized into two main groups based on whether the object model is necessary for the training stage: instance-level pose estimation and category-level pose estimation.

Several reviews of this task already exist [4,6,14,18], but unlike these methods, here, we focus on methods based on deep learning, organize the paper according to different input data types or method types, and summarize commonly used public datasets in the field. We summarize the application scenarios of different methods, organize commonly used datasets in the field, and provide a comprehensive summary of methodologies. Additionally, the paper discusses future expectations based on different application scenarios, which can provide a valuable reference for subsequent research work.

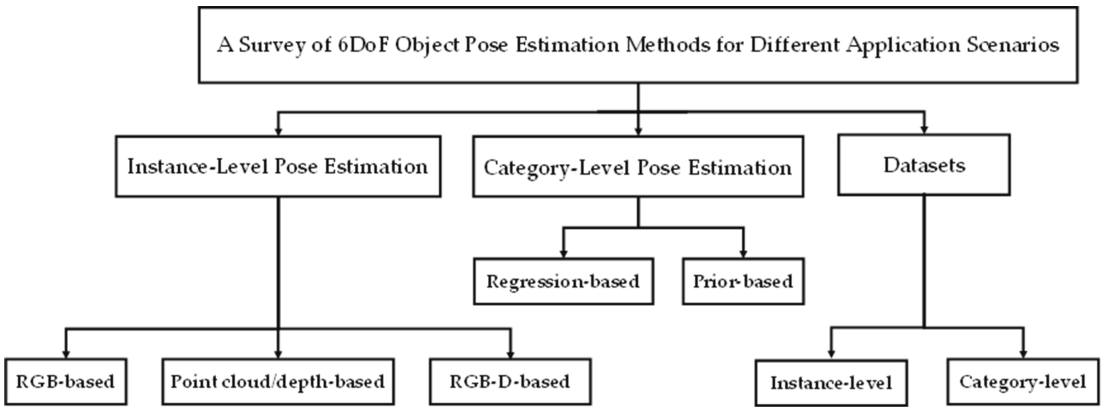

Figure 2. General Structure of the Survey.

2. Instance-Level 6DoF Object Pose Estimation

There has been significant research conducted on instance-level 6DoF object pose estimation. Depending on the input data, we classified these methods into three categories: RGB-based methods, point cloud or depth-based methods, and RGB-D-based methods. Then, we summarized the refinement methods.

2.1. RGB-Based Methods

With the advancement of deep learning, RGB-based 6DoF pose estimation methods have made significant progress in both theoretical and practical aspects. Figure 3 shows an overview of the typical RGB-based pose estimation methods; RGB images provide rich visual information and scene texture features that enable deep learning networks to extract effective representations of object poses. Furthermore, the widespread use and affordability of RGB cameras contribute to their cost advantage.

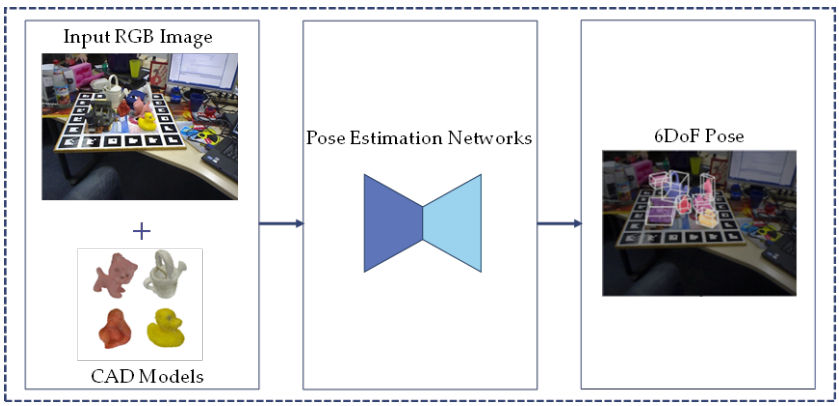

Figure 3. Schematic of a typical method for RGB-based pose estimation.

The task of pose estimation from RGB images faces several challenges, such as occlusion and lack of texture information. In this section, following the common classification, RGB-based methods can be divided into three classes: regression-based methods, template-based methods, and feature-based methods.

2.1.1. Regression-Based Methods

One of the straightforward approaches is that the 6DoF object pose estimation is considered as a regression task, which directly predicts the poses from the input RGB images without intermediate keypoint representations. A typical simplified flow of the regression-based approach is shown in Figure 4.

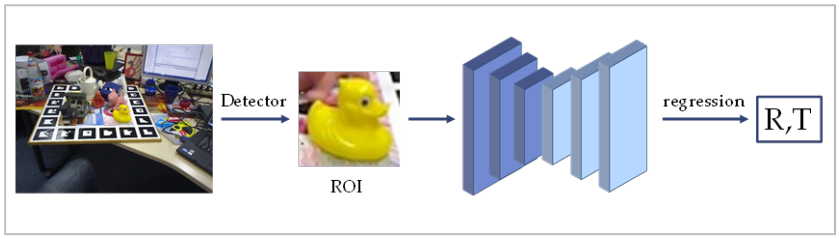

Figure 4. A typical flow of regression-based pose estimation methods.

PoseNet [19] is one of the first works to propose an end-to-end approach that introduces an end-to-end 6DoF pose estimation model using convolutional neural networks(CNNs). The model demonstrates the feasibility of deep learning for pose estimation by directly regressing the orientation and position of RGB images. However, this method is only used for human pose estimation.

PoseCNN [20] pioneers a direct learning approach for regression object 6D pose representation. The method treats rotations and translations separately, incorporates 2D centroid prediction based on hough voting, integrates additional prior knowledge for translation processing, and proposes a symmetric loss function for shape matching. It is worth mentioning that PoseCNN introduces the highly influential YCB-V dataset. However, the ICP algorithm [21] needs to be optimized for better accuracy, and the networks are specific and may have poor generalization. Overall, this is a classical work.

Deep-6DPose [22] extends the instance segmentation network Mask R-CNN [23] to pose estimation. It directly introduces a pose prediction branch into the framework, thus realizing an end-to-end regression-based pose estimation method. Different from PoseCNN [20], Deep-6Dpose does not require subsequent refinement steps, thus simplifying the process and improving efficiency. However, its performance decreases significantly when the pose changes significantly, and its accuracy still needs to be improved.

Hu et al. [24] proposes a segmentation-driven simple pose estimation network, which enables handling multiple objects occluding each other, even in the absence of texture. This method avoids the need for post-processing, but the estimation of small objects needs to be improved. Additionally, there is room for improvement in the network architecture and fusion strategies.

Another noteworthy contribution is YOLO-6D [25], which leverages the YOLO family [26–29]. YOLO-6D [25]converts the pose estimation problem into a nine keypoints regression task and utilizes the real-time framework of YOLO-V2 [30]. This approach has had a significant impact on subsequent research. As the YOLO family continues to evolve, using new YOLO frameworks may produce better results, but application in complex environments may be limited. NeRF [31] introduces a method for generating rendered images without relying on mesh models. Inspired by this, iNeRF [32] predicts the pose from a single RGB image, specifically targeting scenarios where object mesh models are not available during training or testing. iNeRF can be extended to category-level pose estimation, but it is susceptible to lighting and occlusion, and its real-time performance needs to be improved.

DeepIM [33] introduces iterative improvements by regressing the pose difference between the rendered pose assumptions and the input image. Building on DeepIM, CosyPose [34] improves by incorporating rotational continuity representation, symmetry-aware display processing, and network architecture updates. These improvements make it possible to recover a consistent scene across multiple views, thus facilitating 6D attitude estimation for multiple classes of objects. CosyPose obtained the best results on multiple datasets in the 2020 Benchmark for 6D Object Pose Estimation (BOP) Challenge [35] and many subsequent studies built upon it.

To better solve the occlusion problem, ZebraPose [36] proposes a method that uses a dense representation of the object surface with discrete descriptors. The method uses an encoder–decoder architecture for feature extraction and direct regression of pose without post-processing procedures. However, the method has limited generalization to instances with significant appearance differences, although it has also been shown that the dense correspondence method is more effective at solving problems related to occlusion.

Hai et al. [37] address the limitations of existing self-supervised methods, which often require additional depth [38,39] or segmentation mask information [40]. They propose a self-supervised pose optimization framework that employs a synthetic dataset generated from the 3D mesh of the target object. The network is trained solely on this dataset to obtain the initial pose, followed by rendering multiple synthetic images from different viewpoints. To bridge the domain gap between synthetic and real data, a pseudo-label-based optimization strategy is employed for refinement. However, this approach heavily relies on synthetic data for training initial poses, and there is room for improving the generalization of current methods.

To address the problem of the increasing runtime when performing multi-object tasks, EfficientPose [41] proposes an efficient end-to-end pose estimation method that utilizes two additional sub-networks for predicting translations and rotations, thereby reducing computational cost and eliminating post-processing steps. The method also proposes a data enhancement technique involving random rotation and scaling of images to improve generalization to small datasets. However, as EfficientPose relies on overall detection, it may be less effective in heavily occluded scenes.

GDR-Net [15] proposes a simple and effective method for geometrically guided pose estimation. It dynamically scales up the detection results of other methods as the inputs and utilizes intermediate representations of dense correspondences. A modified version of this method utilizes a more robust backbone network and was successful in the BOP2022 challenge [42], demonstrating impressive accuracy and speed.

These regression-based methods have made significant progress in RGB-based pose estimation. However, there are still some challenges, such as dealing with occlusion, coping with illumination variations, improving real-time performance, and enhancing generalization to different scenes and object appearances.

2.1.2. Template-Based Methods

Template-based methods typically require finding the most similar template of the target, this search is conducted among many templates labeled with true poses, and then performing 6D pose estimation. Template matching is also a broad class of direct methods used for pose estimation. A typical simplified flow of the template-based approach is shown in Figure 5.

SSD-6D [43] is an approach that extends the 2D target detection network to handle the pose estimation task. It introduces a pose estimation branch during the detection process and decomposes the pose space of the model, using the multiple detection results obtained, performing a series of pose estimations as templates, and selecting an optimal hypothesis. Additionally, it treats rotational regression as a classification task, which improves the training and learning of symmetric objects. However, an inherent limitation of the method is its reliance on regressing the 2D bounding box corners. This reliance can result in decreased accuracy, particularly for heavily occluded objects. Furthermore, methods such

as data augmentation need to be used to minimize the differences between synthetic images and real data.

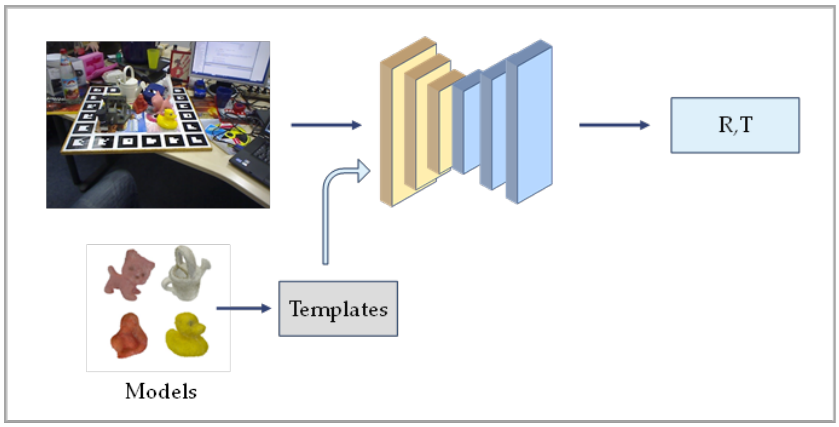

Figure 5. Typical flow of template-based pose estimation methods.

LatentFusion [44] introduces a novel framework for the 6D pose estimation of unseen targets by leveraging learned 3D representations. The network is capable of rendering the target from any viewpoint and directly optimizing the pose of the input image. The method achieves this by training the network on a large dataset of 3D shapes, enabling it to reconstruct and render objects accurately. Moreover, the use of multiple views during modeling allows for robust observations, and the consistency across these views enables the construction of a canonical representation, resulting in improved generalization to unseen targets. However, it's worth noting that LatentFusion's iterative optimization process during inference can be computationally expensive. Additionally, the method is sensitive to occlusions in the input data, which can lead to significant performance degradation when occlusions are present.

DPOD [45] is a pose estimation method that combines detection and matching using a dense matching-based approach. In the first stage, a detector predicts 2D frames. The second stage refers to the voting-based PVNet [46] method for template matching. DPOD can estimate poses from a single RGB image without requiring perfect segmentation. It demonstrates robustness in handling occlusion and lighting changes. However, it may be less effective for objects that lack distinctive color features.

PoseRBPF [47] utilizes a Rao-Blackwellized particle filter that samples object poses and estimates the discretized distribution of each particle's rotations using a pre-computed codebook. This method is effective at tracking object poses and is less susceptible to motion blur and occlusion. However, it may encounter difficulties when objects are heavily occluded or measurements deviate significantly from the synthetic training data. Furthermore, there is room for improvement in the discretized rotation representation in PoseRBPF.

OSOP [48] utilizes semantic segmentation to predict the mask of the visible parts, and renders 2D templates from various viewpoints, using the templates to first locate the approximate viewpoints, and then obtain the final pose after dense matching, which can be generalized to train unseen novel objects. It is applicable for tasks that need to be performed on new objects, although the domain gap between synthetic and real data needs to be handled.

Template-based methods have many advantages, including simplicity, speed, adaptability to changes in appearance, and better handling of weakly textured objects. It is a straightforward and intuitive method that can quickly detect and localize objects. However, template matching may face challenges in complex scenarios with occlusions, lighting variations, or objects lacking distinctive features. Incorporating local representations and addressing these challenges is crucial for robust template matching. Additionally, global matching in template matching can be influenced by the background and may perform

poorly when matching real images of unseen objects [49,50]. Therefore, the use of local representations becomes necessary for template matching of unseen objects.

2.1.3. Feature-Based Methods

The feature-based pose estimation method is widely used in the field of 6DoF pose estimation. The standard procedure involves extracting features from the input image, matching them with corresponding features in an existing 3D model, and then establishing the correspondence of 2D–3D coordinates using the Perspective-n-Point (PnP) algorithm. PnP is the corresponding method for solving 3D to 2D points. It describes how to estimate the pose of a camera when 3D space points and their positions are known. By leveraging the features extracted from the image and matching them with the features of the 3D model, this feature-based method establishes the 2D–3D relationship and enables accurate estimation of the target object's 6D position. A typical simplified flow of the feature-based approach is shown in Figure 6.

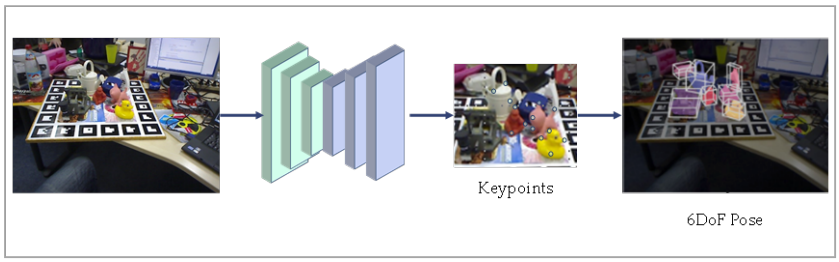

Figure 6. Typical flow of feature-based pose estimation methods.

The feature-based matching method has been extensively studied over time. Traditionally, feature points are extracted from two images, and then these feature points are compared to determine their correspondence. Global matching methods [51,52] perform well, with low computing power requirements, but they are sensitive to occlusion and noise, limiting their practical application. On the other hand, localized features are more robust when dealing with occlusion. Local feature extraction relies on feature detection and descriptors that should possess distinctiveness and invariance to certain transformations.

Traditional descriptors [53–57], such as SIFT [53], are manually designed and have certain limitations. They may not capture sufficient information, primarily describing geometric relationships, and can be less effective when the texture is not rich or the environment undergoes significant changes. In contrast, local feature detection and matching methods [58] based on deep learning have shown a better performance compared with traditional methods that rely on hand-crafted local features. These methods typically involve two steps: the first stage utilizes neural networks for feature extraction and obtaining 2D–3D correspondences, while the second stage solves the PnP problem. The differences between these methods mainly lie in how they establish the correspondence. These methods effectively leverage the advantages of CNN network structures, combining them with traditional computer vision techniques, and resulting in improved accuracy in pose estimation.

Pavlakos et al. [59] proposes a method that utilizes detected semantic keypoints to regress and compute the 6DoF pose in an end-to-end training fashion. This approach avoids the laborious process of point-by-point matching. However, it may be less effective when dealing with small objects and severe occlusion.

BB8 [60] utilizes segmentation methods to predict 3D boundary points based on 2D bounding boxes. It avoids the need for feature extraction and matching. Pose estimation is achieved by regressing the 2D coordinates of the inflection points of the projected 3D bounding box corners. BB8 demonstrates that the accurate and stable 3D pose estimation can be accomplished using only RGB information. The approach is extensible to new object categories without the need for predefined models. The estimation of symmetric objects has always been difficult. To solve this problem, BB8 explicitly handles it by range

transformation and constraining object-labeled poses during training. This approach to handling symmetry has become more widely used in subsequent works [37,61]. However, it may be less effective for untextured objects.

To address the challenges posed by severe occlusion, PVNet [46] builds upon the symmetry handling approach introduced in [60]. The proposed pose estimation framework in PVNet [46] regresses pixel-level vectors that point to the keypoints. These vectors are then used for voting on the keypoints locations, resulting in a spatial probability distribution of the keypoints. Additionally, the network predicts pixel orientations, which allows it to focus more on the local features of the object and mitigate the effects of background clutter. PVNet [46] effectively solves the problem of occlusion and has laid the foundation for much of the subsequent work in the field.

EPOS [62] is a pose estimation method that takes into account the symmetry of objects. It decomposes the pose space into symmetry-invariant and symmetry-related parts. The method discretizes the object's surface into fragments and predicts a probability distribution for each fragment, classifies pixels based on the associated object segments, and regresses coordinates. This approach is adaptable to various types of symmetrical objects, including those with reflective surface symmetry or rotational symmetry.

Pix2Pose [63] uses an untextured 3D model to regress pixel-level 3D coordinates from RGB images. It introduces a transformer loss function specifically designed for symmetric objects and trains a self-coding network with a Generative Adversarial Network (GAN) [64] to denoise the model and recover occluded parts. The method has been evaluated on the T-LESS dataset [12]. It uses the visible surface deviation as a metric, which measures only the distance error of the visible parts; this metric is independent of symmetry and occlusion, and the results of Pix2Pose outperform previous methods significantly. This method has good applicability in industrial-related scenarios.

RNNPose [65] proposes a pose refinement method based on Recurrent Neural Network (RNN) [66] design, using CAD models for rendering. It optimizes the error between the rendered image and the observed image using nonlinear least squares. RNNPose introduces a hybrid network trained with contrast to handle occlusion, making it more robust against errors and occlusion introduced by the initial poses. The method shows substantial improvements over the initial poses obtained from PoseCNN [20]. However, one limitation is that the training model is object-specific, and still needs to be improved if it is to meet the generalization requirements for unseen objects.

Onepose [8] presents a novel approach for 2D–3D feature matching using a graph attention network [67]. This method effectively preserves the graph structure information of feature tracking, resulting in more reliable and faster matching. It achieves a higher accuracy compared to PVNet [46], without requiring instance-specific training on the validation set. Hybridpose [68] proposes a network architecture based on PVNet [46], leveraging a prediction network with three intermediate representations using ResNet [69]. By fusing the features of keypoints, edges, and symmetry points, this approach expresses geometric knowledge with multiple intermediate representations, providing additional constraints on the object. The introduction of edge and symmetry point features improves the stability of position estimation. However, it is worth noting that training the network requires careful design.

CRT-6D [70] employs a sparse set of features based on key points on the object's surface, significantly reducing the impact of noise and computational cost. This method incorporates a fast refinement technique for better real-time performance, utilizing a deformable attention mechanism to handle occlusion robustly. However, it should be acknowledged that the accuracy of CRT-6D is lower compared with the approach proposed in Zebrapose [36].

Among the feature-based methods, pose estimation using the PnP algorithm is widely adopted, and exploring ways to improve the PnP algorithm is also a direction. Previous approaches [25] have employed various techniques such as direct usage of the PnP algorithm, the EPnP [71] method [46], or combining PnP with RANSAC [45].

However, the non-differentiability of the PnP problem at some points poses challenges for convergence during training. To address this issue, the CVPR2022 Student Best Paper, EPro-PnP [72], tackles the problem of solving the camera pose by transforming it into a probability density prediction task. By learning the 2D–3D correlation based on the ground truth pose, EPro-PnP achieves end-to-end training of a network that predicts the probability density of the pose. This approach not only solves the PnP pose optimization problem, but also provides insights for optimizing other networks. It enables stable and flexible training of pose estimation networks based on PnP geometry optimization, surpassing the state-of-the-art performance of the outdated CDPN [73] method.

EPro-PnP [72] essentially applies the multi-class softmax concept to the continuous domain, which can be extended not only to other geometric optimization-based 3D vision problems [40], but also theoretically generalized to train general models with nested optimization layers.

In general, the method based on global feature matching offers advantages in terms of speed, while the method based on local feature matching better fulfills the accuracy requirements. Global features rely heavily on semantic information, providing stronger discriminative capabilities, while local features rely more on texture information, making them more robust to image variations. The sparse counterpart of feature-based methods requires less computational resources and can perform better in some real-time applications. It offers an overall satisfactory performance. In contrast, the dense counterpart utilizes richer information and is particularly effective at handling occlusion problems. However, it demands higher computational resources. Feature-based methods, in general, are fast and robust, especially when dealing with texture-rich objects. However, they may be less effective when applied to weakly textured objects, where the distinction between objects and the background is weak, and detecting keypoints becomes challenging.

2.1.4. Refinement Methods

Refinement methods play a crucial role in improving the performance of pose estimation by refining the initial coarse pose. RGB-based methods often require subsequent optimization, and several popular approaches have been proposed [33,34,65,74]. Figure 7 shows a simplified flow of the refinement method using the renderer.

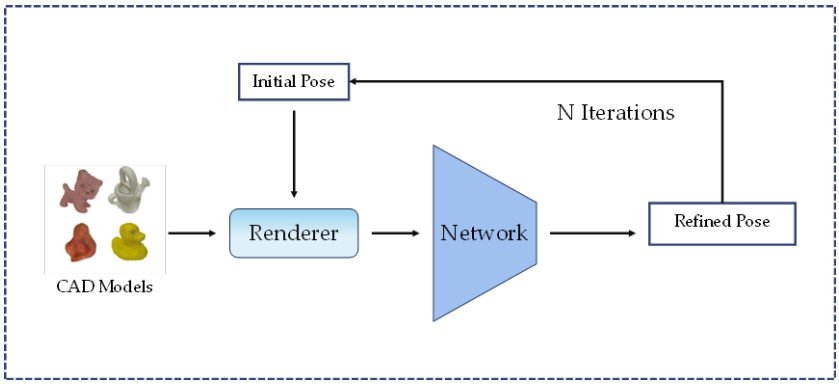

Figure 7. General flow of the refinement method.

PoseCNN [20,75] utilizes the Iterative Closest Point (ICP) algorithm to align known models with the depth map for pose refinement. DeepIM [33] takes an iterative approach, using a pose refinement network to minimize the difference between the observed image in the current pose and the rendered image. Another method [76] introduces a novel visual loss for pose updating, which aligns contours to refine the pose. HybridPose [68] proposes a pose refinement method that utilizes a robust norm optimization of the reprojection error, termed GM robust norm optimization.

DPOD [45] presents a pose refinement network that includes modules for independent regression of rotations and translations. It optimizes the estimation results based on the difference between the image rendered by the predicted pose and the real input. CosyPose [34] draws inspiration from bundle adjustment and globally refines all objects and camera poses by minimizing multi-viewpoint reprojection errors. Repose [74] introduces a faster refinement approach by extracting the image features using U-Net [77]. RNNPose [65] formulates pose optimization as a nonlinear least squares problem.

Some of the RGB-based methods are uniformly compared in Table 1, employing metrics including Average Distance of Detected points (ADD) and Symmetric ADD (ADD-S) on both the LineMod (LM) [78] dataset and LineMod-Occlusion (LM-O) [79] dataset, as well as the area under the curve (AUC) of ADD-S on the YCB-Video (YCB-V) dataset [20].

Table 1. RGB-based pose estimation methods and results.

Methods	Years	Input	Types	LM	LM-O	YCB-V
PoseCNN [20]	2017	RGB	Regression	-	24.9	61.3
SSD-6D [43]	2017	RGB	Refinement	79	-	-
YOLO-6D [25]	2018	RGB	Regression	55.95	-	-
DeepIM [33]	2018	RGB	Refinement	88.6	55.5	81.9
Deep-6DPose [22]	2018	RGB	Regression	65.2	-	-
BB8 [60]	2018	RGB	Refinement	43.6	-	-
PVNet [46]	2018	RGB	Feature	86.27	40.77	73.4
Hu et al. [24]	2019	RGB	Regression	-	27.0	-
CDPN [40]	2019	RGB	Feature	89.86	-	-
DPOD [45]	2019	RGB	Template	95.2	47.3	-
Pix2Pose [63]	2019	RGB	Feature	72.4	32.0	-
Efficientpose [41]	2020	RGB	Regression	97.35	83.98	-
CosyPose [34]	2020	RGB	Regression	-	-	84.5
LatentFusion [44]	2020	RGB	Template	87.1	-	-
Hybridpose [68]	2020	RGB	Feature	91.3	47.5	-
GDR-Net [15]	2021	RGB	Regression	93.7	62.2	84.4
SO-Pose [80]	2021	RGB	Feature	94.0	62.3	83.9
RePose [74]	2021	RGB	Refinement	96.1	51.6	82.0
PoseRBPF [47]	2021	RGB	Template	79.76	-	-
Zebrapose [36]	2022	RGB	Regression	-	76.9	85.3
RNNPose [65]	2022	RGB	Refinement	97.37	60.65	83.1
DPOD-v2 [81]	2022	RGB	Feature	93.59	-	-
EPro-PnP-v2 [72]	2023	RGB	Feature	96.36	-	-
Hai et al. [37]	2023	RGB	Regression	92.2	65.4	-
CRT-6D [70]	2023	RGB	Feature	-	66.3	87.5

2.2. Point Cloud or Depth-Based Methods

In some situations of object pose estimation, such as in industrial environments, the limitations of RGB-based methods are evident due to the lack of color and texture information [82]. In contrast, methods based on point clouds or depth maps may offer unexpected advantages, while RGB images lack geometric data. Depth information or point cloud information contains rich shape geometry information, which is significant for inferring the pose of objects [83,84].

Methods based on depth maps or point clouds may have advantages in training data. Methods based on real images usually require expensive manual labeling. The annotation cost can be reduced when using synthetic images, but the domain gap becomes an important issue. Methods based on depth information or point clouds have smaller domain gaps with more robust results [85]. Figure 8 shows a typical approach using point clouds as inputs.

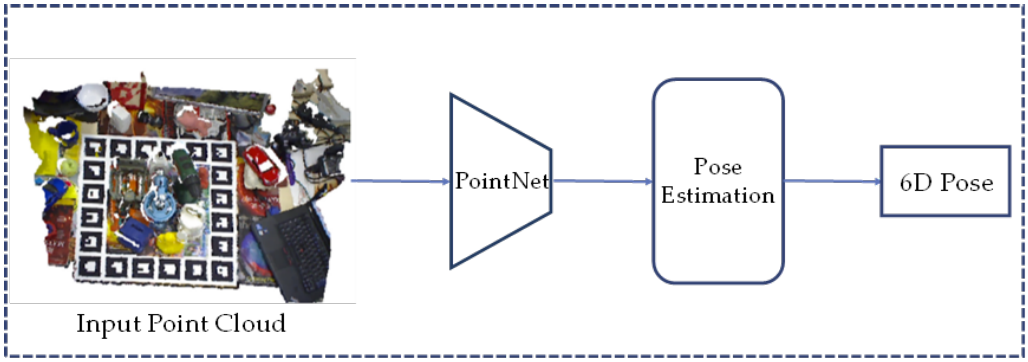

Figure 8. Point cloud-based typical approach flow.

Research on point cloud or depth maps aims to achieve a balance between accuracy and computational speed by combining global and local features, and is therefore particularly suitable for objects with different surface textures [86]. Colored point-pair features have been introduced in traditional methods [87] to improve discrimination and accuracy by exploiting the color information. There are also some works where the point cloud is used directly as an input to achieve the desired results through deep learning.

2.2.1. Point Cloud-Based Methods

Liu et al. [88] proposed a new downsampling method that combines edge and geometric information to estimate complex shapes, oriented to the requirements of medicine, and a pose estimation method based on edge-enhanced point-pair features for the characteristics of the spine structure. This method showed competitiveness when dealing with complex shapes and symmetric objects and is applicable in automated surgery. However, this method may not perform as expected when dealing with tiny objects or asymmetric cylindrical objects.

Previously, 3D data faced inherent challenges when represented using 3D voxel meshes or multi-view projections, including high computational requirements and loss of geometric information. To address this problem, Pointnet [89] proposed a solution based on point cloud data. At the same time, to address the problem of disorganization of the point cloud, the method employs a simple symmetric function to aggregate the vertex information, starting with global feature extraction, and then performs point cloud segmentation or classification.

Based on Pointnet [89], the Pointnet++ algorithm [90] further improves the acquisition and processing of localized information in point clouds. Both networks play an important role in various point cloud-based tasks. Another innovative approach is Pointvotenet [91], which employs a 3D segmentation method, based on Pointnet [89], to estimate the pose directly from a disordered 3D point cloud, unlike traditional projection-based methods. However, Pointvotenet minimizes the keypoints of symmetry in the process, and there is still space to improve the performance, while in real-time demanding scenarios this may not be applicable. The RandLA-Net algorithm [92] introduces stochastic downsampling to point cloud processing, simplifying network complexity while preserving local features through a feature aggregation module.

The PointPoseNet [93] method performs segmentation and vector prediction of point clouds obtained from RGB-D images, which in turn results in optimal pose estimation. It works well for scenes in the presence of occlusions, but the runtime increases when faced with the need to process multiple instances.

Point pair features (PPF) [86] is a method of global modeling and local matching, and PPF-based methods are known for their potential to achieve a high accuracy, while they often come with the drawback of a high computational complexity. In response to this challenge, PPFNet [94] combines PPF with deep learning techniques to enhance 3D point matching and point cloud feature extraction. The experimental results demonstrate that the learned features outperform traditional methods significantly in tasks like 3D shape

retrieval and matching. The deep-learning-based 3D target recognition methods show a superior generalization performance compared with traditional methods.

In real-world applications, such as the task of robot bin-picking applications, where objects are randomly stacked and occlusions in the scene are common, the Point-Wise Pose Regression Network (PPR-Net) [82] is a straightforward and effective solution. This network utilizes input point cloud data to simultaneously process instance segmentation and pose estimation, which can effectively identify occlusion relations and handle symmetric objects, thus achieving favorable results in practical applications.

Hoang et al. [95] introduced a detection method in pose estimation that relies on a voting mechanism designed for point cloud inputs without segmentation. Their method incorporates an attention module for learning rich associations between object parts and instances to improve the pose estimation performance. This method works well when dealing with datasets containing industrial parts, and it may be suitable for industrial scenarios.

CloudAAE [85] proposed a new method for reconstructing point clouds by regressing 6D poses using desired viewpoints and synthetic data based on 3D models, with temporary point locations. By augmenting the autoencoder to generate a noiseless, occlusion-free point cloud, the online approach offers advantages in terms of time efficiency and hardware storage over rendering methods. However, the method relies on the iterative closest point (ICP) algorithm [21] for optimization, which may not be optimized in the case of severe occlusions.

Point cloud-based pose estimation has important applications in robotic 6DoF grasping, and with the co-development of complete and local point cloud methods, the robustness of industrial grasping, as well as its adaptability to the environment, has improved greatly.

2.2.2. Depth-Based Methods

A common method for 6D pose estimation from depth images is to convert the depth image into a point cloud, and then perform pose estimation through the obtained segmentation mask, as shown in Figure 9.

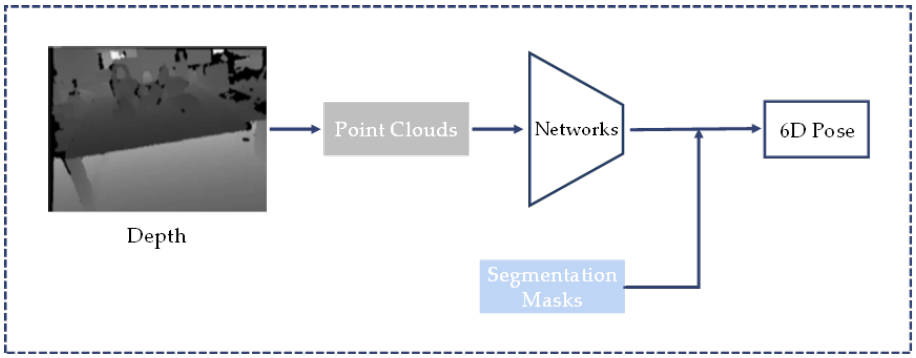

Figure 9. Depth-based typical approach flow.

A brand new framework, SwinDePose, was proposed in [96], which extends the Swin Transformer [97] to pose estimation using depth information. The combination of the Swin Transformer and pose estimation achieves a high accuracy by fully leveraging point cloud information and vector data from the depth map. It also handles occlusion well, but the performance depends on the quality of the annotations.

OVE6D [98] is trained using purely synthetic data, estimated from a single depth map and segmentation mask, and decomposes the pose estimation task into viewpoint, in-plane rotation, and translation. It can be easily generalized without parameter optimization in new objects. It works well on the T-LESS dataset [12], but only applies to the model of the object and to cases where instance segmentation masks are readily available.

Methods based on point clouds or depth maps may encounter challenges when dealing with reflected light on the surface of objects. This reflection problem can hamper the accurate capture of the actual point cloud data of objects, thus affecting the quality of subsequent works and tasks. In addition, there are relatively few methods dedicated to point cloud or depth map-based position estimation. In current industrial applications, most of the point cloud methods are still adapted to cope with the challenges by improving the traditional methods. It is notable, however, that these networks also help support other methods. A typical example is PVNet [46], which utilizes the principles of PointNet [89]. This approach stands out for its efficient location estimation capabilities and has had a significant impact in the field.

Table 2 presents a systematic comparison of point cloud-based or depth-based methods. Evaluation metrics are consistently applied across all three datasets [20,78,79], utilizing the ADD(-S) metric for the assessment of the pose estimation performance.

Table 2. Point cloud or depth-based pose estimation methods and results.

Methods	Years	Input	LM	LM-O	YCB-V
Gao et al. [99]	2020	D	-	-	94.7
Pointvotenet [91]	2020	Point Cloud	96.3	75.1	-
CloudAAE [85] + ICP	2021	Point Cloud	95.5	66.1	94.0
OVE6D [98]	2022	D	96.4	70.9	-
Depth-Based [96]	2023	D	97.5	77.1	-

2.3. RGB-D-Based Methods

Methods that rely solely on RGB images may be susceptible to challenges such as cluttered backgrounds, lighting changes, and texture differences, while methods based solely on point clouds face the problem of difficult data processing. Combining RGB images with depth information can enhance the ability to extract target geometric data, thereby improving the pose estimation performance in complex environments. Figure 10 shows a simplified flow of a typical approach using RGB-D as the input.

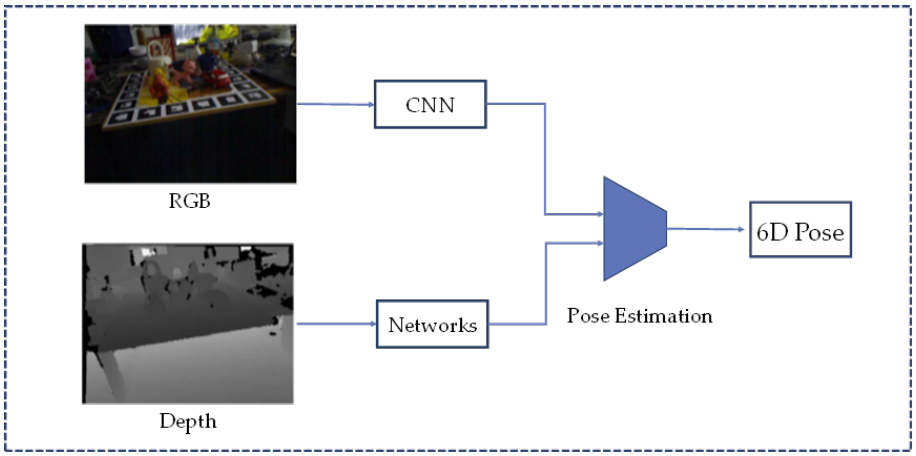

Figure 10. RGB-D-based typical approach flow.

The main challenge in RGB-D-based methods lies in fully utilizing the appearance information from RGB images and the geometric information from depth images. Early RGB-D estimation approaches often required the extraction of information from RGB and depth images separately. For instance, in [100], pose estimation was achieved by clustering 3D feature points in the object model, allowing for the extraction of features in the object shape that are independent of perspective changes and enabling cross-view matching.

However, although cross-view methods can provide richer information, they may require a large amount of storage space [101] or require complex post-processing [20], limiting their availability for complex scenes and real-time applications. In the case of symmetric objects, it is common to restrict the range of viewpoints, necessitating additional processing steps, such as in BB8 [60] for view classification and PoseCNN [20] for average distance computation between transformed models and estimated poses. Nonetheless, the process of finding the nearest 3D point can be time consuming.

On a different note, Li et al. [102] introduced a method to incorporate depth information as an additional channel. The network was designed by combining RGB and depth data in feature channel dimensions. This method has proven to be effective for multi-object instances and handling occlusions, but it may not perform optimally in single-view scenarios. The complexity of the network structure can also result in higher time costs. In this section, RGBD-based methods are classified into three categories: fusion-based methods, keypoints-based methods, and other methods.

2.3.1. Fusion-Based Methods

To tackle challenges related to occlusion and poor lighting conditions, DenseFusion [103] employs separate feature extraction and dense fusion of color and depth information. A pose estimate is generated for each pixel and the final result is obtained by voting. This approach considers the structural information of the depth channel, leading to accurate object pose estimation. Remarkably, it is nearly 200 times faster than the PoseCNN [20] with the ICP combination method. However, DenseFusion is limited to estimating the 6D pose of known objects and demands high-quality depth data and substantial computational resources.

MoreFusion [5] is tailored to scenarios where objects are known in robotics applications, focusing on solving pose estimation problems in the contact and occlusion of different objects. It achieves this by fusing segmentation masks into volumetric maps to represent occupied and free space. This approach enables pose estimation with awareness of the peripheral information, initial rough voxel reconstruction, and multi-object pose estimation, even in cases of occluded contact. Differentiable collision refinement and CAD model alignment support robot planning for grasping tasks in complex scenarios. When compared with DenseFusion [103], MoreFusion [5] also performs well in severe occlusion situation, and integrates more physics knowledge into the optimization framework.

FFB6D [104] enhances DenseFusion [103] with an improved fusion module. Fusion is applied at each coding and decoding layer to maximize the utilization of local and global information from another network. This simplifies keypoint localization and yields accurate pose estimation, resulting in a high accuracy. However, it is important to note that the effectiveness of deep fusion is highly dependent on data quality, with data noise significantly impacting performance. Furthermore, post-processing operations account for over half of the time cost.

From the above, it can be seen that the fusion-based method can utilize the two types of data more elegantly and is more robust to occlusion environments.

2.3.2. Keypoints-Based Methods

Keypoints-based methods are also an influential class of methods, which achieve position prediction by finding keypoints in an object through correspondences. A typical approach to the brief flow is shown in Figure 11.

PointNet [89] serves as a solid foundation for methods like PointFusion [11]. However, it has certain limitations in effectively extracting local point cloud features. To address this, PVN3D [105] introduces a two-stage approach, encompassing a feature extraction module, a keypoint detection module, a semantic segmentation module, and a centroid voting module. These modules work together to identify key points of objects through voting and clustering. Subsequently, after detecting the 3D key points of the target, the least squares method is used to fit the pose. Notably, the combination of 3D keypoints

and semantic segmentation enhances the overall performance, making the approach more robust, especially in the presence of natural occlusion.

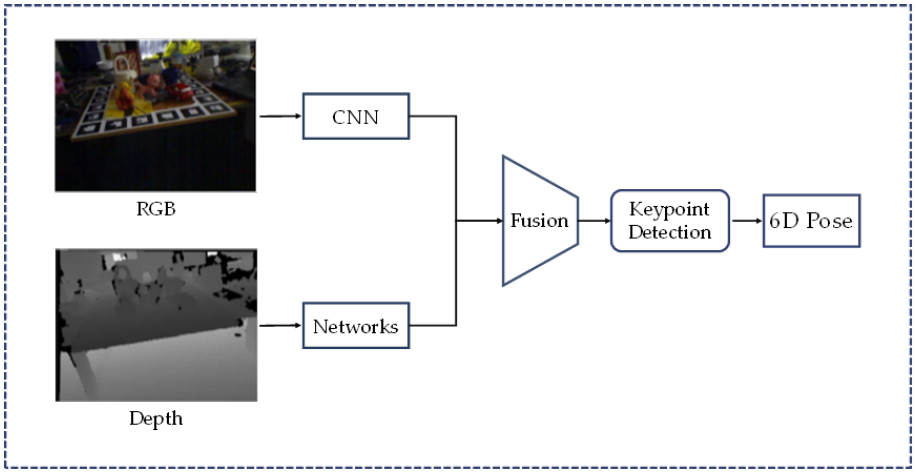

Figure 11. Keypoint-based typical approach flow.

Lin et al. [106] propose an end-to-end regression-based pose estimation method rooted in geometric information. This method supervises the decomposition of keypoint offsets into unit vectors and lengths and introduces an improved keypoint sampling strategy to ensure an adequate number of sampling points for small objects. However, it encounters challenges when addressing symmetric cases due to the lack of clear keypoint definitions for symmetric objects.

Zhou et al. [107] employ Deep Fusion Transformer (DFTr) blocks to elevate pose estimation by aggregating globally enhanced features across different modalities, facilitated by semantic similarity. They introduce a globally optimized voting algorithm to obtain accurate keypoints and exhibit robustness in dealing with various occlusions and symmetries while maintaining real-time performance. But improper selection of DFTr blocks can lead to overfitting, and the computational demands are relatively high.

The keypoints-based approach is more robust to noise and has a relatively good estimation, which makes it more practical, but it requires the determination of suitable keypoints.

2.3.3. Other Methods

In addition to the two methods mentioned above, there are many other studies based on RGB-D. Addressing the issue of previous methods employing separate networks for RGB and depth information extraction, Uni6D [108] introduces a unified CNN framework based on Mask R-CNN [23]. This framework incorporates additional UV data as an input to resolve the projection decomposition problem. Uni6D stands out for its efficiency in terms of time and cost, and achieves approximate accuracy on the YCB-V [20] dataset. It is exciting that it is 7.2 times faster than the FFB6D [104] method. Nevertheless, it has to be recognized that the simplification process may lead to accuracy degradation and further research is necessary, especially when denoising RoI features.

G2L-Net [109] takes a global-to-local approach, focusing on extracting point clouds from RGB-D data through 2D detection. The network performs 3D segmentation and translation prediction based on the coarse point cloud. It also captures viewpoint perception information using point-based features. G2L-Net estimates the initial rotation in the coordinate system transformed from the fine point cloud and further enhances the accuracy by considering rotation residuals between the predicted and true values. Impressively, G2L-Net achieves a good real-time performance despite the multi-step process.

More comprehensive utilization of geometric information has been shown to help mitigate issues related to color and appearance interference, random occlusions, and generalization from unseen instances. Previous methods that leverage geometric information often exhibit weak explanatory and generalization capabilities. In response to this, Stable-

Pose [110] introduces the concept of geometric stability to 6DoF pose estimation for the first time. Operating by the geometric stability principle, Stablepose [110] stands apart from approaches like EPOS [62], which involve sampling from a template model and regressing it as the 3D coordinates of image pixels. Instead, Stablepose learns the pose by focusing on geometrically stabilized portions of the point cloud derived from depth images, particularly emphasizing planar and cylindrical information. It accomplishes this by utilizing a minimum of three patches and predicting the pose for each patch through a sub-network. This approach significantly enhances the robustness of pose estimation in occluded scenes, as well as in objects that are not fully visible.

Building on the principles of EPOS [62], SurfEmb [16] presents a technique to learn a continuous dense distribution with the aid of contrast loss. This allows the model to capture multimodal distributions on an object's surface, making it more effective at handling symmetry and representing positional ambiguity. However, it is important to note that the position optimization process may encounter challenges when surface changes are subtle, and the approach involves four stages.

MegaPose [111] proposes a method that provides pose estimation of novel objects from RGB or RGB-D images. Rough pose estimation is performed first by classification and then refined by rendering synthetic views, which is simple to couple with other detection methods. The method is tested for its performance on multiple datasets and is suitable for real robots operating on unknown objects, but the runtime needs to be considered and not all rough initial poses can be successfully refined.

Lipson et al. [112] propose an end-to-end network that utilizes geometric knowledge to refine the pose and correspondence through coupled iterations and dynamically reject outliers. This method uses a novel bidirectional PnP algorithm, where the entire network can learn to optimize and perform pose updates. The refinement method may result in a less effective local optimal solution when the initial pose rotation error turns out to be large. This method also works well when only RGB images are used as the input.

Numerous pose estimation methods, including those referenced in [16,33,36,103], require object detection methods. However, in complex scenes with a poor detection performance, the estimation results of these methods will be greatly affected. To address this issue, Hai et al. [113] introduce a rigidity-aware detection method. This innovative approach capitalizes on the inherent rigidity property of the task object and formulates bounding boxes by sampling from the visible region rather than including the occluded part. The robustness of the target detection is enhanced, and the detection results can further improve the pose estimation effect.

The methods with RGB-D input offer several advantages and disadvantages. On the positive side, the combination of color information and depth information allows for more accurate and robust pose estimation, particularly in challenging scenarios with occlusion or poor lighting conditions. The depth data provide valuable geometric information, enhancing the recognition and localization of objects. Additionally, this approach can be instrumental in real-world applications such as robotics, where precise pose estimation is crucial. However, there are some drawbacks to consider, including increased computational demands due to processing both RGB and depth data.

In summary, RGBD-based methods provide a unique advantage by harnessing the synergy of RGB and depth information, resulting in a substantial improvement in estimation accuracy. How to effectively utilize different data is the key to these methods.

Table 3 presents a comparison of some RGBD-based methods. Evaluation metrics are consistently applied across all three datasets: LM, LM-O, and YCB-V [20,78,79], utilizing the ADD(-S) metric for the assessment of pose estimation performance.

Table 3. RGB-D-based pose estimation methods and results.

Methods	Years	Input	LM	LM-O	YCB-V
Li et al. [102]	2018	RGB-D	-	-	94.3
DenseFusion [103]	2019	RGB-D	94.3	-	91.2
Morefusion [5]	2020	RGB-D	-	-	91.0
PVN3D [105]	2020	RGB-D	99.4	70.2	91.8
G2L-Net [109]	2020	RGB-D	98.7	-	92.4
PR-GCN [114]	2020	RGB-D	99.6	65.0	95.8
FFB6D [104]	2021	RGB-D	99.7	66.2	92.7
Uni6d [108]	2022	RGB-D	-	-	88.8
E2EK [106]	2022	RGB-D	99.8	75.3	94.4
RCVPose [115]	2022	RGB-D	99.4	70.2	95.2
Deepfusion [107]	2023	RGB-D	99.8	77.7	94.4

3. Category-Level 6DoF Object Pose Estimation

Category-level 6D object pose estimation is designed to predict the complete pose of rotations, translations, and dimensions of object instances observed in a single arbitrary view of a cluttered scene. Estimating the pose and shape of daily objects is also an essential task, the majority of the previously discussed instance-level methods rely on accurate CAD models, but in daily life environments, it is hard to obtain CAD models of objects in advance, whereas category-level pose estimation methods aim to estimate the poses of arbitrary shapes in the same category without a priori assumptions of known CAD models, and it is starting to attract more attention by dealing with multiple instances of real-life scenarios [116–118].

3.1. Regression-Based Methods

The regression-based approach is a single-stage approach—one of the most straightforward. A simplified flow diagram of a typical process for such methods is illustrated in Figure 12.

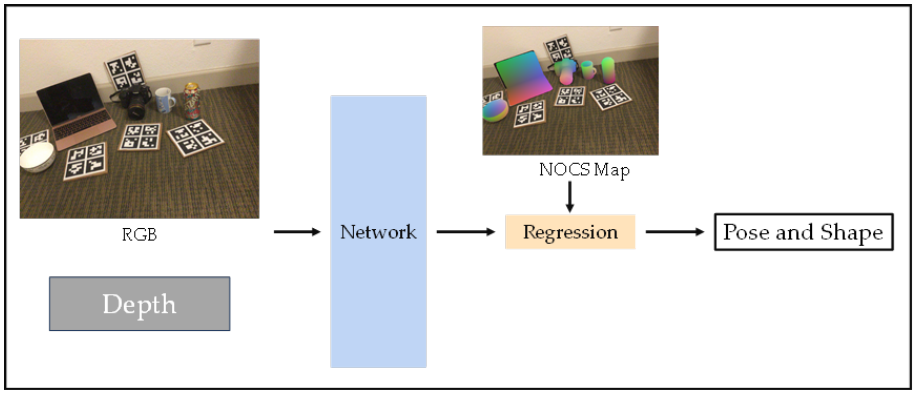

Figure 12. Typical flow of regression-based category-level pose estimation methods.

Category-level pose estimation faces challenges due to the unavailability of ground truth data. NOCS [116] addresses this by introducing a context-aware mixed reality approach, and it can be considered a pioneering work. To handle different and unseen object instances within a category, NOCS proposes a Normalized Object Coordinate Space (NOCS) and a dataset frequently used in category-level pose estimation tasks. Additionally, to cope with the symmetry of real-life objects, an axis of symmetry is defined for each category in the training data, ensuring that predefined rotations result in consistent loss values. While this approach enables robust pose and size estimation for unseen objects in real environments through direct regression, forming a uniform within-category representation remains challenging.

DualPoseNet [119] introduces a novel approach by stacking both an implicit and an explicit pose decoder on a shared pose encoder. This architecture allows for complementary supervision during training, ensuring the consistency of the predicted pose between the two decoders through an adaptive loss term. To further enhance its capabilities, DualPoseNet incorporates a spherical fusion module designed to facilitate more efficient learning from the input appearance and shape features. This method predicts a more compact bounding box, achieves more accurate pose estimation results, and performs well on instance-level tasks.

For addressing intra-class object variations, FS-Net [17] introduces a novel data augmentation method designed to enhance efficient feature extraction. In the context of category-level pose estimation, this approach proves valuable for handling objects with diverse shapes. This method uses only a limited amount of real data for training, demonstrating its proficiency in efficiently extracting category-level features from a small dataset. It proves effective for tasks characterized by a limited number of samples, but its success relies on the use of a high-performance and robust detector.

Centersnap [120] adopts a unique perspective by considering objects as spatial centers, where each center encapsulates a complete representation of an object's shape and pose. Notably, objects within the same category consistently retain the same semantics, even when their shapes are different. This approach is less demanding on computational resources and is better suited for tasks with real-time requirements.

Networks of direct regression methods are easier to deploy, although somewhat less able to cope with change. The subsequent emergence of methods that utilize prior knowledge can significantly improve the overall performance.

3.2. Prior-Based Methods

Exploiting the prior knowledge in category-level pose estimation tasks can be an effective method. By incorporating the prior knowledge learned in the provided instances, more accurate estimates can be obtained. SPD [121] has introduced a prior-based framework to tackle intra-class variation, which has become one of the mainstream approaches. The simplified flow of the method based on prior knowledge is shown in Figure 13.

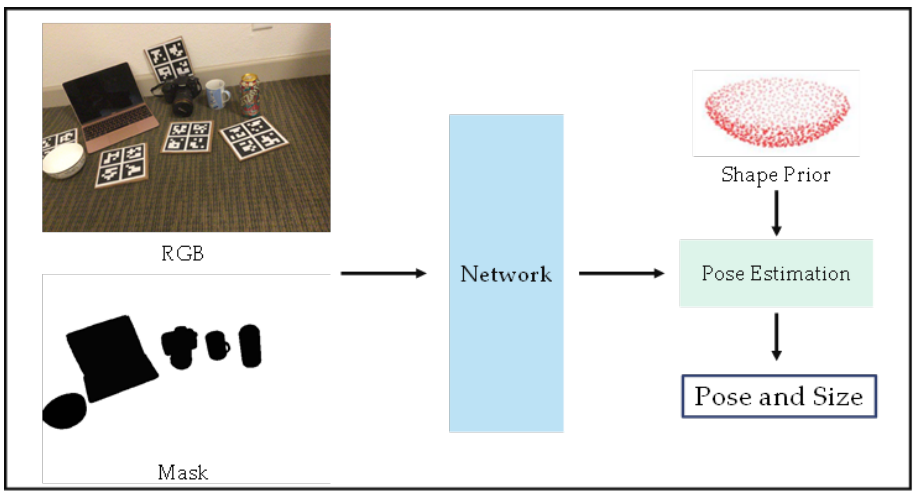

Figure 13. Typical flow of prior-based category-level pose estimation methods.

Subsequent work, ACR-Pose [122], emphasizes the importance of reconstructing canonical NOCS representations. ACR-Pose employs an adversarial training scheme consisting of a reconstructor and a discriminator to improve the network's ability to reconstruct high-quality canonical representations, enhancing the estimation accuracy, especially in challenging intra-class scenarios. The use of adversarial reconstruction loss has influenced subsequent category-level pose estimation methods, seeking to overcome

inherent intra-class variation. It is perhaps less applicable for scenarios where occlusion is common, as this method may fail in cases of occlusion or truncation.

SGPA [123] introduces a canonical prior model with shape deformation for pose estimation, further enhancing the correlation using structural similarity dynamics adaptation. However, the prediction based on the prior model may be less accurate when dealing with significant shape differences between categories.

To minimize the effect of domain gaps caused by the use of synthetic data, DPDN [124] proposes a method based on a deep prior deformation network, which introduces a self-supervised objective through a type of coherent learning to improve the sensitivity to pose changes. The method is suitable for situations where synthetic images are used for training.

RGB features are sensitive to color variations. In contrast, the introduction of additional shape prior features makes the results more robust, and methods based on prior knowledge are becoming popular.

3.3. Other Methods

In addition to the above main methods, there have been many other methods [8,125–127] for category-level pose estimation in recent years. To address the potentially high computational costs of complex multi-stage methods, Li et al. [125] leverage RGB-D images for single-stage object pose and shape estimation. Their method utilizes semantic primitives within a generative model to allow semantic features to model diverse shapes and establish connections between the observed point cloud and implicitly generated shapes. The optimization of an object's shape in arbitrary poses is achieved by using a novel SIM(3) invariant descriptor, delivering superior optimization results. However, it is important to note that this method may not account for occlusion and could result in ambiguous single-view estimations. Furthermore, utilizing implicit representations for shape inference is more complex than direct regression of shape parameters.

Chen et al. [126] introduce a method that leverages category information without relying on CAD models. Their approach involves synthesizing object images from various viewpoints using a generative adversarial network, combining a gradient-based fitting process with a parametric neural image synthesis module. This module can implicitly represent the appearance, shape, and pose of an entire object class, which eliminates the need for explicit CAD models for individual object instances. Notably, using only RGB images as inputs, this method can accurately recover the orientation information. However, achieving full 6DoF poses necessitates the incorporation of additional depth information to overcome scale ambiguity.

OnePose [8] adopts the concept of visual localization, exclusively using RGB images without reliance on CAD models. It can construct representations of specific objects from simple video scans using only a few samples. This unique approach allows it to handle objects from any class without the necessity for instance- or class-specific network training. OnePose excels in delivering fast and accurate position estimation, all without the need for prior knowledge. However, it is worth noting that the method relies on local feature matching and may encounter challenges when dealing with untextured objects.

Lin et al. [127] propose a keypoint-based single-stage method for category-level pose estimation only using a single RGB image as the input. Few studies have been conducted for this task only through RGB images before this approach. This method detects the target object from the input image and then performs pose estimation by predicting the 3D bounding box projection in conjunction with PnP, and finds that accurate bounding box size prediction is critical for category-level tasks. It has notable potential for robotics applications.

Tables 4 and 5 show performance comparison results of the category-level position estimation methods on the NOCS dataset [116].

Table 4. Performance of category level methods on REAL275 dataset.

Methods	Years	Input	5°5 cm	10°5 cm	IoU50	IoU75
NOCS [116]	2019	RGB-D	10.0	25.2	78.0	30.1
SPD [121]	2020	RGB-D	21.4	54.1	77.3	53.2
6-PACK [128]	2020	RGB-D	33.3	-	-	-
DualPoseNet [119]	2021	RGB-D	35.9	66.8	79.8	62.2
FS-Net [17]	2021	RGB-D	28.2	60.8	92.2	63.5
ACR-Pose [122]	2021	RGB-D	36.9	65.9	82.8	66.0
SGPA [123]	2021	RGB-D	39.6	70.0	80.1	61.9
CAPTRA [129]	2021	D	62.16	-	-	-
DPDN [124]	2022	RGB-D	50.7	78.4	83.4	76.0
CATRE [130] + SPD	2022	RGB-D	54.4	73.1	-	43.6
CR-Net [131]	2021	RGB-D	34.3	47.2	79.3	55.9
RBP-Pose [132]	2022	RGB-D	48.1	79.2	-	-
SSP-Pose [133]	2022	RGB-D	44.6	77.8	82.3	66.3
GenPose [134]	2023	D	60.9	84.0	-	-

Table 5. Performance of category level methods on CAMERA25 dataset.

Methods	Years	Input	5°5 cm	10°5 cm	IoU50	IoU75
NOCS [116]	2019	RGB-D	40.9	64.6	83.9	69.5
SPD [121]	2020	RGB-D	59.0	81.5	93.2	83.1
DualPoseNet [119]	2021	RGB-D	70.7	84.7	92.4	86.4
ACR-Pose [122]	2021	RGB-D	74.1	87.8	93.8	89.9
SGPA [123]	2021	RGB-D	74.5	88.4	93.2	88.1
CATRE [130] + SPD	2022	RGB-D	80.3	89.3	-	76.1
CR-Net [131]	2021	RGB-D	76.4	87.7	93.8	88.0
RBP-Pose [132]	2022	RGB-D	79.6	89.5	93.1	89.0
SSP-Pose [133]	2022	RGB-D	75.5	87.4	-	86.8
GenPose [134]	2023	D	84.4	89.6	-	-

4. Datasets and Metrics

4.1. Datasets

Deep-learning-based methods greatly benefit from access to extensive and high-quality training data. In this section, we provide an overview of some of the most commonly used and representative datasets in object position estimation tasks, categorize them according to whether they belong to the instance level or the category level, and describe the application scenarios for which each dataset is suitable. The BOP Challenge [35] serves as a pivotal initiative that organizes multiple 6DoF pose estimation datasets into a standardized format. It is also classified according to whether the dataset belongs to the instance level or the category level, and the application scenarios suitable for each dataset are described. This unification not only simplifies the evaluation of various pose estimation methods, but also fosters significant advancements in the development of deep learning-based pose estimation techniques. The part of the used datasets shown in this paper comes from the BOP website at https://bop.felk.cvut.cz/datasets/ (accessed on 1 February 2024). For a detailed comparison of these datasets and their applicable scenarios, please refer to Table 6. The sample presentations of some of the datasets are shown in Figure 14.

LineMOD (LM) [78], introduced by Hinterstoisser et al. at the 2012 ACCV conference, stands as one of the most widely utilized datasets in the field of 6DoF pose estimation tasks. It also plays a role in object detection tasks. The dataset contains 15 categories of daily objects and comprises more than 18,000 real images, each accompanied by finely labeled poses. LineMOD is suitable for pose estimation in cluttered scenes with minor occlusions.

Figure 14. Examples of some datasets. (**Left**) Examples of LM, LM-O, T-LESS, and ITODD datasets from BOP. (**Right**) Examples of YCB-V, TUD-L, NOCS and Objectron datasets from BOP.

Table 6. Comparison of object pose estimation datasets.

Dataset	Years	Levels	Categories	Suitable Scenes
LM [78]	2012	Instance-Level	15	Objects are cluttered and untextured with limited viewpoints.
LM-O [79]	2014	Instance-Level	8	Objects are cluttered and more severely occluded.
Shapenet [135]	2016	Category-Level	16	Point cloud dataset of common objects in life with fine segmentation.
T-LESS [12]	2017	Instance-Level	30	Industry-related scenes with few object textures, strong symmetry, and mutual occlusion.
ITODD [13]	2017	Instance-Level	28	Industrial scenes with strong and scarce color information in the case of random projections.
Siléane [136]	2017	Instance-Level	8	Different symmetry objects.
YCB-V [20]	2018	Instance-Level	21	Daily objects with occlusion in different light situations, and applicable to the video needs of the object.
TUD-L/TYO-L [35]	2018	Instance-Level	24	Different light conditions.
NOCS [116]	2019	Category-Level	6	Category-level position of common objects, meet real and synthetic dataset requirements.
Fraunhofer [137]	2019	Instance-Level	10	Industrial large-scale dataset, including different modalities, is suitable for grasping tasks.
Objectron [138]	2021	Category-Level	9	Meeting generalizability and tracking task requirements with large-scale multiple views.

Occlusion LineMOD (LM-O) [79], introduced by Brachmann et al. at ECCV 2014, has been proposed to meet the requirements of severely occluded scenes. This dataset extends a test set from LineMOD (LM) and involves photographing objects under three different lighting conditions, introducing significant occlusion across eight object categories.

T-LESS [12], introduced by Hodan et al. at the WACV conference in 2017, is designed for the challenging task of 6DoF pose estimation, particularly focusing on textureless objects. This dataset contains 30 industrially relevant objects that lack apparent texture and color information. These objects share similarities in shape and size, and exhibit symmetry, which pose significant occlusion challenges when multiple objects are combined. T-LESS includes 20 scenes of varying complexity and provides texture-free CAD models for each object. Overall, T-LESS is a very challenging dataset in the 6DoF pose estimation task.

YCB-V [20], introduced in the context of PoseCNN, represents an extension of the YCB dataset. This dataset comprises 21 objects characterized by adjusting the shapes and

textures, and it is derived from 92 videos that encompass scenarios with occlusion and different symmetries. These variations are influenced by image noise and diverse lighting conditions. YCB-V includes a combination of both real and synthesized images, making it suitable for application scenarios that involve daily scenes and 6DoF pose estimation based on video sequences.

ITODD [13], introduced by Drost et al. at ICCVW 2017, is a dataset that contains 28 objects photographed in real industrial environments. This dataset specifically focuses on mechanical parts within industrial settings, where color information is often limited. Notably, each scene in ITODD is captured using two industrial sensors and three grayscale cameras, which results in a high-quality dataset that offers valuable 3D industrial scenes.

TUD-L and TYO-L [35], both introduced in the 2018 BOP Challenge [35], offer distinctive pose estimation datasets designed for various environments and illumination conditions. TUD-L comprises datasets featuring three moving objects, subjecting these objects to eight distinct illumination conditions in the test images. Meanwhile, TYO-L includes 21 objects captured on four different tablecloths and under five diverse illumination conditions. A notable feature of these datasets is their applicability to scenes with varying lighting conditions encountered in daily life.

NOCS [116], which stands for Normalized Object Coordinate Space, was proposed by He Wang et al. in 2019 for Category-Level 6D Object Pose and Size Estimation. It comprises six object categories, including bottle, bowl, camera, can, laptop, and mug, along with a distractor category. The NOCS dataset contains 31 indoor scenes and is divided into two sub-datasets: the real dataset REAL25 and the virtual dataset CAMERA275. Notably, a significant portion of current category-level pose estimation research relies on this dataset.

Siléane dataset [14], introduced in 2017, provides RGB-D images alongside corresponding semantic segmentation labels. It serves as a small yet finely labeled semantic segmentation dataset focused on outdoor scenes, and is provided with different symmetries in eight object categories.

The Fraunhofer IPA Bin-Picking dataset [137], introduced in 2019, contains 10 categories of objects and includes depth maps, point clouds, and RGB maps. This dataset offers large-scale data designed for complex industrial scenarios and multiple parts for industrial grasping. It extends the scope of the Siléane dataset [136] to cover more diverse scenarios suitable for deep learning. Additionally, it introduces two new industrial object categories.

Shapenet [135], proposed in 2016, is a point cloud dataset that comprises 16 large categories and 55 small categories commonly found in daily life. Each large category includes lots of model data, with multiple models corresponding to each category. Shapenet provides various semantic annotations for each model, which supports the segmentation of different instances of parts and is widely utilized in a range of visual tasks based on point clouds.

Objectron [138], proposed in CVPR 2021 by Ahmadyan et al. Contains nine categories of objects with 4 million labeled images from 14,819 videos. The dataset is designed for category-level pose estimation, with each category consisting of hundreds of examples captured under different lighting conditions. Significantly, these videos showcase stationary objects from various perspectives, consistently providing bounding boxes, also making them well-suited for tracking tasks. The data collection took place in the wild environment, enhancing its real-world generalizability.

4.2. Metrics

Different algorithms can be evaluated more fairly under the same evaluation metrics, and the following evaluation metrics are commonly used in 6DoF pose estimation:

ADD (Average Distance of Model Points): ADD measures whether the average deviation of the transformed model points is less than a certain value of the diameter of the object. The commonly used index value is ADD-0.1d, and it is considered that the estimation is correct when the distance is less than 10% of the size of the model diameter.

2D Projection Metric: This metric calculates the average distance between the projections of the 3D model points given the estimated pose and the ground truth pose. If the distance between projections is less than five pixels, the pose is considered correct. Note that a CAD model of the target object needs to be known for this metric.

n○n cm Metric: The effectiveness of the pose estimation is tested by measuring the rotation angle and translation distance errors. A common metric is 5°5 cm, meaning that the pose estimation is considered correct if the absolute value of the error of each rotation angle does not exceed 5°, and the absolute value of the error of the translation position from the real data does not exceed 5 cm. Additionally, 5°10 cm and 10°10 cm are commonly used as numerical settings. For symmetric objects, the ADD (-S) metric is used, i.e., the distance to the closest point is used instead of the average distance calculation.

VSD (Visible Surface Difference): VSD considers only the visible object part and treats indistinguishable poses as equivalent. It is applicable for symmetric and occlusion cases. The higher the overlap between the estimated pose and the true value in the visible region, the lower the error. This metric is often used in research work on the T-LESS dataset.

MSSD (Maximum Symmetric Surface Distance): MSSD measures the maximum distance deviation of a surface point measured in 3D space. It is relevant for robotics applications. The smaller the maximum distance deviation between the surface of the 3D object in the estimated pose and the surface points in the true value pose, the smaller the error.

MSPD (Maximum Symmetric Projection Distance): MSPD measures the maximum deviation perceivable on the image plane and is relevant for augmented reality applications. It calculates the maximum deviation of the 2D profile of an object on the image plane. It is similar to MSSD, but calculates 2D projection. The smaller the maximum deviation of the 2D contour in the estimated pose from the true pose contour, the smaller the error.

AR (Average Recall): In the BOP Challenge [35], the pose error is measured by the average of three error functions: Visible Surface Difference (VSD), Maximum Symmetric Surface Distance (MSSD), and Maximum Symmetric Projection Distance (MSPD).

5. Analysis and Possible Future Directions

5.1. Analysis of Task

Convolutional neural networks, are good at establishing mapping relationships between 2D images to 3D objects, and methods utilizing deep learning of the networks have been shown to significantly improve pose estimation accuracy and robustness [139–141]. It can be known from the results of Tables 1 and 3 that, in general, the same method framework performs better in terms of accuracy when based on RGB-D, but it has an efficiency advantage when based on RGB images only. As shown in Table 7, based on the previous content, we summarize the applicable scenarios and limitations of different algorithms.

Estimating rotations is significantly more complex and difficult than estimating translations. Common rotation representations, such as rotation matrix, quaternion, and Euler angle, are usually discontinuous in 3D Euclidean space, which is very challenging for neural network training. Sundermeyer et al. [75] utilize autoencoders to learn implicit 3D orientation features directly from images, embedding rotation information in the latent representation. Zhou et al. [142] introduce a continuous representation definition that is of great benefit to deep-learning-based methods.

According to the characteristics of different data sets in Table 6, in industrial applications, metal parts often lack color information, the surface shows different reflections under different lighting conditions, and the objects are mostly symmetric, with smooth surfaces and less texture information. In these situations, methods based on RGB images may not be so effective, the use of RGB-D or point cloud data needs to be considered. Symmetry is also a widespread problem in pose estimation tasks. To handle it, methods such as defining a suitable loss function [20,63], utilizing geometric constraints and transformations designed for rotationally symmetric objects [60], or employing multi-view fusion [34] can be considered.

Table 7. Comparison of pose estimation algorithms based on deep learning.

Methods	Level	Advantages or Applicable Scenarios	Limitation
Regression-based methods	Instance-level	Simple design and wide application.	Applicability to complex environments may be limited.
Feature-based methods	Instance-level	Situations with rich features and not severe occlusion.	Symmetry needs to be considered.
Fusion-based methods	Instance-level	Industrial applications, are suitable for occlusion.	The method design is relatively complex.
Point cloud-based methods	Instance-level	Robot grabbing-related tasks.	Surface reflections may result in poorer results.
Regression-based methods	Category-level	Everyday objects, perform better in generalization.	Poor handling of intra-category differences.
Prior-based methods	Category-level	More robust to intra-class differences and color changes.	High demand for computing resources.

While in daily life scenarios, CAD models or depth maps are not readily available for many objects, color images are very easy to capture. Here, it is possible to consider using only RGB images as the input or a category-level method. Compared with instance-based methods, category-level pose estimation methods have a better generalization ability to different shapes when category information is known. From the information in Tables 4 and 5, we can know that there is still much room for improvement in category-level pose estimation.

When considering training samples, tools for labeling object poses are provided by Label Fusion [143], among others. However, labeling 6D poses in real images is both expensive and unavoidably subject to a not insignificant percentage of errors [16,61]. In contrast, synthetic images are advantageous due to their low time cost and storage efficiency. Some approaches [32,38] utilize synthetic images for training or explore self-supervised learning methods to deal with the problem that labeling real images is difficult. However, the drawbacks of synthetic-to-real must not be overlooked, and the domain gap generated by only training on synthetic images could affect their use in real-world scenarios. To address this, referencing techniques used in the BOP Challenge [144], which generate synthetic images through physically based rendering methods, can help minimize the domain gap between synthetic and real images. Also, the treatment of DPDN [124] in category-level pose estimation can be referenced.

The refinement method has proven very effective for initially rough poses. There are already some methods that improve accuracy through refinement, the usage of PoseCNN [20] results in DeepIM [33] and the combination of Repose [74] and PVNet [46] are successful examples. However, while refinement enhances accuracy, it also entails some efficiency trade-offs for the method.

To meet the real-time requirements of the task, an end-to-end pose estimation method based on sparse feature matching can be considered, or a method combined with detection results can be designed. Generally, sparse feature matching methods are faster than dense matching. Compared with segmentation-based methods, object detection-based methods can better meet the speed requirements of processing before pose estimation. But time efficiency is not the only goal, multi-stage approaches may be more time-consuming than end-to-end approaches, but each module can be optimized independently to improve accuracy and are easier to modularize for different specific tasks.

5.2. Challenges and Possible Future Directions

In recent years, driven by the rapid development of computer vision, deep-learning-based pose estimation methods have made significant progress. However, these methods still face various challenges and a lot of research space exists.

One of the widespread challenges in various application scenarios is how to achieve accurate pose estimation under low texture, severe occlusion, cluttered background, or changing lighting conditions. To solve these problems, previous methods have made many efforts. For example, using RGB-D images may address the shortcomings of only using RGB or point clouds. Or incorporating geometric constraints and domain-specific knowledge into the network architecture and loss function design to enhance the model's ability to utilize prior information. But there is still room for improvement.

Another challenge is the need for methods based on small samples, or even zero samples, to estimate the pose of new objects, and how to improve the generalization performance of the method. Achieving accurate pose estimation for specific object instances often demands extensive training data, limiting generalization capability. Therefore, a significant trend in this field is the development of approaches that address small sample challenges. One promising approach is category-level pose estimation, which does not rely on specific objects, can be trained with a smaller sample size, and offers robust generalization to unseen instances. Data augmentation techniques, such as symmetry transformations and illumination variations, contribute to enhanced model robustness against pose and appearance variations. Bridging the gap between synthetic and real images is an ongoing challenge in this context, and Blenderproc's method [144] offers a valuable reference.

The third possible direction is that when facing tasks with high accuracy or speed requirements, integration with other advanced knowledge may be required. There are an increasing number of methods that use object detection [145] or segmentation methods [146] in the initial stage of pose estimation. For example, SAM [146] can be effectively utilized during the training process of POPE [147]. Convolutional Neural Networks (CNNs) excel at capturing local information, while transformer architectures [148] are adept at handling global information. Although transformers have gained prominence in tasks like human body pose estimation, their adoption in 6DoF pose estimation is relatively limited. Transformer has shown excellent performance in many areas, which means that combining CNN and Transformer may lead to better performance. While CNNs currently dominate the landscape of network architectures, recent efforts [96,107,149,150] have explored the integration of attention mechanism modules, yielding promising results. Transformers have also been used by several methods [151–153] to address structural irregularities present in point cloud data.

The fourth is to use multi-view information. Important factors such as physical and semantic information in the same scene are shared in different views, and better representation can be obtained by utilizing multiple views [154]. In the field of deep-learning-based pose estimation, the limitations of a single perspective are gradually emerging. The utilization of multiple viewpoints is poised to be an important direction for the future of pose estimation. These multiple viewpoints offer a more comprehensive representation of the target object, effectively alleviating visual ambiguities, and multi-view data are easily accessible in tasks such as industrial object manipulation. [155] aggregats 2D–3D Distributions of Multiple Views for Initial Position Estimation and Refinement One promising avenue involves the learning of optimal observation viewpoints. For instance, Gen6D [156] initially extracts target region features through a dedicated target detector and subsequently employs a viewpoint selection module to match these features, pinpointing the reference viewpoint most akin to the target viewpoint. This process simplifies regression challenges. Alternatively, different views can be harnessed to iteratively optimize pose estimation. By initially conducting coarse pose estimation using global features and then fine-tuning it through the alignment of detailed features from various viewpoints, a step-wise refinement of the pose is realized. This progression, from coarse to fine, has practical applications in tasks such as robot grasping, where multi-view data can inform the selection of optimal grasping positions, enhancing overall operational efficiency [157].

The final possibility is to propose new datasets to meet the demands of the task for changes in usage scenarios. Existing datasets, as shown in Table 6, while widely used, may have limitations in terms of scenario diversity and data types. To cater to the varied

demands of applications across different scenarios, future datasets with richer scenarios and diverse data types are expected to emerge. Such datasets will play a key role in advancing the development of a unified framework for cross-modal pose estimation, accommodating a wide range of usage requirements.

5.3. Conclusions

In this study, we categorize the 6DoF pose estimation methods into two groups: instance-level and category-level. We analyze the applicable scenarios of different methods in each category, and also provide method recommendations based on the challenges faced by different application scenarios. Although the 6DoF method has developed rapidly in recent years, there is still a lot of room for in-depth research. This article also provides simple suggestions for possible future research directions. In the future, we would like to extend this work to the video field, as well as real-time pose estimation and robot grabbing.

Author Contributions: Conceptualization, J.G. and Y.H.; methodology, J.G.; formal analysis, J.G.; investigation, J.G.; resources, Y.H. and Q.W.; writing—original draft preparation, J.G.; writing—review and editing, Y.H. and S.L.; supervision, Q.W. and Y.F. All authors have read and agreed to the published version of the manuscript

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

References

1. Liu, X.; Bai, H.; Song, G.; Zhao, Y.; Han, J. Augmented reality system training for minimally invasive spine surgery. In Proceedings of the 2017 IEEE International Conference on Robotics and Biomimetics (ROBIO), Macau, China, 5–8 December 2017; pp. 1200–1205.
2. Kalia, M.; Navab, N.; Salcudean, T. A real-time interactive augmented reality depth estimation technique for surgical robotics. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 8291–8297.
3. Arnold, E.; Al-Jarrah, O.Y.; Dianati, M.; Fallah, S.; Oxtoby, D.; Mouzakitis, A. A survey on 3D object detection methods for autonomous driving applications. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 3782–3795. [CrossRef]
4. Fan, Z.; Zhu, Y.; He, Y.; Sun, Q.; Liu, H.; He, J. Deep learning on monocular object pose detection and tracking: A comprehensive overview. *ACM Comput. Surv.* **2022**, *55*, 1–40. [CrossRef]
5. Wada, K.; Sucar, E.; James, S.; Lenton, D.; Davison, A.J. Morefusion: Multi-object reasoning for 6D pose estimation from volumetric fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14540–14549.
6. Du, G.; Wang, K.; Lian, S.; Zhao, K. Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: A review. *Artif. Intell. Rev.* **2021**, *54*, 1677–1734. [CrossRef]
7. Pérez, L.; Rodríguez, Í.; Rodríguez, N.; Usamentiaga, R.; García, D.F. Robot guidance using machine vision techniques in industrial environments: A comparative review. *Sensors* **2016**, *16*, 335. [CrossRef] [PubMed]
8. Sun, J.; Wang, Z.; Zhang, S.; He, X.; Zhao, H.; Zhang, G.; Zhou, X. Onepose: One-shot object pose estimation without cad models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6825–6834.
9. Cui, Y.; Chen, X.; Zhang, Y.; Dong, J.; Wu, Q.; Zhu, F. Bow3D: Bag of words for real-time loop closing in 3D lidar slam. *IEEE Robot. Autom. Lett.* **2022**, *8*, 2828–2835. [CrossRef]
10. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-view 3D object detection network for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1907–1915.
11. Xu, D.; Anguelov, D.; Jain, A. Pointfusion: Deep sensor fusion for 3D bounding box estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 244–253.
12. Hodan, T.; Haluza, P.; Obdržálek, Š.; Matas, J.; Lourakis, M.; Zabulis, X. T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 880–888.

13. Drost, B.; Ulrich, M.; Bergmann, P.; Hartinger, P.; Steger, C. Introducing mvtec itodd-a dataset for 3D object recognition in industry. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 2200–2208.
14. Zhu, Y.; Li, M.; Yao, W.; Chen, C. A review of 6D object pose estimation. In Proceedings of the 2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 17–19 June 2022; Volume 10, pp. 1647–1655.
15. Wang, G.; Manhardt, F.; Tombari, F.; Ji, X. Gdr-net: Geometry-guided direct regression network for monocular 6D object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16611–16621.
16. Haugaard, R.L.; Buch, A.G. Surfemb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6749–6758.
17. Liu, C.; He, L.; Xiong, G.; Cao, Z.; Li, Z. Fs-net: A flow sequence network for encrypted traffic classification. In Proceedings of the IEEE INFOCOM 2019-IEEE Conference on Computer Communications, Paris, France, 29 April–2 May 2019; pp. 1171–1179.
18. Marullo, G.; Tanzi, L.; Piazzolla, P.; Vezzetti, E. 6D object position estimation from 2D images: A literature review. *Multimed. Tools Appl.* **2023**, *82*, 24605–24643. [CrossRef]
19. Kendall, A.; Grimes, M.; Cipolla, R. Posenet: A convolutional network for real-time 6-dof camera relocalization. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2938–2946.
20. 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.
21. Besl, P.; McKay, N.D. A method for registration of 3-D shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **1992**, *14*, 239–256. [CrossRef]
22. 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.
23. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
24. Hu, Y.; Hugonot, J.; Fua, P.; Salzmann, M. Segmentation-driven 6D object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3385–3394.
25. Tekin, B.; Sinha, S.N.; Fua, P. Real-time seamless single shot 6D object pose prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 292–301.
26. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
27. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
28. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
29. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
30. Redmon, J.; Farhadi, A. YOLO9000: better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
31. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* **2021**, *65*, 99–106. [CrossRef]
32. Yen-Chen, L.; Florence, P.; Barron, J.T.; Rodriguez, A.; Isola, P.; Lin, T.Y. inerf: Inverting neural radiance fields for pose estimation. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 1323–1330.
33. Li, Y.; Wang, G.; Ji, X.; Xiang, Y.; Fox, D. Deepim: Deep iterative matching for 6D pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 683–698.
34. Labbé, Y.; Carpentier, J.; Aubry, M.; Sivic, J. Cosypose: Consistent multi-view multi-object 6D pose estimation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 574–591.
35. Hodan, T.; Michel, F.; Brachmann, E.; Kehl, W.; GlentBuch, A.; Kraft, D.; Drost, B.; Vidal, J.; Ihrke, S.; Zabulis, X.; et al. Bop: Benchmark for 6D object pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 19–34.
36. Su, Y.; Saleh, M.; Fetzer, T.; Rambach, J.; Navab, N.; Busam, B.; Stricker, D.; Tombari, F. Zebrapose: Coarse to fine surface encoding for 6Dof object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6738–6748.
37. Hai, Y.; Song, R.; Li, J.; Ferstl, D.; Hu, Y. Pseudo Flow Consistency for Self-Supervised 6D Object Pose Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 14075–14085.
38. Wang, G.; Manhardt, F.; Liu, X.; Ji, X.; Tombari, F. Occlusion-aware self-supervised monocular 6D object pose estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [CrossRef]

39. Wang, G.; Manhardt, F.; Shao, J.; Ji, X.; Navab, N.; Tombari, F. Self6D: Self-supervised monocular 6D object pose estimation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part I 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 108–125.
40. Sock, J.; Garcia-Hernando, G.; Armagan, A.; Kim, T.K. Introducing pose consistency and warp-alignment for self-supervised 6D object pose estimation in color images. In Proceedings of the 2020 International Conference on 3D Vision (3DV), Fukuoka, Japan, 25–28 November 2020; pp. 291–300.
41. Bukschat, Y.; Vetter, M. EfficientPose: An efficient, accurate and scalable end-to-end 6D multi object pose estimation approach. *arXiv* **2020**, arXiv:2011.04307.
42. Sundermeyer, M.; Hodaň, T.; Labbe, Y.; Wang, G.; Brachmann, E.; Drost, B.; Rother, C.; Matas, J. Bop challenge 2022 on detection, segmentation and pose estimation of specific rigid objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2784–2793.
43. 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; pp. 1521–1529.
44. Park, K.; Mousavian, A.; Xiang, Y.; Fox, D. Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10710–10719.
45. Zakharov, S.; Shugurov, I.; Ilic, S. Dpod: 6D Pose object detector and refiner. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1941–1950.
46. Peng, S.; Liu, Y.; Huang, Q.; Zhou, X.; Bao, H. Pvnet: Pixel-wise voting network for 6Dof pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4561–4570.
47. Deng, X.; Mousavian, A.; Xiang, Y.; Xia, F.; Bretl, T.; Fox, D. PoseRBPF: A Rao–Blackwellized particle filter for 6-D object pose tracking. *IEEE Trans. Robot.* **2021**, *37*, 1328–1342. [CrossRef]
48. Shugurov, I.; Li, F.; Busam, B.; Ilic, S. Osop: A multi-stage one shot object pose estimation framework. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6835–6844.
49. Nguyen, V.N.; Hu, Y.; Xiao, Y.; Salzmann, M.; Lepetit, V. Templates for 3D object pose estimation revisited: Generalization to new objects and robustness to occlusions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6771–6780.
50. Balntas, V.; Doumanoglou, A.; Sahin, C.; Sock, J.; Kouskouridas, R.; Kim, T.K. Pose guided RGBD feature learning for 3D object pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3856–3864.
51. Rusu, R.B.; Bradski, G.; Thibaux, R.; Hsu, J. Fast 3D recognition and pose using the viewpoint feature histogram. In Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 18–22 October 2010; pp. 2155–2162.
52. Marton, Z.C.; Pangercic, D.; Blodow, N.; Beetz, M. Combined 2D–3D categorization and classification for multimodal perception systems. *Int. J. Robot. Res.* **2011**, *30*, 1378–1402. [CrossRef]
53. Ng, P.C.; Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **2003**, *31*, 3812–3814. [CrossRef] [PubMed]
54. Bay, H.; Tuytelaars, T.; Van Gool, L. Surf: Speeded up robust features. In Proceedings of the Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; Proceedings, Part I 9; Springer: Berlin/Heidelberg, Germany, 2006; pp. 404–417.
55. 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.
56. 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 1, pp. 886–893.
57. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
58. Lindenberger, P.; Sarlin, P.E.; Pollefeys, M. LightGlue: Local Feature Matching at Light Speed. *arXiv* **2023**, arXiv:2306.13643.
59. Pavlakos, G.; Zhou, X.; Chan, A.; Derpanis, K.G.; Daniilidis, K. 6-dof object pose from semantic keypoints. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May 2017–3 June 2017; pp. 2011–2018.
60. 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 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3828–3836.
61. Hu, Y.; Fua, P.; Salzmann, M. Perspective flow aggregation for data-limited 6D object pose estimation. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 89–106.
62. Hodan, T.; Barath, D.; Matas, J. Epos: Estimating 6D pose of objects with symmetries. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11703–11712.
63. Park, K.; Patten, T.; Vincze, M. Pix2pose: Pixel-wise coordinate regression of objects for 6D pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October 2019–2 November 2019; pp. 7668–7677.

64. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *arXiv* **2014**, arXiv:1406.2661.
65. Xu, Y.; Lin, K.Y.; Zhang, G.; Wang, X.; Li, H. Rnnpose: Recurrent 6-dof object pose refinement with robust correspondence field estimation and pose optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14880–14890.
66. Elman, J.L. Finding structure in time. *Cogn. Sci.* **1990**, *14*, 179–211. [CrossRef]
67. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv* **2017**, arXiv:1710.10903.
68. Song, C.; Song, J.; Huang, Q. Hybridpose: 6D Object pose estimation under hybrid representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 431–440.
69. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
70. Castro, P.; Kim, T.K. Crt-6D: Fast 6D object pose estimation with cascaded refinement transformers. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 5746–5755.
71. Lepetit, V.; Moreno-Noguer, F.; Fua, P. EP n P: An accurate O (n) solution to the P n P problem. *Int. J. Comput. Vis.* **2009**, *81*, 155–166. [CrossRef]
72. Chen, H.; Tian, W.; Wang, P.; Wang, F.; Xiong, L.; Li, H. EPro-PnP: Generalized End-to-End Probabilistic Perspective-n-Points for Monocular Object Pose Estimation. *arXiv* **2023**, arXiv:2303.12787.
73. 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 IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7678–7687.
74. Iwase, S.; Liu, X.; Khirodkar, R.; Yokota, R.; Kitani, K.M. Repose: Fast 6D object pose refinement via deep texture rendering. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 3303–3312.
75. Sundermeyer, M.; Marton, Z.C.; Durner, M.; Brucker, M.; Triebel, R. Implicit 3D orientation learning for 6D object detection from rgb images. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 699–715.
76. Manhardt, F.; Kehl, W.; Navab, N.; Tombari, F. Deep model-based 6D pose refinement in rgb. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 800–815.
77. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
78. Hinterstoisser, S.; Lepetit, V.; Ilic, S.; Holzer, S.; Bradski, G.; Konolige, K.; Navab, N. Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes. In Proceedings of the Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Republic of Korea, 5–9 November 2012; Revised Selected Papers, Part I 11; Springer: Berlin/Heidelberg, Germany, 2013; pp. 548–562.
79. Brachmann, E.; Krull, A.; Michel, F.; Gumhold, S.; Shotton, J.; Rother, C. Learning 6D object pose estimation using 3D object coordinates. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part II 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 536–551.
80. Di, Y.; Manhardt, F.; Wang, G.; Ji, X.; Navab, N.; Tombari, F. So-pose: Exploiting self-occlusion for direct 6D pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 12396–12405.
81. Shugurov, I.; Zakharov, S.; Ilic, S. Dpodv2: Dense correspondence-based 6 dof pose estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7417–7435. [CrossRef] [PubMed]
82. 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.
83. Masoumian, A.; Rashwan, H.A.; Cristiano, J.; Asif, M.S.; Puig, D. Monocular depth estimation using deep learning: A review. *Sensors* **2022**, *22*, 5353. [CrossRef] [PubMed]
84. Ding, Z.; Sun, Y.; Xu, S.; Pan, Y.; Peng, Y.; Mao, Z. Recent Advances and Perspectives in Deep Learning Techniques for 3D Point Cloud Data Processing. *Robotics* **2023**, *12*, 100. [CrossRef]
85. Gao, G.; Lauri, M.; Hu, X.; Zhang, J.; Frintrop, S. Cloudaae: Learning 6D object pose regression with on-line data synthesis on point clouds. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 11081–11087.
86. Drost, B.; Ulrich, M.; Navab, N.; Ilic, S. Model globally, match locally: Efficient and robust 3D object recognition. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 998–1005.
87. Choi, C.; Christensen, H.I. RGB-D object pose estimation in unstructured environments. *Robot. Auton. Syst.* **2016**, *75*, 595–613. [CrossRef]

88. Liu, C.; Chen, F.; Deng, L.; Yi, R.; Zheng, L.; Zhu, C.; Wang, J.; Xu, K. 6DOF Pose Estimation of a 3D Rigid Object based on Edge-enhanced Point Pair Features. *arXiv* **2022**, arXiv:2209.08266.
89. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3D classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
90. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv* **2017**, arXiv:1706.02413.
91. Hagelskjær, F.; Buch, A.G. Pointvotenet: Accurate object detection and 6 dof pose estimation in point clouds. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 2641–2645.
92. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; Markham, A. Randla-net: Efficient semantic segmentation of large-scale point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11108–11117.
93. 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 IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass, CO, USA, 1–5 March 2020; pp. 2824–2833.
94. Deng, H.; Birdal, T.; Ilic, S. Ppfnet: Global context aware local features for robust 3D point matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 195–205.
95. Hoang, D.C.; Stork, J.A.; Stoyanov, T. Voting and attention-based pose relation learning for object pose estimation from 3D point clouds. *IEEE Robot. Autom. Lett.* **2022**, *7*, 8980–8987. [CrossRef]
96. Li, Z.; Stamos, I. Depth-based 6DoF Object Pose Estimation using Swin Transformer. *arXiv* **2023**, arXiv:2303.02133.
97. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
98. Cai, D.; Heikkilä, J.; Rahtu, E. Ove6D: Object viewpoint encoding for depth-based 6D object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6803–6813.
99. Gao, G.; Lauri, M.; Wang, Y.; Hu, X.; Zhang, J.; Frintrop, S. 6D object pose regression via supervised learning on point clouds. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 3643–3649.
100. Xiang, Y.; Choi, W.; Lin, Y.; Savarese, S. Data-driven 3D voxel patterns for object category recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1903–1911.
101. Kehl, W.; Milletari, F.; Tombari, F.; Ilic, S.; Navab, N. Deep learning of local rgb-d patches for 3D object detection and 6D pose estimation. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part III 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 205–220.
102. Li, C.; Bai, J.; Hager, G.D. A unified framework for multi-view multi-class object pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 254–269.
103. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3343–3352.
104. He, Y.; Huang, H.; Fan, H.; Chen, Q.; Sun, J. Ffb6D: A full flow bidirectional fusion network for 6D pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3003–3013.
105. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11632–11641.
106. Lin, S.; Wang, Z.; Ling, Y.; Tao, Y.; Yang, C. E2EK: End-to-end regression network based on keypoint for 6D pose estimation. *IEEE Robot. Autom. Lett.* **2022**, *7*, 6526–6533. [CrossRef]
107. Zhou, J.; Chen, K.; Xu, L.; Dou, Q.; Qin, J. Deep Fusion Transformer Network with Weighted Vector-Wise Keypoints Voting for Robust 6D Object Pose Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 13967–13977.
108. Jiang, X.; Li, D.; Chen, H.; Zheng, Y.; Zhao, R.; Wu, L. Uni6D: A unified cnn framework without projection breakdown for 6D pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11174–11184.
109. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4233–4242.
110. Shi, Y.; Huang, J.; Xu, X.; Zhang, Y.; Xu, K. Stablepose: Learning 6D object poses from geometrically stable patches. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15222–15231.
111. Labbé, Y.; Manuelli, L.; Mousavian, A.; Tyree, S.; Birchfield, S.; Tremblay, J.; Carpentier, J.; Aubry, M.; Fox, D.; Sivic, J. Megapose: 6D Pose estimation of novel objects via render & compare. *arXiv* **2022**, arXiv:2212.06870.

112. Lipson, L.; Teed, Z.; Goyal, A.; Deng, J. Coupled iterative refinement for 6D multi-object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6728–6737.
113. Hai, Y.; Song, R.; Li, J.; Salzmann, M.; Hu, Y. Rigidity-Aware Detection for 6D Object Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 8927–8936.
114. Zhou, G.; Wang, H.; Chen, J.; Huang, D. Pr-gcn: A deep graph convolutional network with point refinement for 6D pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 2793–2802.
115. Wu, Y.; Zand, M.; Etemad, A.; Greenspan, M. Vote from the center: 6 Dof pose estimation in rgb-d images by radial keypoint voting. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 335–352.
116. Wang, H.; Sridhar, S.; Huang, J.; Valentin, J.; Song, S.; Guibas, L.J. Normalized object coordinate space for category-level 6D object pose and size estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2642–2651.
117. Deng, X.; Geng, J.; Bretl, T.; Xiang, Y.; Fox, D. iCaps: Iterative category-level object pose and shape estimation. *IEEE Robot. Autom. Lett.* **2022**, *7*, 1784–1791. [CrossRef]
118. Lin, H.; Liu, Z.; Cheang, C.; Fu, Y.; Guo, G.; Xue, X. Sar-net: Shape alignment and recovery network for category-level 6D object pose and size estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6707–6717.
119. Lin, J.; Wei, Z.; Li, Z.; Xu, S.; Jia, K.; Li, Y. Dualposenet: Category-level 6D object pose and size estimation using dual pose network with refined learning of pose consistency. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 3560–3569.
120. Irshad, M.Z.; Kollar, T.; Laskey, M.; Stone, K.; Kira, Z. Centersnap: Single-shot multi-object 3D shape reconstruction and categorical 6D pose and size estimation. In Proceedings of the 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022; pp. 10632–10640.
121. Tian, M.; Ang, M.H.; Lee, G.H. Shape prior deformation for categorical 6D object pose and size estimation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 530–546.
122. Fan, Z.; Song, Z.; Xu, J.; Wang, Z.; Wu, K.; Liu, H.; He, J. ACR-Pose: Adversarial canonical representation reconstruction network for category level 6D object pose estimation. *arXiv* **2021**, arXiv:2111.10524.
123. Chen, K.; Dou, Q. Sgpa: Structure-guided prior adaptation for category-level 6D object pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 2773–2782.
124. Lin, J.; Wei, Z.; Ding, C.; Jia, K. Category-level 6D object pose and size estimation using self-supervised deep prior deformation networks. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 19–34.
125. Li, G.; Li, Y.; Ye, Z.; Zhang, Q.; Kong, T.; Cui, Z.; Zhang, G. Generative category-level shape and pose estimation with semantic primitives. In Proceedings of the Conference on Robot Learning, Auckland, New Zealand, 14–18 December 2022; pp. 1390–1400.
126. Chen, X.; Dong, Z.; Song, J.; Geiger, A.; Hilliges, O. Category level object pose estimation via neural analysis-by-synthesis. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXVI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 139–156.
127. Lin, Y.; Tremblay, J.; Tyree, S.; Vela, P.A.; Birchfield, S. Single-stage keypoint-based category-level object pose estimation from an RGB image. In Proceedings of the 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022; pp. 1547–1553.
128. Wang, C.; Martín-Martín, R.; Xu, D.; Lv, J.; Lu, C.; Fei-Fei, L.; Savarese, S.; Zhu, Y. 6-pack: Category-level 6D pose tracker with anchor-based keypoints. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 10059–10066.
129. Weng, Y.; Wang, H.; Zhou, Q.; Qin, Y.; Duan, Y.; Fan, Q.; Chen, B.; Su, H.; Guibas, L.J. Captra: Category-level pose tracking for rigid and articulated objects from point clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 13209–13218.
130. Liu, X.; Wang, G.; Li, Y.; Ji, X. Catre: Iterative point clouds alignment for category-level object pose refinement. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 499–516.
131. Wang, J.; Chen, K.; Dou, Q. Category-level 6D object pose estimation via cascaded relation and recurrent reconstruction networks. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 4807–4814.

132. Zhang, R.; Di, Y.; Lou, Z.; Manhardt, F.; Tombari, F.; Ji, X. RBP-Pose: Residual bounding box projection for category-level pose estimation. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 655–672.
133. Zhang, R.; Di, Y.; Manhardt, F.; Tombari, F.; Ji, X. SSP-Pose: Symmetry-Aware Shape Prior Deformation for Direct Category-Level Object Pose Estimation. In Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022; pp. 7452–7459.
134. Zhang, J.; Wu, M.; Dong, H. GenPose: Generative Category-level Object Pose Estimation via Diffusion Models. *arXiv* **2023**, arXiv:2306.10531.
135. Chang, A.X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. Shapenet: An information-rich 3D model repository. *arXiv* **2015**, arXiv:1512.03012.
136. Brégier, R.; Devernay, F.; Leyrit, L.; Crowley, J.L. Symmetry aware evaluation of 3D object detection and pose estimation in scenes of many parts in bulk. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 2209–2218.
137. Kleeberger, K.; Landgraf, C.; Huber, M.F. Large-scale 6D object pose estimation dataset for industrial bin-picking. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 2573–2578.
138. Ahmadyan, A.; Zhang, L.; Ablavatski, A.; Wei, J.; Grundmann, M. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7822–7831.
139. Hoque, S.; Arafat, M.Y.; Xu, S.; Maiti, A.; Wei, Y. A comprehensive review on 3D object detection and 6D pose estimation with deep learning. *IEEE Access* **2021**, *9*, 143746–143770. [CrossRef]
140. Sahin, C.; Garcia-Hernando, G.; Sock, J.; Kim, T.K. A review on object pose recovery: From 3D bounding box detectors to full 6D pose estimators. *Image Vis. Comput.* **2020**, *96*, 103898. [CrossRef]
141. Fu, M.; Zhou, W. DeepHMap++: Combined projection grouping and correspondence learning for full DoF pose estimation. *Sensors* **2019**, *19*, 1032. [CrossRef] [PubMed]
142. Zhou, Y.; Barnes, C.; Lu, J.; Yang, J.; Li, H. On the continuity of rotation representations in neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5745–5753.
143. Marion, P.; Florence, P.R.; Manuelli, L.; Tedrake, R. Label fusion: A pipeline for generating ground truth labels for real rgbd data of cluttered scenes. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018; pp. 3235–3242.
144. Denninger, M.; Sundermeyer, M.; Winkelbauer, D.; Zidan, Y.; Olefir, D.; Elbadrawy, M.; Lodhi, A.; Katam, H. Blenderproc. *arXiv* **2019**, arXiv:1911.01911.
145. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
146. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. *arXiv* **2023**, arXiv:2304.02643.
147. Fan, Z.; Pan, P.; Wang, P.; Jiang, Y.; Xu, D.; Jiang, H.; Wang, Z. POPE: 6-DoF Promptable Pose Estimation of Any Object, in Any Scene, with One Reference. *arXiv* **2023**, arXiv:2305.15727.
148. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
149. Amini, A.; Selvam Periyasamy, A.; Behnke, S. YOLOPose: Transformer-based multi-object 6D pose estimation using keypoint regression. In *Proceedings of the International Conference on Intelligent Autonomous Systems*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 392–406.
150. Zhang, Z.; Chen, W.; Zheng, L.; Leonardis, A.; Chang, H.J. Trans6D: Transformer-Based 6D Object Pose Estimation and Refinement. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 112–128.
151. Lee, J.; Lee, Y.; Kim, J.; Kosiorek, A.; Choi, S.; Teh, Y.W. Set transformer: A framework for attention-based permutation-invariant neural networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 3744–3753.
152. Zhao, H.; Jiang, L.; Jia, J.; Torr, P.H.; Koltun, V. Point transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 16259–16268.
153. Wu, X.; Lao, Y.; Jiang, L.; Liu, X.; Zhao, H. Point transformer v2: Grouped vector attention and partition-based pooling. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 33330–33342.
154. Tian, Y.; Krishnan, D.; Isola, P. Contrastive multiview coding. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 776–794.
155. Haugaard, R.L.; Iversen, T.M. Multi-view object pose estimation from correspondence distributions and epipolar geometry. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; pp. 1786–1792.

156. Liu, Y.; Wen, Y.; Peng, S.; Lin, C.; Long, X.; Komura, T.; Wang, W. Gen6D: Generalizable model-free 6-DoF object pose estimation from RGB images. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 298–315.
157. Mousavian, A.; Eppner, C.; Fox, D. 6-dof graspnet: Variational grasp generation for object manipulation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2901–2910.