Displays 89 (2025) 103069

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

Review

From visual understanding to 6D pose reconstruction: A cutting-edge review of deep learning-based object pose estimation☆

Jing Wang, Guohan Liu *, Wenxin Ding, Yuying Li, Wanying Song

Xi'an University of Science and Technology, 710054, Xi'an, China

ARTICLE INFO

Keywords:
Computer vision
Deep learning
Feature matching
Multi-modal fusion
Object pose estimation

ABSTRACT

Object pose estimation, as a key problem in computer vision, plays an important role in tasks such as autonomous driving and robot navigation. However, most of the existing reviews discuss both traditional and deep learning methods and fail to comprehensively define instance-level and category-level object pose estimation methods. To help researchers better understand this field, this paper summarizes instance-level, category-level, and unseen object and articulated body pose estimation methods in detail, filling the gap in the discussion of these emerging areas in existing reviews. Depending on the different modalities of the input data, the implementations, application domains, training paradigms, network architectures, and their strengths and weaknesses of the deep learning-based object position estimation methods are highlighted, and the performance of these methods on different datasets is compared. In addition, this paper comprehensively combs through the evaluation metrics and benchmark datasets in this field, deeply analyzes their application scope and applicability in different scenarios, and reveals the key roles of these metrics and datasets in promoting technological progress and solving practical problems. Facing the current technical bottlenecks, this paper also looks forward to the future development direction from the cutting-edge explorations of multi-view fusion, cross-modal data integration and novel neural networks, which provide brand new ideas and references to push forward the breakthrough progress in the field of object attitude estimation.

Contents

☆ This paper was recommended for publication by Prof. Guangtao Zhai.

* Corresponding author.
E-mail address: 23207223106@stu.xust.edu.cn (G. Liu).

https://doi.org/10.1016/j.displa.2025.103069
Received 21 October 2024; Received in revised form 22 February 2025; Accepted 28 April 2025
Available online 14 May 2025

1. Introduction

Object pose estimation, an important branch of computer vision, has attracted attention since the 1980s. The aim is to accurately estimate the 6DoF pose of an object, i.e., the object's 3D rotation and 3D translation information [1]. Early research focused on geometric methods and feature point matching, which rely on the correspondence between the image and the model to compute the object's pose. Later studies introduced approaches using predefined templates to match object poses [2] or estimating poses by utilizing 3D models and projection transformations [3]. In recent years, with the increase in computational power and the emergence of big data, deep learning-based methods for object pose estimation have gradually become dominant. These increasingly comprehensive theoretical knowledge, as well as state-of-the-art techniques, have led to a wide range of applications of object pose estimation in fields such as industrial automation [4], augmented reality [5], and human–computer interaction [6].

Estimating the pose of an object is a crucial aspect of augmented reality for determining and tracking the 3D position and orientation of objects in the real world [7]. Specifically, the AR system first captures visual and spatial information of objects in the real world through sensors in real-time. Then, the system applies object pose estimation techniques to analyze and understand the specific orientation of the object in space. This process allows the AR system to accurately align virtual objects with real-world objects. In addition, the system dynamically adjusts the object's pose information to ensure that it perfectly matches the object's position, thus enhancing the user's experience of the real-world environment.

In industrial automation, robots perform tasks with stringent requirements for precision and reliability, such as accurately grasping and placing parts or assembling parts in a set order and position. Object pose estimation technology enables robots to accurately recognize each part's spatial position and orientation of each part to perform delicate operations and improve productivity. In addition, the introduction of object pose estimation technology enables automated systems to perform more complex industrial tasks. This not only enhances the flexibility of the production line so that it can quickly adapt to fluctuations in production demand but also significantly reduces the reliance on manual operations in the production process, promoting the development of industrial automation to a higher level.

Traditional methods for object pose estimation correspond the extracted feature points to the feature points in the 3D model and the Perspective-n-Point (PnP) algorithm [8] to solve for the coordinates of the target under the camera coordinate system. Alternatively, they match the object contours in a 2D image with the 3D model to compute the object's pose. In addition, edge- and contour-based methods use edge detection algorithms such as Canny [9] or Sobel [10] to extract the edges in the image and then match the edges with the projections of the 3D model. The above three methods have significant advantages in dealing with local features, geometric transformations, etc. However, their high computational complexity and high dependence on the CAD model of the object make the traditional methods have certain limitations in real-time, complex environments and dynamic scenes. On the other hand, deep learning methods can generate stable feature descriptors by sensing the target features through multilayer networks, and the gradient propagation between networks enables the model to automatically learn the features in the image, thus improving the performance of object pose estimation. In addition, large-scale data training, various lightweight networks, and pruning and quantization techniques enhance the generalization ability of the model while achieving real-time estimation performance. Deep learning-based object pose estimation methods have been applied to various industries and have made significant progress in several fields, which will remain competitive in the future.

Based on whether researchers use the object's CAD model, we classify object pose estimation methods into instance-level object pose estimation and category-level object pose estimation. Category-level object pose estimation has made rapid progress after 2019. The advantage of these methods lies in their ability to estimate the pose of new objects without the need for object models while also handling structurally complex articulated bodies, such as robotic arms, demonstrating greater adaptability. Additionally, leveraging generative models and self-supervised learning techniques, these methods can reduce the reliance on large-scale labeled data, thereby improving data utilization efficiency. As a result, an increasing number of researchers are focusing on category-level, unseen objects, and articulated body pose estimation.

As discussed in many methods, the quality of the input data plays a crucial role in the model's training effectiveness and final performance. High-quality training data not only provides rich feature information but also effectively reduces noise and bias [11–14]. Especially in real-world application scenarios with complex backgrounds, occlusions, and varying object shapes, high-quality data will significantly enhance the model's performance and reliability. Therefore, this paper will also explore in detail the relationship between data quality and model performance and analyze how to optimize the input data through data preprocessing, data augmentation, and shape deformation techniques, thereby improving the accuracy of object pose estimation.

In conclusion, this paper will focus on relevant techniques in recent years, such as instance-level, category-level, and object pose estimation for unseen and articulated objects. Fig. 1 illustrates the structure of this paper to clearly present the classification and connections of these research directions.

The main contributions of this paper are as follows:

(1) Select high-impact computer vision papers focusing on deep learning-based object pose estimation methods from recent years. Clearly define the latest advancements in accuracy and the scope of application of these methods.

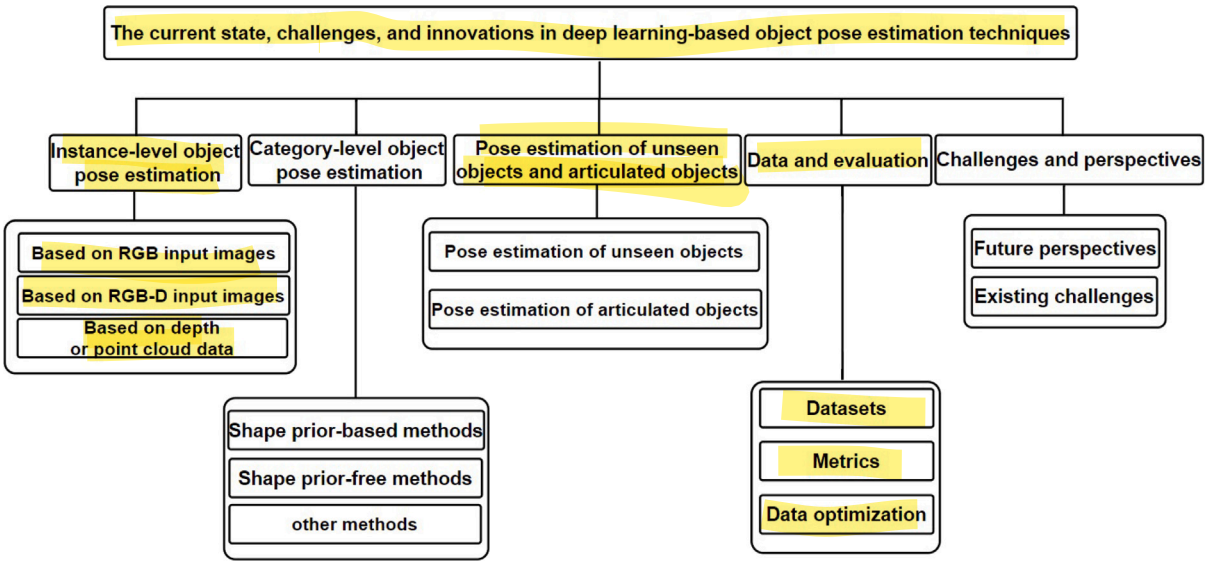

Fig. 1. Content structure diagram. This paper summarizes deep learning-based object pose estimation methods from the past five years, and introduces them from the perspectives of instance-level, category-level, unseen objects and articulated object pose estimation, datasets and evaluation metrics, as well as challenges and future directions. For instance-level methods, we categorize them based on data input, further dividing them into methods based on RGB images, RGB-D images, and depth or point cloud information. Category-level methods are categorized into methods based on shape priors and those without shape priors. The paper then proceeds to discuss methods for unseen and articulated objects. Next, this paper reviews the mainstream datasets, evaluation metrics, and training data optimization methods in the field of object pose estimation. Finally, based on the key challenges still existing in the field, the paper outlines potential future research directions.

(2) This paper synthesizes three main directions in object pose estimation: instance-level, category-level, and object pose estimation for unseen and articulated objects. Based on different input data types and implementation approaches, this paper elaborates on the basic principles and network structure design, as well as the advantages and limitations of each method. It provides an overview of the connections and progression between different methods.

(3) A comprehensive summary of the key datasets in the task of object pose estimation is presented, with an in-depth analysis of their characteristics, size, and application value, and the role of evaluation metrics in assessing the accuracy, robustness, and real-time performance of the algorithms is also explored. By comparing and analyzing performance across different datasets and evaluation metrics, this paper provides insights for readers on selecting evaluation criteria, helping them make more informed decisions for specific application scenarios.

(4) Finally, this paper addresses the existing challenges in this field and discusses the future development directions of object pose estimation from multiple perspectives, including network architecture, learning methods, and application areas.

2. Instance-level object pose estimation

Researchers have made significant technological breakthroughs in object pose estimation methods for specific instances. They utilize structures such as Convolutional Neural Networks (CNN) [15], Residual Networks (ResNet) [16], and Dynamic Graph Convolutional Networks (DGCNN) [17] to extract features from RGB images, RGB-D images or point cloud data, aiming to address the challenges of object pose estimation in various environments. This paper analyzes these methods based on different input data types, focusing on RGB images, RGB-D images, and depth or point cloud data, as shown in Table 1. It also provides a detailed summary of the implementation methods, advantages and disadvantages, and the problems these methods aim to solve.

2.1. RGB-based input image methods

RGB images contain rich color and texture information [18], which facilitates the network to extract the coarse and fine features of the image and enhance the network's perception of the object's pose. Much research is currently focused on developing deep processing and optimization algorithms for RGB images, aiming to recover more accurate pose information. We can systematically classify these methods into three categories: coordinate correspondence-based methods, feature-based methods, and template-based methods. Fig. 2 provides a chronological overview of RGB image-based methods for object pose estimation, showing the lineage of these methods as they have evolved.

2.1.1. Coordinate correspondence-based methods

Coordinate correspondence-based object pose estimation method maps between 2D image and 3D space (dense or sparse correspondence) by training a deep network to predict the corresponding 3D spatial locations of pixel points in the image. Then, the target pose is computed using the Random Sampling Consistency (RANSAC) and PnP algorithms [19]. Fig. 3 illustrates this process in detail.

The Pix2Pose method proposed in 2019 [20] effectively addresses the challenge of pose estimation for poorly textured objects [21] by establishing a mapping between coordinates. Specifically, Pix2Pose utilizes the codec structure to extract an image's coarse and delicate features. Meanwhile, the network uses jump connections during the decoding process to combine the outputs of the encoder and decoder to improve the prediction accuracy of geometric boundaries. In addition, Pix2Pose further improves the quality and accuracy of 3D coordinate images through GAN [22] training. The encoder–decoder structure, skip connections, and generative adversarial networks make Pix2Pose more robust in estimating object poses. PVNet [23] is a proposed method for the presence of occlusion or truncation of objects in an image. Unlike Pix2Pose, PVNet does not directly regress the coordinates of the key points but predicts pixel-level vectors pointing to the key points. Compared to traditional coordinate-based or heat map-based representations, vector field-based representations allow the network to focus more on local features and spatial relationships of objects, and to generalize better to different object poses and viewpoints. However, PVNet does not explicitly propose a method to deal with symmetric objects, so there are still some shortcomings in dealing with the pose ambiguity problem brought by symmetric objects. The HybridPose algorithm [24] consists of an intermediate feature prediction network and a pose regression network. Unlike the aforementioned methods,

Table 1
Instance-level object pose estimation methods. For each method, we introduce its implementation form, release year, training input, processing categories, and performance metrics on different datasets.

Methods	Data	Years	Type	Evaluation metrics(%)		
				LM	O-LM	YCB-Video
Pix2Pose	RGB	2019	Correspondence	72.4	32.0	–
PVNet	RGB	2019	Correspondence	86.27	40.77	73.4
DPOD	RGB	2019	Correspondence	95.2	47.3	–
HybridPose	RGB	2020	Correspondence	94.5	79.2	–
CosyPose	RGB	2020	Correspondence	–	–	84.5
EfficientPose	RGB	2021	Correspondence	97.35	83.98	–
NeRF-Pose	RGB	2022	Correspondence	95.1	49.2	–
SO-Pose	RGB	2023	Correspondence	96.0	62.3	56.8
[25]	RGB	2024	Correspondence	–	–	–
[26]	RGB	2024	Correspondence	88.4	71.7	55.2
CheckerPose	RGB	2023	Correspondence	97.1	77.5	81.4
CDPN	RGB	2019	Feature	89.86	–	–
TexPose	RGB	2022	Feature	91.7	66.7	–
[32]	RGB	2023	Feature	–	43.3	53.9
Crt-6D	RGB	2023	Feature	–	66.3	72.1
EPro-PnP	RGB	2023	Feature	95.80	–	–
[33]	RGB	2024	Feature	–	75.6	–
[34]	RGB	2024	Feature	–	65.8	68.1
ZeroPose	RGB	2023	Feature	–	76.90	80.5
PoseRBPF	RGB	2021	Template	79.76	–	–
OSOP	RGB	2022	Template	–	46.2	54.2
[35]	RGB	2023	Template	99.1	79.4	–
GigaPose	RGB	2024	Template	–	63.1	65.2
[41]	RGB	2024	Template	–	66.89	–
GS-Pose	RGB	2024	Template	97.3	80.54	–
ZS6D	RGB	2024	Template	–	52.7	49.9
DenseFusion	RGB-D	2019	Template	94	91.8	–
PVN3D	RGB-D	2020	Vote	99.4	91.8	–
KDFNet	RGB-D	2021	Vote	–	–	66.5
PR-GCN	RGB-D	2021	Vote	99.6	–	65.0
ES6D	RGB-D	2022	Vote	–	93.2	–
[48]	RGB-D	2023	Vote	99.8	96.7	77.7
[49]	RGB-D	2024	Vote	99.9	91.5	–
HiPose	RGB-D	2024	Vote	–	79.9	90.7
RDPN6D	RGB-D	2024	Vote	99.97	79.5	94.6
Self-6D	RGB-D	2020	Regression	58.9	–	–
GDR-Net	RGB-D	2021	Regression	93.7	84.4	62.2
[55]	RGB-D	2021	Regression	88.5	80.0	64.7
Unif6d	RGB-D	2022	Regression	–	88.8	–
6D-Diff	RGB-D	2024	Regression	–	83.8	79.6
[57]	RGB-D	2024	Regression	–	78.6	84.4
DCL-Net	RGB-D	2022	depth or point cloud	99.5	96.6	70.6
PointPoseNet	RGB-D	2023	depth or point cloud	98.4	93.2	79.5
SymFM6D	RGB-D	2023	depth or point cloud	–	94.1	–
[52]	RGB-D	2023	depth or point cloud	–	–	–
[53]	RGB-D	2024	depth or point cloud	69.0	52.0	–

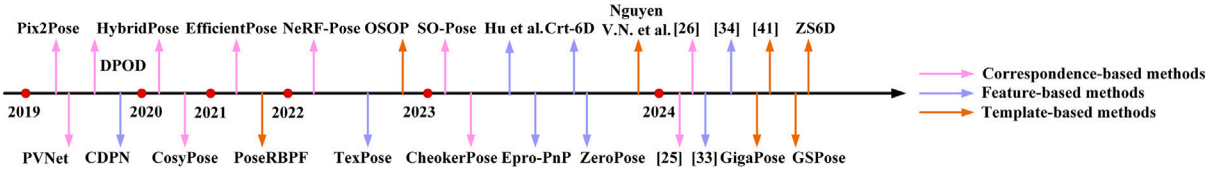

Fig. 2. Timeline of RGB image-based object pose estimation methods. Notably, the pink arrows, blue arrows, and orange arrows represent methods based on coordinate correspondence, feature-based methods, and template-based methods, respectively.

HybridPose not only uses key points as intermediate representations but also incorporates edge vectors and symmetry correspondences to capture richer geometric information from images. Additionally, it designs an optimization submodule that uses GM robust norms to enhance the robustness of pose estimation. However, in practical applications, this method requires generating key point and symmetry labels and providing segmentation templates and PVNet data.

NeRF-Pose [25] achieves optimal pose estimation by representing an object's 3D shape and color as an implicit function and optimizing this function using multi-view image [26] training. Due to the utilization of multi-view information, NeRF-Pose has high robustness in processing complex scenes. However, its volume rendering process requires a lot of computational resources and time. Unlike traditional methods, SO-Pose [27] enhances the representation of 3D objects by introducing self-obscuration information [28] and establishing a two-layer observer-centered representation structure. The first layer deals with the correspondence between visible points and their projections, and the second layer utilizes a self-attention mechanism [29] to generate a self-obscuring perceptual map to integrate the occlusion information. Additionally, SO-Pose introduces a cross-layer consistency loss term to align the self-obscuring, the correspondence field, and the 6D pose to improve the accuracy and robustness of the pose estimation. However, efficient computational performance also needs to be considered in practical applications to meet the demands of real-time

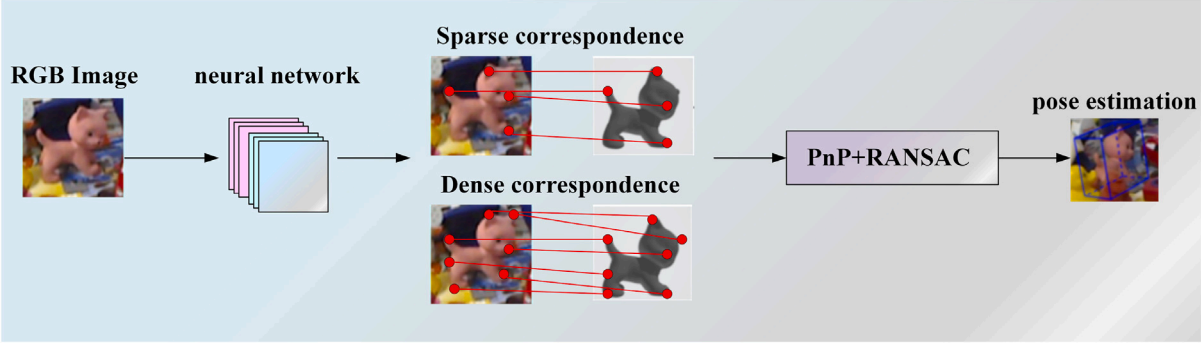

Fig. 3. Correspondence-based methods. This method primarily establishes a dense or sparse point correspondence between the input image and the object's CAD model, and then solves the pose using the PnP and RANSAC algorithms.

applications. EfficientPose [30] balances computational efficiency with performance by simultaneously scaling depth, width, and resolution. The network adds sub-networks to EfficientNet [31] to predict object rotation and translation to recover rotation and translation information, respectively. Although EfficientPose utilizes pose loss and feature loss to optimize the accuracy of pose estimation, its complex network structure requires careful tuning of parameters and training strategies to achieve optimal performance.

CosyPose [32] proposed a method that first reconstructs before regressing. It initially generates pose hypotheses from a single view and then matches these hypotheses across different images to jointly estimate the camera viewpoint and object pose. Additionally, CosyPose uses a shared encoder and two separate decoder networks to predict self-occlusion information and 2D–3D point correspondences, respectively. This method has significantly improved single-view and multi-view 6D pose estimation on the YCB-Video and T-LESS datasets. DPOD [33] focuses more on estimating and refining the pose of individual objects using deep learning to compute 6D poses through dense 2D–3D correspondence maps. To improve the generalization of the model, DPOD uses online data generation [34] and background enhancement techniques [35] and creates a bidirectional mapping model through correspondence texture mapping. Compared with traditional methods, DPOD improves the robustness and accuracy of pose estimation by creating dense point correspondences.

CheckerPose [36] improves the matching accuracy of corresponding points in 2D images by uniformly sampling critical points on the surface of 3D objects and using a graph neural network (GNN) [37] to model the interrelationships between these points. In addition, CheckerPose captures finer image features through the GNN network. Although CheckerPose is an efficient single-stage pose estimation method, further optimization regarding computational efficiency, hyperparameter tuning, and generalization capability is needed. [38] proposed a method called VAPO for the visibility of key points. The method focuses on localizing the visible key points in the input image to improve the 3D–2D point correspondence. Specifically, the VAPO method employs a personalized PageRank (PPR) algorithm to evaluate the perceived importance of visibility of key points and selects key points for localization accordingly. Then, a Graph Neural Network (GNN) is used to model the interactions between key points, and positional encoding [39] is applied to enhance the key point embeddings, further improving localization accuracy. This visibility-aware keypoint selection strategy enables VAPO to handle occlusions and viewpoint changes exceptionally well.

In order to accurately recover the object's positional information from a single RGB image, [40] designed the SymNet network. SymNet works by extracting the region of interest (RoI), generating full masks, visible masks, and the binary code of the object's surface (SymCode) as the intermediate variables, and then utilizes a correspondence-based pose regression (CPR) module to recover the 6D pose of the object directly. The method mainly targets the pose estimation problem of symmetric objects and effectively solves the ambiguity of pose estimation [41] due to symmetry by introducing symmetry-aware surface coding and one-to-many correspondences. In addition, SymNet does not rely on traditional PnP-RANSAC methods, giving it significant application potential.

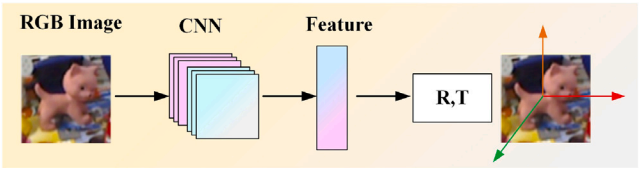

Fig. 4. Feature-based methods.

In summary, the correspondence-based instance-level object pose estimation method accurately estimates the pose of an object by establishing the correspondence between image feature points and 3D model points. The advantages of this method include high accuracy and robustness, especially when dealing with complex scenes and partial occlusions. However, dealing with many feature points and complex models requires substantial computational resources and time. Additionally, this method relies on precise feature point detection and matching, which demands high quality in feature point extraction.

2.1.2. *Feature-based methods*

Feature-based object pose estimation techniques are central to academic research and practical applications, especially for target objects with rich surface textures. This approach derives the spatial pose of an object by identifying key visual features in an image (e.g., edges, corner points, texture regions, feature descriptors, etc.) and establishing a correspondence with a 3D model or directly using network regression. Fig. 4 illustrates the implementation process of feature-based object pose estimation methods.

In object pose estimation tasks, separately handling rotation and translation information is a method to improve estimation accuracy and robustness. However, many earlier methods overlooked this, leading to poor performance in complex scenes. CDPN [42] introduces a decoupled method [43], which separately estimates the rotation and translation information in 6D pose estimation. It also employs dynamic scaling techniques to address the issue of object size variation in images. Additionally, it improves the accuracy of coordinate prediction through foreground region coordinate loss and global confidence loss. However, using larger models or running on resource-constrained devices may require significant computational resources. TexPose [44] implements an effective self-supervised learning strategy by combining texture learning and pose estimation with high-quality viewpoint synthesis using NeRF techniques [45]. In the texture learning phase, the network predicts the exact texture of an object even without precise

pose information. These generated texture data supervise the pose estimator to capture more accurate appearance information. Unlike TexPose, Crt-6D [46] proposed a rough-refinement-based pose estimation method. The method first estimates the initial rough pose of the object by a multilayer perceptron (MLP) [47] and then extracts features to generate keypoint features on the object surface. In order to improve the accuracy, Crt-6D introduces the Transformer module, which optimizes the pose offsets through self-attention and cross-attention mechanisms and gradually refines the pose estimation through cascade training. Although this cascade refinement structure can improve performance, it may adversely affect the final pose if the initial pose estimation is too coarse.

EPro-PnP [48] interprets the PnP output as a probability distribution of poses on SE(3) space [49] and thus deals with the traditional PnP problem. This approach allows EPro-PnP to deal with pose uncertainty and ambiguity, improving the model's ability to generalize to unknown objects. Also, the design of EPro-PnP allows it to be seamlessly integrated into existing network architectures, such as CDPN, to enhance performance. In addition, it provides a basis for developing entirely new network architectures. MRC-Net [50] learns the pose features of target objects through two stages: pose classification and pose residual prediction. Between these two stages, a multiscale residual correlation (MRC) layer captures the correspondences between the input and rendered images from the first stage, effectively applying the classification results to predict the residual pose accurately. MRC-Net performs superior on multiple BOP benchmark datasets, surpassing all RGB-based methods. Additionally, it does not require iterative post-processing, making it highly efficient. ZeroPose [51] proposed a pose estimation method based on surface encoding [52] and a hierarchical learning strategy [53]. It employs a layered binary encoding system to assign unique binary identifiers to surface vertices of the object, achieving dense feature representation. Hierarchical learning begins with coarse segmentation and gradually refines to more detailed levels. Compared to traditional dense correspondence methods, ZeroPose's encoding approach allows direct pixel-to-surface matching, simplifying the matching process. However, real-time applications might require further optimization to meet speed requirements. [54] presents a single-stage [55] pose estimation method. The method includes a local feature extraction module, a feature aggregation module, and a global inference module, which extracts local features by sharing network parameters, performs feature aggregation, and estimates the final 6D pose through a fully connected layer. The end-to-end training strategy eliminates the iterative RANSAC process in traditional two-stage methods, improving computational speed. However, single stage production relies on 3D–2D correspondence, and if the quality of the correspondence is poor, the network's performance will be greatly reduced. Instance-level object pose estimation methods rely heavily on selecting key points, and an accurate set of key points can effectively improve model performance. KeyGNet [56], based on graph convolutional networks, optimizes the location and distribution of key points by processing non-obscuring visible surface point clouds while combining the distributional similarity loss and dispersion loss functions of the Wasserstein distance [57] to optimize the location and distribution of key points. Experimental results show that KeyGNet can effectively reduce the performance gap from single-object to multi-object training scenarios, demonstrating its effectiveness and good generalization ability in 6DoF pose estimation tasks.

In summary, feature-based instance-level object pose estimation methods extract local or global features from images and match them with pre-stored object model features to estimate the object's pose. This approach is more suitable for objects with distinct features. However, its performance may significantly degrade in cases of occlusion. Moreover, the effectiveness of this method primarily depends on the quality and consistency of feature extraction, and the feature-matching process often requires handling many parameters.

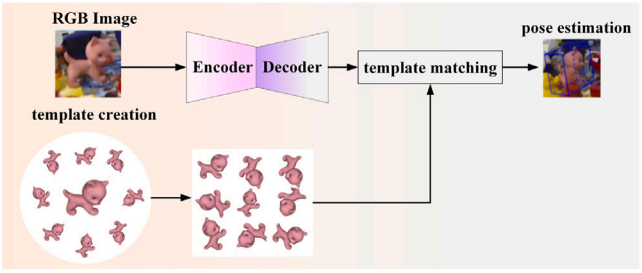

Fig. 5. Template-based methods. This method queries the input image against a pre-established template library, and then finds the template in the library whose pose is closest to that of the input image. The pose of this template is considered the pose of the target object.

2.1.3. Template-based methods

Template-based object pose estimation is a classical and widely used technique that utilizes pre-defined object templates for pose estimation. This method is robust when targeting targets with insignificant texture features. Its matching process and results are intuitive and interpretable, which makes it easy to understand and debug, and it is especially suitable for applications in scenarios such as industrial inspection and robotic grasping, which require high accuracy and reliability. Fig. 5 shows an overview diagram of the template-based object pose estimation method.

[58] presents a template matching-based pose estimation method that matches input images with candidate templates by learning local target representations. Unlike previous global feature representation methods, this method focuses on local features. It retrieves the most similar templates through local features to estimate the pose, which makes it better able to deal with the interference and occlusion problems of the background. However, considering only local features and ignoring the global features of the object has some limitations when dealing with complex objects. ZS6D [59] utilizes pre-trained Vision Transformers (ViT) to extract image features and achieves zero-sample pose estimation through template matching and local correspondence estimation without task-specific fine-tuning. The proposed method demonstrates the effectiveness of pre-trained ViTs in extracting generic image descriptors. It provides a new research direction for future refinement of pose assumptions without fine-tuning. GigaPose [60] utilizes two feature extractors, Fae and Fist, to extract dense features from the input image and the template. Fae is trained through local contrastive learning to be invariant to in-plane and scale but sensitive to out-of-plane rotations. Fist and lightweight MLPs are then used to predict in-plane rotation, scale, and 2D translation from matched image patches. The predicted affine transformations are optimized and validated through the RANSAC algorithm to find the best pose candidates. This method can be integrated into any existing refinement method for higher accuracy.

OSOP [61] is a one-time pose estimation method aiming to infer the pose of an object from a single observation. OSOP employs a multi-stage processing flow to realize one-time segmentation, template matching, dense 2D–2D matching, and pose estimation. Unlike other methods, OSOP introduces an attention mechanism in the segmentation stage to enhance the feature representation and further improve the segmentation accuracy. However, additional domain adaptation techniques may be required to improve the generalization ability of the feature extractor due to the differences between the synthetic data and the actual scene. PoseRBPF [62] combines Rao-Blackwellized particle filtering [63] and an autoencoder network to improve efficiency by decoupling 3D rotation and translation. It optimizes the rotation distribution using feature embedding codebook and cosine distance and supports initialization from 2D target detection or global sampling, which can efficiently handle symmetric targets and maintain accurate posterior distribution. However, the quality of the feature embedding

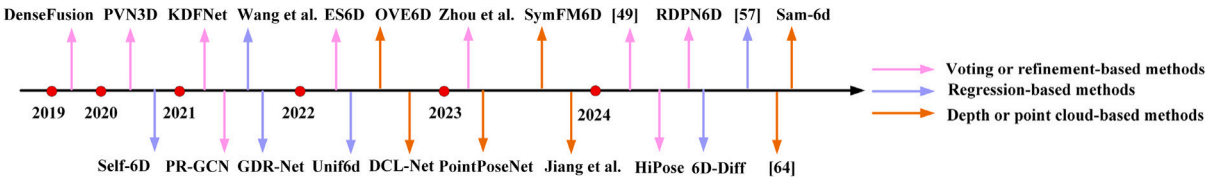

Fig. 6. Timeline of RGB-D image-based and depth or point cloud data-based methods. Notably, the pink arrows represent methods based on voting or refinement, the blue arrows indicate direct regression methods, and the orange arrows represent methods based on depth or point cloud data.

codebook constructed by the autoencoder network directly impacts the accuracy of rotation estimation, requiring ample training data and fine-tuning.

[64] presents a shared template representation learning method for multi-object pose estimation. The method improves the network's generalization performance to complex scenes by combining metric learning and reconstruction learning as similarity constraints and introducing semantic feature constraints to enhance the network's response to objects and reduce the response to the background. However, the additional network modules introduced may lead to overly complex models and increase the consumption of storage resources. Unlike existing methods, GS-Pose [65] employs a novel two-stage framework to optimize object localization, initial pose estimation, and pose refinement. The key innovation is that it utilizes a combination of object representations, including object semantic representations [66], rotation-aware embedding vectors, and 3D Gaussian object representations. The combination of these representations significantly improves the performance of the inference phase. To further enhance the effectiveness of GS-Pose in real-world applications, future work could explore extending it to the field of 6D pose tracking to adapt to the challenges of dynamic environments.

In summary, the template-based object pose estimation method performs well in specific domains, and with the development of deep learning and multimodal fusion technology, its application prospect is even broader. However, this template-based object pose estimation method is highly dependent on the template and is challenging when dealing with severely occluded targets. Continuous robustness and computational efficiency optimization are still needed to cope with more complex and diverse practical application requirements.

2.2. RGB-D based input image

The instance-level object pose estimation method based on RGB-D input images combines the visual features of RGB images and the three-dimensional geometric features of depth images, providing richer data for pose estimation. The depth information can effectively complement the spatial information that is difficult to obtain from RGB images alone, making the method more robust in dealing with complex scenes and occlusion situations. Researchers have explored various fusion methods, including early fusion, late fusion, and intermediate feature fusion. They also introduced transformer-based architectures to capture long dependencies and global features more effectively. Fig. 6 shows a chronological overview diagram of object pose estimation methods based on RGB-D images.

2.2.1. Voting-based or refinement methods

The voting-based object pose estimation method determines the final pose of an object by voting on the pose estimation results of localized feature points. This method utilizes the local information of each feature point and enhances the robustness of occlusions and complex backgrounds through a cumulative voting mechanism. In recent years, voting-based object pose estimation has been gradually combined with deep learning techniques to extract more discriminative features through neural networks. There are also some methods that combine the confidence level with the point voting strategy to give the high-confidence points a more considerable voting weight, thus reducing the negative impact of noisy or low-confidence points on the pose estimation results. In addition, the method also shows excellent performance in multi-target scenarios since the voting mechanism can effectively separate and recognize multiple objects. Fig. 7 is an overview diagram of the voting-based object pose estimation method.

The innovation of PVN3D [67] lies in its keypoint voting mechanism and multi-task learning strategy, which enables it to excel in handling occlusion scenes. It uses a deep Hough voting network [68] to identify 3D key points of objects, combines PSPNet [69] and pre-trained ResNet34 to extract appearance features of RGB images, and utilizes PointNet++ [70] to extract geometric information of point clouds. By simultaneously training the 3D keypoint detection, instance semantic segmentation, and center point voting modules, PVN3D achieves multi-task learning. However, keypoint selection significantly impacts pose estimation accuracy and must be carefully designed to ensure optimal performance. Unlike PVN3D, KDFNet [71] introduces a "Keypoint Distance Field (KDF)" to improve the processing of slender objects. The method uses a fully convolutional neural network to regress the KDF of each keypoint and proposes a distance-based voting scheme to localize the keypoints by computing the intersection of circles for RANSAC voting. Ultimately, the predicted 2D and 3D keypoint coordinates are used to recover the 6D pose. On the Occlusion LINEMOD(O-LM) [72], KDFNet achieves state-of-the-art performance at the time.

ES6D [73] proposed a shape representation method based on grouped primitives and designed a symmetry-invariant pose distance metric to handle the symmetry of objects, which solves the estimation ambiguity problem of symmetric objects. In the point-by-point feature fusion stage, ES6D extracts local features by 2D convolution and splices them with the XYZ mapping map to recover the spatial structure and encodes each point's local features and coordinates by convolution. Although not strictly a voting-based approach, ES6D performs 6D pose estimation by fusing RGB and depth information point-by-point and iteratively optimizes at each point, similar to an implicit voting mechanism. DenseFusion [74] is also an implicit voting method like ES6D. It uses a heterogeneous architecture to preserve the native structure of RGB images and depth maps and semantically segments [75] the RGB images through an encoder–decoder network [76]. Then, the geometric features of the point cloud extracted by PointNet [77] are fused with color features to form a dense feature representation. During pose estimation, DenseFusion uses an end-to-end iterative refinement approach to improve accuracy, but this complex architecture and iterative refinement may increase the difficulty of training and debugging. [78] uses the DFTr network to perform cross-modal fusion of color and depth images. Unlike traditional data fusion methods, it employs a Transformer structure to capture the semantic correlation between the two modalities. It also models long-range dependencies of cross-modal features through self-attention mechanisms. Additionally, it introduces a new weighted vector voting algorithm, achieving precise 3D keypoint localization and near-real-time inference speed. This algorithm enhances robustness to occlusion and complex scenes, maintaining high accuracy even in cases of surface reflections or lack of texture.

PR-GCN [79] achieves 6D pose estimation based on graph convolutional networks, which efficiently capture the local structural information of the point cloud through graph convolutional structures. The method also includes a point cloud refinement module trained by multiresolution regression loss to optimize the original point cloud,

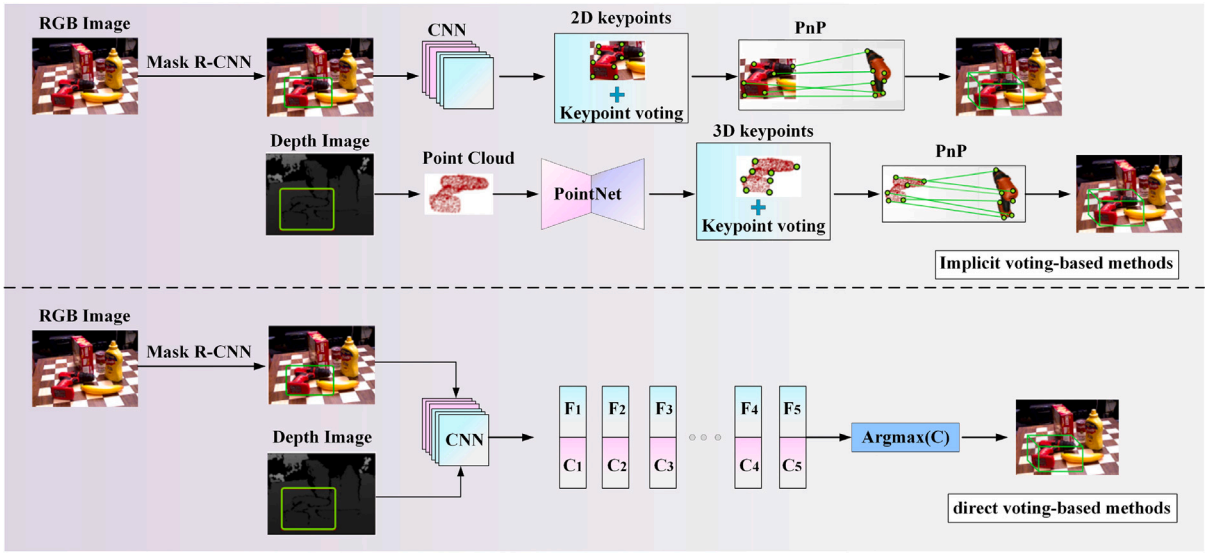

Fig. 7. Voting-based or refinement methods. Two common types are implicit voting and direct voting methods. Implicit voting determines the object pose through pixel-level or point-level voting schemes. Direct voting methods often incorporate confidence strategies, assigning higher voting weights to points with higher confidence, thereby enabling a more accurate recovery of the object pose.

recover missing parts, and remove noise. Overall, PR-GCN is a powerful tool for 6D pose estimation, but further improvements and validations are still needed regarding flexibility, scalability, training efficiency, and generalization ability. FoundationPose [80] utilizes a Transformer-based architecture [81] and a comparative learning formulation to achieve strong generalization capabilities by training large-scale synthetic data. It initializes the global pose by uniform sampling and improves the pose quality through iterative optimization. Finally, a ranking network is also used to score the optimized poses and select the best ones. However, the performance of the model is still somewhat dependent on the quality and diversity of the training data. HiPose [82] uses a two-branch network [83] to process RGB images and point cloud data separately and generates a binary encoding for each point through a two-way fusion network using the ConvNext architecture as a feature extractor. HiPose then establishes a dense correspondence from the RGB-D image to the object model through a hierarchical binary surface coding and iterative process. As a result, it has high estimation accuracy but still faces challenges in dealing with highly occluded objects. Like HiPose, RDPN6D [84] also adopts dense point-to-point correspondences to predict object poses. It proposes a strategy based on residual representation and anchor-based techniques, effectively handling such dense correspondences, reducing the output space, and improving performance. However, learning to predict their 3D coordinates may lead to a decline in model performance for objects with complex shapes or symmetrical features.

In summary, the voting-based object pose estimation method improves accuracy by aggregating multiple predictions and has strong robustness and adaptability. It can reduce the impact of individual prediction errors, improve the adaptability to complex scenes and objects with different angles, and have a high tolerance to noise. However, this method is computationally expensive, may affect real-time performance when dealing with a large amount of voting data, and relies on a large amount of training data, which may not be effective when data is scarce. Additionally, the voting strategy and parameter tuning process are complex, increasing the difficulty of system debugging.

2.2.2. *Direct regression-based methods*

Regression-based object pose estimation methods regress the 6DoF pose parameters of an object directly from an image via a deep neural network. Compared to coordinate or voting-based methods, regression methods directly map image features to pose parameters through

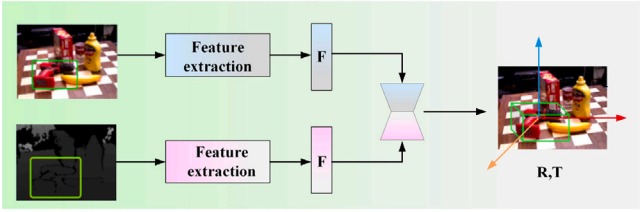

Fig. 8. Direct regression-based methods. Direct regression-based methods refer to neural networks that extract features from RGB images and depth maps, and then directly regress the object's rotation and translation information. Compared to two-stage methods, this approach reduces the consumption of computational resources.

end-to-end training, avoiding complex intermediate steps, which results in faster processing speed. Although it still faces challenges in dealing with occlusion and complex scenes, regression-based methods have become an important direction in the research of object pose estimation due to their efficiency and end-to-end learning. Fig. 8 shows an overview diagram of regression-based methods for object pose estimation.

6D-Diff [85] introduces a diffusion model-based approach for 6D object pose estimation, where a training signal is generated by gradually adding noise through a forward diffusion process, which converts a defined distribution of keypoint coordinates into an uncertain distribution. The inverse diffusion process then progressively denoises and recovers the accurate 2D keypoint coordinates from the noisy distribution. This method provides an end-to-end framework for directly estimating 6D object poses with the advantages of efficiency and accuracy. GDR-Net [86] is a network that combines direct regression and geometric feature representation. It uses a CNN network [87] to extract image features and predict 2D–3D correspondence maps, surface area attention maps, and critical geometric feature maps to improve the accuracy and robustness of pose estimation. Then, direct regression is guided by encoding 2D–3D correspondence and geometric shape information, and symmetry ambiguity is solved using symmetry attention maps. However, GDR-Net relies on large-scale labeled data for training. Self6D [88] combines a two-stage training method and a deep bootstrap network to improve the efficiency of pose estimation through a self-supervised mechanism. First, Self6D utilizes a CNN network to extract features from RGB-D images and generate initial pose estimates.

Subsequently, the network performs synthetic view reconstruction via a self-supervised module and reduces the reliance on large amounts of labeled data. [89], similar to Self6D, it also adopts a two-stage training and self-supervised strategy. In the first stage, fully supervised training is conducted on a synthetic RGB dataset to generate an initial 6D pose estimation. In the second stage, a large amount of unlabeled real RGB-D images is used for self-supervised training to further optimize the model and reduce the domain gap between synthetic and accurate data. Unlike Self6D, [45] introduces a teacher–student network [90] collaboration to enhance the robustness of the model. Although this method reduces dependence on large-scale labeled data, its high demand for computational resources and complex training process needs to be further evaluated for practical applications.

Uni6D [91] solves the projection problem of depth images by introducing UV data and extracts both RGB and depth image features using a single CNN network. Based on Mask R-CNN, Unif6D adds an RT header for predicting 6D pose and an ABC header for auxiliary mapping, realizing end-to-end multi-task learning. This method simplifies the network structure and speeds up the inference. Although its performance may be slightly lower than that of existing methods, it has significant advantages in efficiency and practicality, making it particularly suitable for applications requiring real-time processing. EPRO-GDR [92] enables estimating the probability density distribution of an object's pose by combining existing GDRNPP algorithms [93] with the probabilistic modeling capabilities of EPro-PnP [94], thus allowing multiple meaningful pose candidates to be sampled for each detected object. The end-to-end training approach means the method can learn directly from the data without complex post-processing steps.

In summary, regression-based object pose estimation methods achieve efficient and accurate pose estimation by directly predicting pose parameters. Unlike coordinate alignment-based methods, regression methods can perform fast pose prediction on large-scale datasets and fine-grained prediction in continuous pose space, improving the accuracy of the estimation. However, these methods have a high dependence on the quality and diversity of training data and are sensitive to noise and outliers, which may affect the stability of the model. In addition, the training and optimization process is complex and requires careful design of the loss function and selection of appropriate regression strategies.

2.3. Methods based on depth image or point cloud data

Currently, object pose estimation based on depth images and point clouds has been widely used in real-world scenarios such as robot grasping, medical image processing, and aerospace. These methods utilize 3D information for accurate pose estimation and have rich object shape and distance description capabilities, which enable them to maintain high accuracy in complex scenes and occlusion situations. In recent years, researchers have continued to optimize the accuracy, speed, and robustness of the algorithms and explored lightweight models to accommodate resource-constrained embedded devices. This paper selected several representative methods, as shown in Fig. 6.

SyMFM6D [95] generates compact scene representations by fusing visual and geometric features from multi-view RGB-D images. It integrates information across viewpoints and employs the SIFT-FPS algorithm [96] for 3D keypoint detection and instance semantic segmentation. To enhance the robustness of symmetric objects, SyMFM6D introduces symmetry-aware training and a new objective function to optimize the accuracy of keypoint detection. DCL-Net [97] utilizes two Feature Decoupling and Alignment (FDA) modules to establish a feature space mapping between observable object features and a complete CAD model, effectively separating pose features from corresponding features used for depth matching. Moreover, pose estimation accuracy is enhanced by computing confidence scores to weigh the pose features through feature matching. PointPoseNet [98] employs a multi-stage processing flow in which it first localizes the object by means of a 2D detector and then performs segmentation and keypoint prediction tasks on the point cloud. Its network utilizes a 3D vector field representation to capture the local geometric features of the object and selects the best pose hypothesis through a pose-scoring mechanism. However, this scoring mechanism also implies high computational resources. Future work will aim to overcome these limitations and further improve the robustness and efficiency of the network.

[99] achieved high-precision and high-robustness 6D object pose estimation by progressively mapping the source point cloud to the target point cloud through a diffusion process in the SE(3) space. The authors handle complex transformations in the reverse process linearly and train the model by optimizing the variational lower bound objective function to predict the optimal transformation parameters accurately. However, using the diffusion model increases computational complexity and poses challenges in training and tuning. [100] uses a matching normalization method for point cloud alignment to achieve highly accurate 6D pose estimation through feature extraction, matching and normalization processes. The method first extracts high-dimensional features from the point cloud using a CNN network, then aligns the features of the source point cloud with those of the target point cloud through a matching normalization module, and finally optimizes the network parameters by minimizing the matching error to compute the 3D rotation and translation of the object.

In summary, current instance-level object pose estimation methods aim to address key challenges such as occlusion and symmetry by improving network architectures, optimizing training strategies, and incorporating advanced technologies like diffusion models and Generative Adversarial Networks (GANs). These methods enhance the accuracy and robustness of pose estimation through various approaches, with the introduction of attention mechanisms offering new perspectives on feature extraction and modality fusion. The goal of instance-level methods is to estimate the six degrees of freedom (6D) pose of a specific known object, typically requiring precise identification and localization of a single object's pose. This means the model must accurately locate and describe the position and orientation of the target object within an image or scene. Since the object types are known and fixed, attention mechanisms in this task are typically used to strengthen the focus on specific regions of the object, such as edges, corners, or salient areas. Unlike traditional attention mechanisms, MSRA [101] focuses on the geometric regions of the object's surface by predicting attention maps for surface regions and assigning probability values to each region, implicitly representing the symmetry of the object, thus improving the accuracy of pose estimation. In SO-Pose, the self-attention mechanism does not directly act on the feature map but enhances the 3D representation of the object through self-occlusion information. This method is similar to multi-layer models used in 3D reconstruction but primarily targets monocular image-based pose estimation. Additionally, a deformable attention mechanism [102] can be used to process the surface keypoint features of objects, allowing the network to flexibly sample within the neighborhoods of keypoints, thus better handling occlusions and pose estimation errors. The pixel-level attention mechanism in OSOP calculates the correlation between each pixel and a pre-rendered template, generating an attention weight map that improves the feature extraction capability of the visible parts of the target object. OA-Pose [103] utilizes global semantic similarity to fuse RGB and depth features and performs bidirectional interaction between the two modalities through a cross-attention module, enhancing the feature representation. In these methods, attention mechanisms enhance feature extraction and modality fusion for object pose estimation in different ways. However, these methods often require high computational resources and storage space when processing large-scale 3D data, which may limit the real-time processing capability of the system.

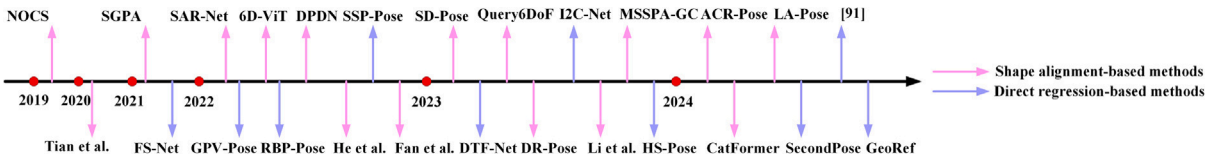

Fig. 9. Timeline overview of shape prior-based methods. Notably, in shape prior-based pose estimation methods, the pink arrows and blue arrows represent two key techniques: shape alignment methods and direct regression methods, respectively.

3. Category-level object pose estimation

All of the above methods are instance-level object pose estimation, which usually relies on the CAD model of the object to be measured. In contrast, many objects in real life do not have CAD models, which limits the application of the above methods. In recent years, researchers have shifted their focus to category-level object pose estimation, aiming to improve the generalization ability of pose estimation. Unlike instance-level methods that rely on a single data source, category-level methods tend to use multimodal data fusion, such as semantic information. Additionally, some methods introduce visual–language models to learn object category information better. According to the existing methods, researchers have made various improvements in terms of network models, data fusion, training methods, and loss optimization. Based on this, this paper will describe two aspects of the methods based on shape a priori and no shape a priori information.

3.1. Shape prior-based methods

Shape priori information is pre-known geometric or structural information about an object class, typically including a 3D model of the object, a silhouette, a keypoint, or a template [104]. By combining shape prior information, the model can more accurately match the object's observations (e.g., images or point clouds) with known 3D shapes, thus improving the accuracy of pose estimation. Especially when dealing with occlusions or complex backgrounds, it can help the model better understand and infer the complete shape of the object. Especially in the multimodal data fusion process, shape prior information can be used to guide the fusion of different data sources and ensure the consistency of data acquired from different viewpoints and sensors. The chronological overview diagram of the method is shown in Fig. 9.

3.1.1. Shape alignment-based methods

Shape alignment-based methods involve aligning observed object data with known shape priors to estimate the object's pose in 3D space. They have obvious advantages in object pose estimation. First, they significantly improve the accuracy and robustness of pose estimation by combining the shape prior information and especially excel in complex scenes and partial occlusion. Second, unlike traditional ICP algorithms, shape alignment methods do not fall into local optima and provide an excellent initial pose estimate, making the optimization process more efficient and stable. The general flow of the shape alignment-based object pose estimation method is shown in Fig. 10.

Objects in different instances usually have different sizes and orientations, and such differences increase the complexity of model learning, making it necessary for the model to deal with both shape and pose changes. NOCS [105] eliminates such differences by normalizing all objects to a uniform coordinate space. It introduces a context-aware mixed reality technique [106] that generates a large amount of synthetic data, which improves the model's generalization ability in natural scenes. SAR-Net [107] proposed a lightweight network structure that handles 3D rotation and shape reconstruction of objects [108] through shape alignment and symmetry correspondence. Unlike SAR-Net, SD-Pose [109] extracts semantic and geometric features of image blocks and point cloud blocks by using PSPNet and PointNet++, respectively, and designs a semantic dynamic fusion module to fuse this information to enhance robustness to noise points. In addition, SD-Pose introduces an information exchange enhancement module to learn the structural relationship between instance features and category prior so that it has both the properties of instance shapes and the commonalities of category structures. However, it also increases the complexity of the network, leading to slower inference.

DR-Pose [110] designed a two-stage deformation and alignment process for the pose estimation problem of occluded objects. In the assisted deformation phase, DR-Pose utilizes the PoinTr network to recover the unseen portion of the target object and improve the model's performance in dealing with occluded objects. In the alignment phase, the system uses the KPFCN as a feature encoder to extract geometric features and enhances them with a position-aware transformer. This two-phase design allows the network to focus more on network design and optimization for each task, thus improving overall performance. Semantic information can provide category and contextual information to the model, thus helping the network to better understand and distinguish between different class objects. [111] utilizes a local segmentation network to divide the observed point cloud into components with explicit semantic labels, thus establishing a mapping between the actual observed data and the abstract representation space. The 9D pose information is then recovered through semantic correspondences. Compared with other methods, this method requires fewer parameters, which helps to reduce model complexity and the risk of overfitting. SGPA [112] also incorporates semantic information, but unlike [59], the semantic information in SGPA is derived from RGB images. Specifically, SGPA combines semantic information with a priori information to dynamically adapt to each particular object. In addition, SGPA employs a low-rank transformer [113] with structural regularization to guide the a priori adaptation process, ensuring that the most representative features are taken into account in the adaptation process. However, inappropriate keypoints may affect the accuracy of the final pose estimation.

6D-ViT [114] designed Pixelformer and Pointformer structures to extract the appearance representation of RGB images and the geometric features of 3D point clouds, respectively, and combine them with the category shape prior to generating dense instance representations. Finally, the correspondence between the dense representation, shape prior, and instance point cloud is utilized to compute the 6D pose of the object. This approach, based on the dual-stream Transformer architecture and multi-source feature fusion, improves the accuracy of pose estimation. CatFormer [115] achieves high-precision object pose prediction through three key steps: coarse deformation, fine deformation, and iterative refinement. Its core module employs self-attention and cross-attention mechanisms [116] to deform and complement point clouds. Query6DOF [117] implements category-level 6D pose estimation through a query deformation estimator, a query point correspondence estimator, and a pose size estimator. The query deformation estimator adapts the category-specific query to the representation of the target object through the attention mechanism, the query point correspondence estimator establishes the correspondence between the deformation query and the object features, and the pose size estimator predicts the rotation, translation, and size of the object through global average pooling [118] and multi-layer perceptron. Despite its effectiveness, the attention mechanism imposes high computational resource consumption and thus needs to be optimized in the future to reduce computational burden and improve inference speed.

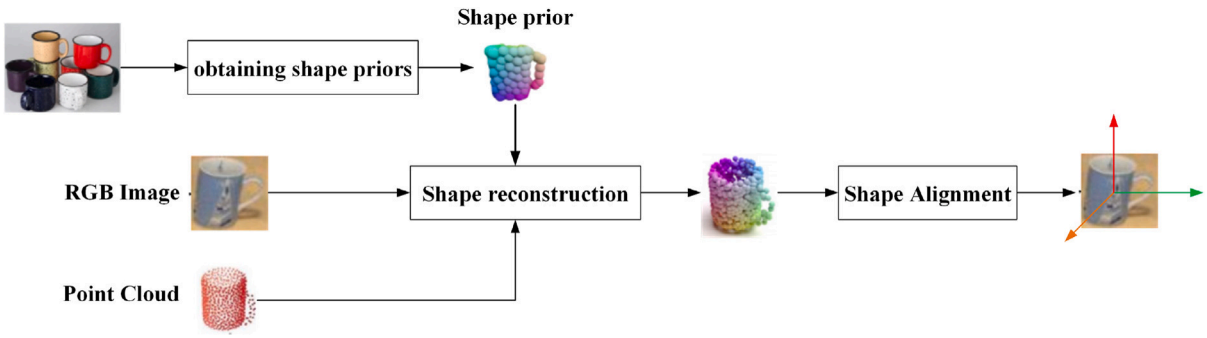

Fig. 10. Shape alignment-based methods. Taking RGB-D image input as an example, the NOCS shape alignment method first learns a model to predict the target's NOCS shape/map, and then aligns the target point cloud with the NOCS shape/map using non-differentiable pose-solving methods, such as the Meshing algorithm, to solve for the target pose.

ACR-Pose [119] addresses the problems of intra-class shape change [120], canonical representation reconstruction, and pose vs. shape estimation by introducing an adversarial training mechanism for reconstructors and discriminators. The method utilizes pose-independent and relational reconstruction modules, focuses on learning shape information, and generates high-quality canonical representations with multiple loss constraints, significantly improving pose estimation's performance. Although it performs well on the CAMERA25 [105] and REAL275 [105] datasets, its ability to generalize to real-world scenarios requires further validation. DPDN [121] contains three main modules: a triple feature extractor, a depth a priori deformer, and a pose-size estimator. By extracting point-level features from the target image, point cloud, and category shape prior while using depth correspondence to deform the prior in feature space to match the object observation, it finally predicts the pose and size of the object directly. Its self-supervised learning-based approach reduces the gap between the synthesized and real data domains. However, DPDN relies on the quality of the category shape prior and may be limited by the accuracy of the prior data.

[122] presents an advanced network structure. The structure is realized in three key phases: an instance segmentation phase to obtain foreground masks, a 3D model reconstruction phase to learn shape priors using autoencoders, and a pose estimation phase. The network cleverly combines a PointNet encoder, a fully-connected decoder, and a fully-convolutional network to process point cloud and image data, and the authors introduce a "Map operator" [123] operator to solve the ambiguity problem caused by object symmetry. [124] is used to solve the problem of object pose estimation in the absence of accurate pose labeling. The method first deforms a template mesh using a differentiable shape deformation network to remove differences in shape, pose, and scale. Then, a point cloud alignment network [125] is applied to estimate pose and scale parameters. Finally, the deformed mesh is rendered using a differentiable renderer [126] to reinforce the geometric consistency between the point cloud extracted from the rendered depth image and the observed scene. The algorithm's robustness to various variations is enhanced by progressively removing shape, pose, scale, and part-to-complete inconsistencies.

OLD-Net [127] is a network for monocular RGB images that overcomes the limitation of relying on depth sensors by directly predicting depth information at the object level. The network solves the scale ambiguity problem by providing absolute depth information through the Normalized Global Position Hints (NGPH) module. Shape points and depth translations are independently predicted using the Shape-aware Decoupled Depth Reconstruction (SDDR) module, preserving shape details. LaPose [128] is an innovative network framework that quantifies shape uncertainty by modeling object shapes as Laplace Mixture Models (LMMs) [129], effectively addressing the challenges caused by the lack of depth information. In addition, LaPose proposes a scale agnostic pose representation that solves the scale-ambiguity problem in a single RGB image and improves pose estimation's training efficiency and stability. Despite its high computational complexity, it demonstrates excellent performance across multiple datasets.

MSSPA-GC [130] uses a PointNet-like MLP network to extract shape-related object features from a priori shape point clouds. Two decoders are designed to regress the deformation field and the correspondence matrix, respectively, so as to reconstruct the normalized object coordinate space coordinates. Finally, the reconstructed NOCS coordinates and the observed point cloud are utilized to recover the rotation and translation information. MSSPA-GC shows strong competitiveness through its innovative network structure and multi-scale feature extraction method, but it still has room for improvement in integrating complementary information from multiple input sources and data enhancement.

In summary, shape-alignment-based category-level object pose estimation methods usually employ advanced network structures such as CNN networks, Transformer architecture, and PointNet to extract features of the target object. In order to improve the model performance, researchers have adopted various strategies, including fusing semantic and geometric information, introducing a self-attention mechanism and cross-attention mechanism, integrating a graph convolution module, and applying iterative refinement techniques. When confronted with the object occlusion problem, adversarial generative networks (GANs) are used to generate objects in the occluded part, thus enhancing the robustness of the model. Through these methods, researchers not only improved the quality of the input images but also optimized the network structure, which ultimately achieved a significant improvement in the performance of category-level object pose estimation. However, attention should also be paid to the complexity of the model, the speed of inference, and how it can be deployed to real-world applications.

3.1.2. Direct regression-based methods

The direct regression-based category-level object pose estimation method trains deep learning models to directly regress the object's pose parameters without the need for multi-stage processes or post-processing like PnP for pose refinement. This approach typically adopts an end-to-end training manner, simplifying both the training and deployment of the model. Moreover, the entire network can be optimized using a unified loss function, making the training process more efficient. Especially in the pose prediction stage, direct regression-based methods output pose parameters with a single forward pass, making them suitable for real-time applications. The general workflow of direct regression-based object pose estimation methods is shown in Fig. 11.

FS-Net [131] is an end-to-end bit pose estimation network that reconstructs the input point cloud via a 3D graph convolutional network. To cope with intra-class shape variations, FS-Net introduces an approach based on online shape morphing, which simulates instances of different shapes by dynamically adjusting the object bounding box without additional storage resources. Meanwhile, to improve the accuracy and robustness of pose estimation in complex scenes, FS-Net uses three independent branches to recover the rotation and translation information separately, and this decoupled approach reduces the interference of rotation and translation information and enhances the

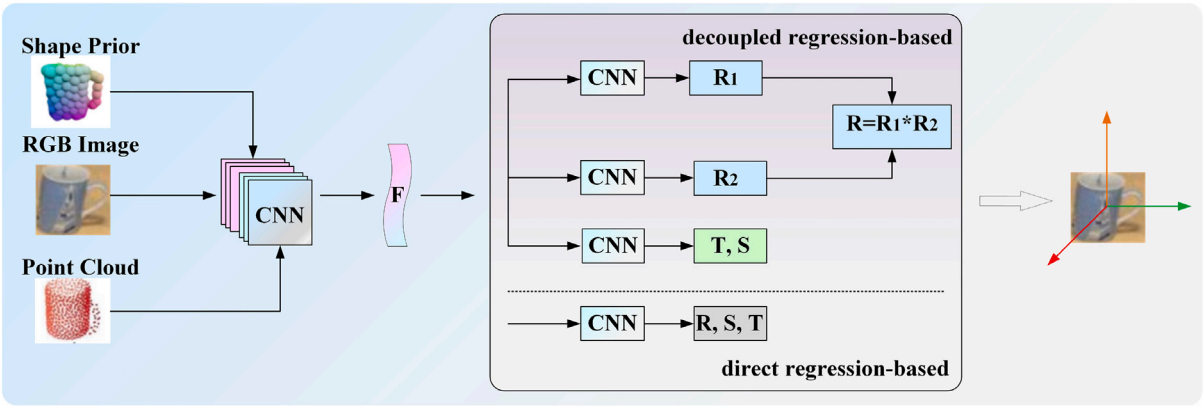

Fig. 11. Direct regression-based methods. Unlike instance-level direct regression methods, category-level direct regression methods utilize richer data sources, such as point cloud information. Additionally, when regressing the object pose, this method adopts a decoupled approach, where rotation and translation information are processed separately, effectively avoiding mutual interference between the two.

model's ability to adapt to pose changes. GPV-Pose [132] is also an approach based on 3D graph convolution, and unlike FS-Net, GPV-Pose enhances the robustness of 3D rotation recovery by introducing a decoupled rotation representation with confidence. The method decomposes the rotation matrix into plane normals of the object bounding box and predicts the confidence level for each normal. A geometrically guided point-voting mechanism [133] predicts the direction, distance, and confidence level of each observation point with respect to the bounding box face and computes the planar parameters of the bounding box using a weighted least squares method [11]. The geometric relationship between the pose, point cloud, and the bounding box is ultimately combined as a supervised term to optimize the network. GPV-Pose achieves inference speeds of up to 69 FPS on the REAL275 dataset, making it suitable for real-time applications. RBP-Pose [134] proposed an online data enhancement method, but unlike FS-Net, RBP-Pose's data enhancement method is a nonlinear one. Specifically, for objects with reflective symmetry properties, such as laptops, PBP-Pose changes the shape of the object by changing the angle between the upper and lower planes. For objects with rotational symmetry, such as bottles, it deforms the object shape along the direction of the symmetry axis by a nonlinear function in order to adjust the scale. In addition, RBP-Pose realizes real-time inference at a frame rate of 25 Hz, which exhibits high processing efficiency.

HS-Pose [135] improves the 3D graph convolutional network by introducing a sensory field module with feature distances, which enhances the extraction of global and local geometric features. Meanwhile, in order to improve robustness, HS-Pose uses an outlier robust feature extraction layer (ORL), which extracts features through local area bootstrapping and maximum pooling [136], combines global average pooling to generate global features, and adjusts the features through a linear layer to reduce the effect of noise on the model. SSP-Pose [137] focuses on pose estimation of symmetric objects by utilizing the symmetry information of the object to guide the deformation of the shape a priori and introduces symmetry-aware loss to reduce ambiguity in the matching process. SSP-Pose excels in symmetric object processing, but its generalization ability to complex and asymmetric objects still needs to be improved. SecondPose [138] extracts color and geometric features from RGB images and point cloud data, respectively, via a dual-stream structure, fuses semantic information using a cross-modal cross-attention mechanism and introduces SE(3) consistency constraints to ensure consistency under spatial transformations. The model is optimized for translation loss, rotation loss, and SE(3) consistency loss through end-to-end training, demonstrating superior performance on synthetic and real datasets but may not fully reflect performance in diverse real-world scenes.

DTF-Net [139] innovatively solves the problem of shape change of category-level objects by introducing a deformable template field. This deformable template field includes a deformation network and a template network for learning geometric deformation features between category template features and observation instances. Meanwhile, DTF-Net combines a pose regression module with a shape-invariant training strategy to achieve end-to-end inference, demonstrating good robustness and real-time performance. Despite its high dependence on training data and high computational resource requirements, it performs well in public benchmarks and is successfully applied to real-world robot grasping tasks. i2c-net [140] differs from other category-level methods in that it achieves accurate estimation of object pose from monocular RGB images by combining instance-level networks, custom synthetic datasets, 3D model reconstruction, and depth-informed post-processing corrections. Among them, the custom dataset is generated using the BlenderProc tool [141], which provides realistic images and 6D pose annotations. GeoReF [142] effectively handles large variations in object shapes within categories by integrating shape-prior information and innovative graph convolution mechanisms. It utilizes learnable affine and cross-cloud transformation mechanisms [143] to dynamically adjust the input point cloud and features to establish precise geometric correspondences between different object shapes. [144] introduced a network architecture based on self-supervised 6D pose estimation, which consists of a part-level shape reconstruction module (PSR) and a coarse-to-fine correspondence optimization module (CFCO) to capture local shape changes of objects and optimize the correspondence between pixels and point clouds. However, it is highly complex, especially when dealing with large-scale point cloud data.

In summary, regression-based category-level object pose estimation methods typically employ technologies such as 3DGCNs, dual-stream architectures, shape deformation, and symmetry-aware constraints to enhance the recognition capability of complex objects. They also use decoupled pose representation methods to reduce interference between rotation and translation. These methods have shown good performance in handling intra-class shape variations, pose estimation in complex scenes, and pose estimation of symmetric objects. However, they still face challenges such as reliance on training data and high computational resource demands. Future improvements may include enhancing model generalization, employing self-supervised learning methods, and increasing model inference speed and real-time capabilities to ensure effective performance in various practical applications.

Table 2 summarizes the above shape a priori information-based category-level object pose estimation methods.

3.2. Shape prior-free methods

In category-level object pose estimation, while the introduction of shape-prior information can improve the accuracy of pose estimation, methods based on a shape-free prior also have their unique advantages.

Table 2
Shape alignment-based methods. For each method, we introduce its implementation form, release year, training input, processing categories, and performance metrics on category-level datasets (CAMERA25 and REAL275).

Methods	Years	Data	Type	Evaluation metrics(%)			
				REAL275		CAMERA25	
				IOU_{50}	10° 5 cm	IOU_{50}	10° 5 cm
NOCS	2019	RGB-D	alignment	78.0	25.2	83.9	64.6
[76]	2020	RGB-D	alignment	77.3	54.1	93.2	81.5
SGPA	2021	RGB-D	alignment	80.1	70.7	93.2	88.4
SAR-Net	2022	RGB-D	alignment	79.3	68.3	86.8	80.3
6D-ViT	2022	RGB-D	alignment	83.06	67.89	93.46	87.98
DPDN	2022	RGB-D	alignment	83.4	78.4	83.0	72.1
[74]	2022	D	alignment	65	42.6	–	–
[77]	2022	RGB	alignment	25.4	–	32.1	–
SD-Pose	2023	RGB-D	alignment	83.2	71.2	93.4	87.7
Query6DOF	2023	RGB-D	alignment	82.5	83.0	91.9	90.0
DR-Pose	2023	RGB-D	alignment	78.9	76.3	92.7	89.7
MSSPA-GC	2023	D	alignment	81.7	77.0	92.8	88.3
[70]	2023	RGB-D	alignment	79.0	76.3	95.7	86.9
ACR-Pose	2024	RGB-D	alignment	82.8	65.9	93.8	87.8
CatFormer	2024	RGB-D	alignment	83.1	79.5	93.5	90.2
LA-Pose	2024	RGB	alignment	48.8	55.4	49.4	73.1
FS-Net	2021	D	regression	81.1	69.1	–	60.8
GPV-Pose	2022	D	regression	83.0	73.3	93.4	89.0
RBP-Pose	2022	RGB-D	regression	–	79.2	93.1	89.5
SSP-Pose	2022	RGB-D	regression	82.3	77.8	–	87.4
DTF-Net	2023	RGB-D	regression	84.0	79.7	94.5	85.5
I2c-net	2023	RGB	regression	92.46	49.42	–	–
HS-Pose	2023	D	regression	82.1	82.7	93.3	89.4
SecondPose	2024	RGB-D	regression	–	86.0	–	–
[91]	2024	RGB-D	regression	44.5	30.4	–	–
GeoRef	2024	RGB-D	regression	77.7	75.2	–	90.5

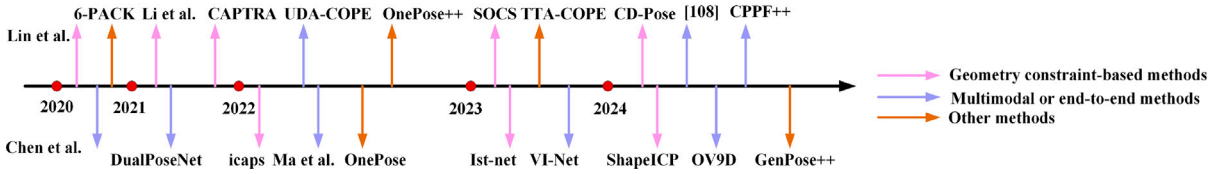

Fig. 12. Timeline overview of category-level object pose estimation methods based on shape-agnostic priors. The pink arrows, blue arrows, and orange arrows represent methods based on geometric constraints, multimodal or end-to-end methods, and other methods, respectively.

These methods do not depend on pre-defined templates or models, making them more flexible when dealing with unknown objects or instances with significant intra-class shape variations. By utilizing strategies such as geometric constraints and multimodal information fusion, they enhance performance, maintaining robustness while reducing dependence on shape priors. The paper categorizes the methods into those based on geometric constraints, those based on multimodal or end-to-end approaches, and other methods. Fig. 12 shows a chronological overview diagram of the category-level object pose estimation methods based on shape-free a priori.

3.2.1. *Geometry-constrained methods*

In the absence of shape prior information, geometric constraints provide additional guidance for pose estimation by utilizing the geometric properties of objects, such as plane normals, symmetry, bounding boxes, and projection relationships. This ensures that the estimated pose aligns with the object's geometric structure and spatial arrangement. Geometric constraints play a multifaceted role in category-level object pose estimation. They not only help reduce uncertainties caused by viewpoint changes and occlusions but also improve the model's generalization ability to new instances by guiding the learning process. Additionally, geometric constraints are crucial for estimating the pose of objects in complex environments with occlusions or partial visibility. The general flow of the geometric constraint-based object pose estimation method is shown in Fig. 13.

iCaps [145] improves the estimation accuracy of shape and pose by using a category-level self-encoder network and a particle filtering

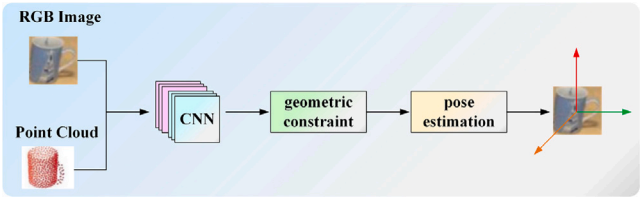

Fig. 13. Geometry-constrained methods. In pose estimation, common geometric constraint methods (such as symmetry-aware constraints) significantly enhance the model's understanding of the object's geometric shape by introducing symmetry-related geometric constraints, thereby improving the accuracy and robustness of pose estimation.

framework to estimate and track the 6D pose of an object, combined with an implicit shape representation based on the signed distance function (SDF). The method has good category-level generalization capability and gradually improves the estimation accuracy through iterative refinement. SOCS [146] deforms and aligns objects using a sparse set of key points and semantic correspondences, creating a semantically consistent coordinate space and overcoming the inaccuracies of traditional NOCS methods in large shape variations. It introduces a multi-scale coordinate attention network and a surface-independent point sampling strategy to extract features and handle occlusions while using contrastive training and pose consistency loss

to enhance robustness. As a result, SOCS performs excellently in generalization and understanding complex scenes, though there is still room for improvement in handling complex shapes and asymmetric objects.

IST-Net [147] proposes an implicit spatial transformation network. It realizes 6D pose estimation without shape a priori information through a feature extractor, implicit spatial transformation module, camera space enhancer, world space enhancer, and pose estimator. The use of integrated loss function optimization during training ensures an overall improvement in network performance, but also requires balancing performance with network complexity. [148] uses an SE(3) equivariant point cloud network as the backbone for feature extraction to ensure equivariance to 3D rotations and translations. Similar to FS-Net, this method estimates object shape and pose through an invariant shape reconstruction module and an equivariant pose estimation module, respectively. It improves prediction accuracy by minimizing the consistency loss between observed and reconstructed shapes through self-supervised learning. However, attention needs to be paid to its handling of symmetric objects and its performance under highly deviated viewpoint distributions in practical applications.

[149] proposed a pose estimation method based on Instance-Adaptive Keypoint Detection (IAKD) and Geometric-Aware Feature Aggregation (GAFA). The IAKD module dynamically detects sparse keypoints for different instances to capture the geometric structure of each instance. The GAFA module integrates local and global geometric information to establish stable keypoint-level correspondences. CD-Pose [150] tackles the problem of intra-category variation, inter-category similarity, and complex structured objects by combining two modules, the Pose-Consistent Module for geometric consistency learning and the Pose-Discrepant Module for geometric disparity learning. ShapeICP [151] effectively handles the coupled attitude and shape problem and improves the algorithm's robustness using a grid-based Active Shape Model (ASM) and an alternating attitude and shape optimization strategy. Despite the high computational complexity and sensitivity to the initial attitude, ShapeICP demonstrates excellent performance without needing attitude labeling data, providing new directions for future research, such as improving the accuracy of local minima processing and shape initialization network.

In summary, the field of category-level object pose estimation based on geometric constraints has seen an influx of innovative methods, including iCaps, SOCS, IST-Net, CAPTRA, and others. These methods improve the generalization ability to unseen instances and pose estimation accuracy by exploiting deep learning, implicit spatial transformations, isomorphic networks, pose normalization, and keypoint detection techniques. However, there is still a need to improve the performance in terms of computational resource requirements, generalization ability, and handling of complex topologies and asymmetric objects to achieve more efficient and accurate pose estimation.

3.2.2. Multimodal or end-to-end based methods

Data from a single modality often cannot adequately capture all the information about an object, so researchers have proposed multimodal and end-to-end approaches to integrate data from different modalities more comprehensively. In recent years, many new fusion strategies have been proposed to significantly improve the accuracy of pose estimation through the continuous improvement of multimodal fusion techniques, especially in RGB-D fusion and multiview fusion. The end-to-end method utilizes CNN networks, PointNet, transformer, self-encoder networks, etc., to recover the object pose information directly from the input features, which simplifies the traditional complex processing flow. Fig. 14 illustrates the basic flow of the method.

Genpose [152] applies diffusion models to the problem of object pose estimation by generating potential pose distributions through a score-based generative model and progressively optimizing these poses using an energy-based diffusion model to reduce noise and error. At the core of OV9D [153] is a framework based on a variety of pre-trained models, which encodes textual descriptions using pre-trained

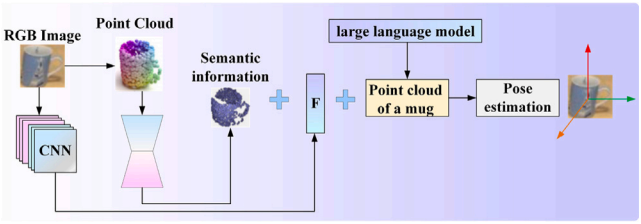

Fig. 14. Multimodal or end-to-end based methods. This method typically employs a dual-branch structure that integrates RGB appearance features, semantic information, and object geometric features to enhance the model's comprehensive understanding of the object. Additionally, large language models (LLMs) can be used to generate labels for each category as a supplementary information source, further boosting the network's feature extraction capabilities and category generalization performance.

visual language models and generates visual features consistent with the textual descriptions via a diffusion model to infer a normalized object coordinate space (NOCS) map of the target and can deal with classes of objects that have not been seen during the training phase. CLIPose [154] improved pose estimation accuracy by integrating three modalities: 3D point clouds, image patches, and textual descriptions. It utilized a pre-trained CLIP model to generate text descriptions corresponding to images and aligned features from different modalities through contrastive learning while fine-tuning the image encoder to enhance sensitivity to pose estimation. Experimental results showed that CLIPose achieved leading performance on mainstream benchmark datasets. However, its performance is somewhat dependent on the pre-trained CLIP model and requires significant computational resources for multimodal feature extraction and alignment.

DualPoseNet [155] designed a spherical convolution-based encoder to learn pose-related shape features and integrate color and geometric features via a spherical fusion module. Its direct regression of the rotation matrix, translation vectors, and dimensional information is accomplished through an explicit decoder and an implicit decoder for reconstructing the input point cloud and implicitly predicting the pose. This dual decoder structure improves the accuracy of pose prediction through complementary supervision. VINet [[156] also proposed a spherical convolution-based method, simplifying the complex rotation problem by projecting point clouds onto a sphere and addressing rotation angles on the sphere. It used two independent networks to predict rotation angles in pitch, yaw, and roll directions. This rotation decoupling approach effectively avoided error accumulation caused by rotation coupling while reducing computational complexity. UDA-COPE [157] designed an unsupervised domain adaptive strategy that utilizes RGB-D data combined with an instructor-student self-supervised learning framework and bi-directional point filtering techniques to achieve efficient category-level 6D object pose estimation without manual annotation. Although it has some limitations in terms of object segmentation accuracy, single-frame image dependency, and demand for computational resources, it shows great potential for multimodal data integration and cross-domain knowledge transfer.

[158] extracts multi-scale feature maps from RGB images using a progressively refined neural feature rendering technique. In this process, feature maps of different resolutions are fused to ensure feature richness and completeness of details. Finally, the pose estimation network recovers the refined features as 6D pose information and optimizes the model with multiple loss functions. [159] reconstructs 3D point clouds of normalized poses from RGB-D images via variational self-encoders. This process does not rely on class-specific 3D models but achieves view-invariant 3D shape representations through cross-class training and a large 3D shape library. This approach reduces the reliance on exact correspondences and improves the overall estimation performance by integrating multimodal data. Despite the challenges, it demonstrates state-of-the-art performance on category-level 6D pose estimation tasks with single-view shape reconstruction.

CPPF++ [160] introduces probabilistic uncertainty modeling based on CPPF and improves the robustness and accuracy of the model through several novel modules such as N-point tuple feature extraction, noise pair filtering, online alignment optimization, and tuple feature integration. The method demonstrates good generalization ability on real images. [161] utilizes stereo images to simultaneously handle category-level object detection, pose estimation, and 3D shape reconstruction. The model effectively fuses stereo image features and 3D spatial information to handle objects with multiple surface properties, including diffuse, specular, transparent, and mixed materials. In addition, CODERS' end-to-end learning framework eliminates error accumulation in traditional multi-stage processes, improving operational accuracy and efficiency.

In summary, the multimodal fusion and end-to-end category-level object pose estimation method significantly improves the accuracy and robustness of pose estimation by combining multimodal data, such as RGB images, point clouds, and textual descriptions, and by utilizing techniques such as generative modeling, spherical convolution, self-supervised learning, and variational self-encoder. However, future improvements still include reducing the computational resource requirements, improving the real-time performance and generalization ability of the model, and considering the introduction of time series analysis for a more comprehensive and dynamic understanding of object pose.

3.3. Other methods

In addition to the aforementioned methods based on shape-prior information and shape-free prior information, researchers have developed a variety of innovative methods to address different challenges. By combining self-supervised learning, feature matching, and multiview consistency, these methods have pushed the boundaries of category-level object pose estimation and have shown excellent performance, especially when dealing with complex environments and low-texture objects. In the following paper, the implementation details of these methods are described in detail.

OnePose [162] draws on the idea of visual localization [163] by capturing object data through an AR tool, including determining the object's center position, size, rotation angle around the Z-axis, and camera pose. SfM is then used to reconstruct the sparse point cloud and extract 2D keypoints and descriptors. During the localization phase, the system captures a series of new images in real-time to match queries and recover the object pose. In this way, OnePose is able to efficiently recover the 3D pose of the camera from the captured data, providing a novel solution for object pose estimation. OnePose++ [164] proposed a pose estimation process based on OnePose that does not require the detection of keypoints, making it more effective in dealing with low-textured objects. In the feature matching phase, it utilizes the LoFTR technique [165] for local feature matching without a detector and improves the point cloud accuracy through a two-stage 3D structure reconstruction process. In the testing phase, OnePose++ uses a sparse-to-dense 2D–3D matching network, enhances the robustness of 2D–3D matching through self-attention and cross-attention mechanisms, and ultimately achieves high-accuracy object pose estimation without CAD models.

TTA-COPE [166] proposes a test-time adaptive method to solve the domain gap problem and improve the accuracy of object pose estimation by combining pre-training and online adaptation. The method utilizes unlabeled target data for adaptive updating at test time, optimizes with self-training loss of pose-aware confidence, and applies point filtering [167] to enhance the robustness of the model. TTA-COPE does not require labeled target data or source data access, which makes it well-suited for real-time application scenarios. 6-PACK [168] utilizes an attention mechanism on a grid of 3D anchor points to identify and attend to anchor points around the predicted object position that capture RGB-D features of the surrounding volume. Through unsupervised learning, 6-PACK generates 3D keypoints from the most likely anchor points without manual labeling and trains the network with multiview consistency loss and pose estimation loss to ensure that the keypoints are consistent across consecutive frames and accurately compute pose changes. GenPose++ [169] enhances pose estimation accuracy through semantic-aware feature extraction and a clustering aggregation strategy, particularly excelling with objects that exhibit discrete symmetry. It also achieves significant progress in sim-to-real generalization. Additionally, introducing the Omni6DPose dataset offers a valuable resource for the 6D object pose estimation field.

In summary, category-level object pose estimation without shape a priori information is evolving towards multimodal fusion, unsupervised and self-supervised learning, diffusion model, neural rendering, and refinement. Despite the challenges in terms of computational resource requirements, model complexity, and generalization capabilities, these technological advances provide new ways to improve the accuracy and robustness of pose estimation and are expected to play an essential role in future applications such as robot vision and augmented reality. Table 3 summarizes the above category-level object pose estimation methods based on shapeless a priori information.

Unlike the attention mechanisms in instance-level methods, the training data for category-level methods comes from images of different objects within the same category, which may vary in appearance, shape, and size. In this task, the role of the attention mechanism is to help the model identify and focus on shared features among objects within a category while handling the diversity and deformation issues between different objects. For example, [170] adopts a cross-attention mechanism that interacts the query information with the input point cloud features. Through this mechanism, the query information transforms from a category-specific representation to an instance-specific representation. This dynamic selection of relevant features allows the model to adapt to the diversity of objects within the category. CatFormer uses both cross-attention and self-attention mechanisms to complete and deform the point cloud, enabling the model to better capture both local and global geometric information in the point cloud, thereby improving the ability to model the shape of the target object. SecondPose employs an implicit attention mechanism, enabling the model to focus more on the semantic and geometric consistency of the object during pose estimation. Additionally, 6-PACK proposes an anchor-based attention mechanism, which allows the network to quickly locate the center of the object in 3D space and better handle the position variations of the object in 3D space. Overall, these methods have shown significant advantages in capturing the variability of object shapes, geometric information, and semantic consistency. Although these methods have different focuses, they collectively advance category-level object pose estimation techniques and provide more robust technical support for future object pose estimation in complex scenarios.

4. Pose estimation of unseen objects and articulated bodies

In computer vision, the study of pose estimation for unseen objects and articulated bodies is significant and challenging. The difficulty of these tasks lies in dealing with the unique shapes and diverse appearance features of unknown objects and the complex joint movements of articulated bodies. For unseen objects, the lack of pre-existing geometric models and sufficient training samples makes pose estimation even more difficult. Fortunately, researchers have significantly advanced this area through the application of methods such as data-driven strategies, generative modeling, and unsupervised learning. Similarly, the multi-joint complex motion of articulated bodies has been more accurately understood and predicted through advanced techniques such as deep learning, graph optimization, and physical modeling. Research in these directions will further improve the performance in estimating the pose of unseen objects and articulated bodies, providing more reliable technical support for applications such as robot vision, aerospace, and

Table 3
Shape prior-free methods for category-level object pose estimation. For each method, we introduce its implementation form, release year, training input, processing categories, and performance metrics based on the 5° 5 cm criterion.

Methods	Years	Data	Type	Evaluation metrics(%)	
				REAL275	CAMERA25
				5° 5 cm	5° 5 cm
[97]	2020	RGB-D, CAD	Geometry	61.7	82.8
[95]	2021	D	Geometry	–	–
CAPTRA	2021	RGB-D	Geometry	62.16	–
icaps	2022	D	Geometry	31.59	–
SOCS	2023	RGB-D	Geometry	56	–
Ist-net	2023	RGB-D	Geometry	53.4	–
CD-Pose	2024	RGB-D	Geometry	44.9	73.0
ShapeICP	2024	RGB-D	Geometry	36.5	–
[106]	2020	RGB-D	multi-modal	23.5	–
DualPoseNet	2021	RGB-D	multi-modal	35.9	70.7
UDA-COPE	2022	RGB-D	multi-modal	34.8	–
[89]	2022	RGB	end to end	–	–
VI-Net	2023	RGB-D	multi-modal	57.6	81.4
CLIPose	2024	RGB-D, text	multi-modal	58.3	82.2
GenPose	2024	D	end to end	71.5	–
[108]	2024	Stereo image	end-to-end	–	–
OV9D	2024	RGB, text	multi-modal	–	–
CPPF++	2024	RGB-D	end-to-end	32.3	–
6-PACK	2020	RGB-D	other	33.3	–
OnePose	2022	RGB	other	–	–
OnePose++	2022	RGB	other	–	–
TTA-COPE	2023	RGB-D	other	35.9	–
Genpose++	2024	RGB-D	other	–	–

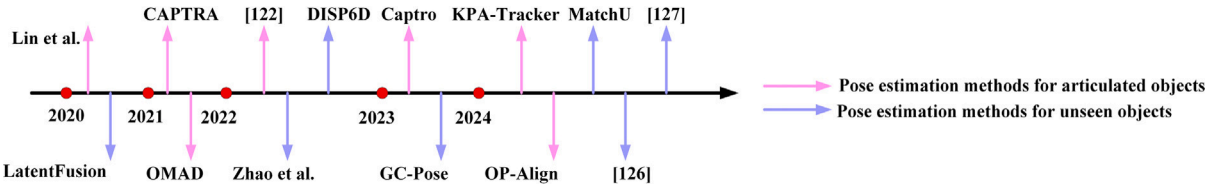

Fig. 15. Timeline overview of pose estimation methods for unseen objects and articulated objects.

others. Fig. 15 shows a time-sequential overview plot of the pose estimation method based on unseen objects and articulated bodies.

[171] proposed the GCPose framework, whose core idea is to establish 3D-3D correspondences between object-scene point clouds and object-model point clouds through the concept of point cloud registration. This allows the estimation of the 6D pose of any unseen object without the need for retraining. LatentFusion [172] is also an approach for pose estimation for unseen objects by reconstructing a potential 3D representation of an object from a reference view using 2D U-Net and 3D transformation units and integrating multi-view features using a view fusion module. Its validity and accuracy of pose estimation are demonstrated on both LINEMOD, ModelNet, and MOPED datasets. MatchU [173] estimates the 6D pose of unseen objects by combining a transformer-based RoITr architecture and a CNN to extract 3D rotation invariant features and 2D local visual features. The method fuses features into the latent space and utilizes the LatentFusion Attention Module to enhance the representation of texture and geometric features. However, it relies on external object localization methods, which may have limitations in practical applications.

DISP6D [174] simplifies the pose estimation task using a pose normalization module and constructs a shape space by contrasting metric learning [175], allowing the network to generalize to new object instances. In the inference phase, the system retrieves object rotations by comparing potential pose codes with the pose code base. [176] solved the problem of 3D orientation estimation of unseen objects in monocular images by multi-scale feature extraction and local similarity computation combined with an adaptive fusion module. The method enhances the key features, reduces the influence of outliers, and introduces a fast retrieval strategy to improve computational efficiency. [177] introduces a target pose estimation method based on diffusion feature aggregation, effectively capturing and aggregating diffusion features of different granularities through three distinct network architectures. Although it has a high computational complexity and may require fine-tuning for specific tasks, its strong generalization and feature representation capabilities enable it to perform well on unseen objects. SAM-6D [178] combines SAM's zero-shot segmentation capability with an innovative pose estimation design, providing an effective solution for zero-shot 6D object pose estimation. Its multi-dimensional matching score and two-stage point matching design significantly improve the accuracy and robustness of segmentation and pose estimation. [179] utilizes visual–language models (VLMs) and natural language descriptions to improve pose estimation accuracy. The architecture uses high-resolution feature maps and effective feature fusion strategies to improve generalization ability and robustness over multiple datasets. However, it relies on depth information and accurate textual descriptions, which may limit its application in some cases. OVE6D [180] solves object pose information through three stages: viewpoint estimation, rotation estimation, and translation estimation. It utilizes a cascaded network architecture that combines a viewpoint encoder and a view codebook to deal with the occlusion problem and capture the object's symmetry. The method performs well on datasets such as T-LESS and LINEMOD with real-time performance. However, the performance of OVE6D relies on accurate 3D models and object segmentation masks, and the generalization ability may be limited when dealing with unseen objects.

[181] introduces an ANCSH normalized representation for different articulated objects. It consists of two levels: the NAOCS at the root level specifies the orientation, scale, and joint state of the object, and the NPCS at the leaf level specifies the pose and scale of each part. And based on the idea of a two-step approach, [181] predicts

Table 4
Pose estimation methods for unseen objects and articulated bodies. We introduced the implementation forms, release years, training inputs, processing categories of different methods, and compared their performance on the T-Less, O-LM, LM, and SAPIEN datasets.

Methods	Years	Data	Type	Evaluation metrics(%)			
				T-LESS	O-LM	LM	SAPIEN
Catogray	2020	D	articulated body	–	–	–	–
CAPTRA	2021	D	articulated body	–	–	–	98.74
OMAD	2021	D	articulated body	–	–	–	–
[122]	2022	RGB-D	articulated body	–	–	–	–
Captro	2023	RGB	articulated body	–	–	–	–
KPA-Tracker	2024	RGB-D	articulated body	–	–	–	–
OP-Align	2024	RGB-D	articulated body	–	–	–	–
LatentFusion	2020	RGB-D	unseen object	–	–	78.0	–
[99]	2022	D	unseen object	78.73	50.34	81.52	–
DISP6D	2022	RGB	unseen object	65.45	–	–	–
OVE6D	2022	RGB-D	unseen object	–	82.5	98.7	–
GC-Pose	2023	D, CAD	unseen object	73.8	74.1	93.6	–
SAM-6D	2024	RGB-D, CAD	unseen object	51.5	69.9	–	–
[126]	2024	RGB	unseen object	71.03	97.8	78.1	–
[127]	2024	RGB-D, text	unseen object	–	–	–	–
MatchU	2024	RGB-D, CAD	unseen object	66.8	68.0	–	–

the joint parameters in NAOCS first and then converts them to the camera space, which accurately estimates the pose and joint parameters of each part of the jointed object and realizes the pose estimation of unknown jointed object instances. [182] achieves part-level pose estimation for multiple instances by combining RGB-D input data with the PointNet++ architecture. It includes stages for object detection, part segmentation, NOCS prediction, and joint attribute prediction, allowing it to handle objects with different motion structures within the same semantic category. In addition, creating the real-world validation dataset ReArtVal offers a new research direction for articulated object pose estimation. Captro [183] is an innovative monocular stereo RGB image reconstruction method capable of handling multiple articulated objects' categories and joint agnostic reconstruction. It infers the 3D shape, 6D pose, size, joint type, and state of an object from a single image via an encoder–decoder architecture. CARTO's strengths lie in its ability to generalize to unseen instances, and its real-time processing capabilities. However, its dependence on shape a priori information limits its ability to generalize to objects with large class differences.

Unlike the above methods, OP-Align [184] employs an object-level and part-level alignment strategy to accurately estimate the pose of articulated objects using point clouds from a single frame, effectively addressing the pose estimation challenges posed by objects with complex shapes. Moreover, it achieves precise pose estimation capabilities through self-supervised learning without manual annotations. OMAD [185] specifies the shape deformation and pose deformation of the target object through linear shape function and nonlinear joint function and predicts the initial shape parameters and joint states through shape branching, joint state branching and key point branching. CAPTRA [186] is a category-level pose-tracking method that handles both rigid and articulated body objects. The method uses a pose normalization module to normalize the input point cloud, simplifying the pose estimation task. The rotation network module then directly regresses the incremental rotations for highly accurate rotation estimation. At the same time, the coordinate network module predicts dense normalized coordinates and segmentations to provide accurate translation and dimension information. KPA-Tracker [187] automatically generates ordered 3D key points through unsupervised learning and utilizes these key points for pose updating between consecutive frames to enable category-level pose estimation and tracking of articulated objects. The method requires no manual annotation and can be generalized to unseen articulated objects. However, it is sensitive to the number of key points, and its performance in complex scenes needs to be improved.

In summary, researchers dealing with the problem of pose estimation for unknown objects often rely on zero-sample learning and meta-learning techniques, which adapt pose estimation for new objects by learning generic features of the object. For articulated bodies, pose estimation methods, on the other hand, combine image and point cloud data and utilize multimodal feature fusion and graph neural networks to capture their complex motion patterns. Although some progress has been made in existing studies, as shown in Table 4, how to effectively cope with the wide range of object shapes and ranges of motion and to overcome occlusion and noise interference are still the main challenges facing the field.

5. Methodological evaluation

5.1. Datasets

Effective datasets are the foundation for advancing research and development in the field of object pose estimation. In recent years, researchers have constructed a variety of representative datasets that not only contain diverse object classes but also cover a variety of challenging factors such as different pose changes, complex backgrounds, occlusions, and noise. With these datasets, researchers are able to train and evaluate various algorithmic models to continuously improve the accuracy and robustness of object pose estimation. In this paper, we will overview the current mainstream instance-level, category-level, and unseen object and articulated body pose estimation datasets and analyze their characteristics, application scenarios, and roles in algorithm development.

5.1.1. Instance-level datasets

The LineMod (LM) dataset is a standardized benchmark dataset for the task of object pose estimation, covering 15 common everyday objects, each of which has a corresponding 3D model and RGB-D images taken from multiple viewpoints. This dataset includes partially occluded objects and complex backgrounds, making it suitable for evaluating the performance of pose estimation algorithms in handling occlusion and complex scenes. The LM dataset provides a unified benchmark for the comparison and evaluation of different algorithms, and it is an important tool that has been widely used in the research of object pose estimation.

The Occlusion LineMod (O-LM) dataset is an extension of the LineMod dataset specifically designed to evaluate the performance of object pose estimation under occlusion conditions. The O-LM dataset contains eight object classes, each of which partially overlaps with the background or with other objects, increasing the challenge of pose estimation. The dataset is widely used to evaluate the robustness and accuracy of pose estimation algorithms in the face of high occlusion situations and is an ideal platform for developing and testing robust algorithms.

The T-LESS dataset is suitable for the task of pose estimation of untextured industrial objects and contains 30 untextured industrial

objects providing RGB-D image data under different lighting conditions and backgrounds. The objects in this dataset are typically highly symmetric and weakly textured, making it suitable for measuring the performance of the algorithm in dealing with industrial environments that are textureless, symmetric, and have complex backgrounds. It also helps researchers to develop pose estimation algorithms for industrial applications, advancing applications in industrial automation and robotics.

The YCB-Video (YCB-V) dataset [188] contains 21 everyday objects from the YCB (Yale-CMU-Berkeley) Objects and Models set, each of which is modeled in 3D and recorded in a real-world scene via RGB-D video. The YCB-V dataset captures the multi-view pose changes of the objects in a variety of backgrounds, lighting conditions, and occlusions. The main applications include evaluating the performance of algorithms for simultaneous pose estimation and tracking of multiple objects in complex dynamic environments, and it is an essential resource for evaluating and advancing the development of 6D pose estimation algorithms.

5.1.2. Category-level datasets

The CAMERA25 dataset is primarily used for training and evaluation of synthetic data to enhance the model's ability to generalize to a variety of object poses and backgrounds. The dataset contains synthetic images of 25 different objects, which are generated to ensure a high level of diversity and realism, allowing the model to learn the features of the objects in a wide range of poses and contexts. Researchers often use the CAMERA25 dataset for initial training of the model, after which the applicability and accuracy of the model in real-world scenarios are further improved by fine-tuning it on real-world data.

The REAL275 dataset contains 275 3D object models of different classes. The dataset consists of multi-view images and corresponding real 6D object pose labels and is designed to provide a high-quality benchmark for training and evaluation. Each object model is captured in multiple environments, including different lighting conditions and backgrounds, enhancing the diversity and challenge of the dataset. The REAL275 dataset is particularly suitable for validating and comparing novel object pose estimation algorithms and their performance in real applications.

The Wild6D dataset [189] focuses on pose estimation of unlabeled objects in real-world scenes. The dataset contains a large number of real images that capture multiple views of various everyday objects in scenes that often have complex backgrounds and lighting variations. The Wild6D dataset is designed to aid in the development and evaluation of 6D pose estimation algorithms for real-world applications, especially those that can handle unseen objects in unknown environments.

5.1.3. Unseen object datasets

OnePose contains a large amount of robot data grabbed through AR tools, including the robot's center position, dimensions around the Z-axis, rotation angle, and camera pose information. The dataset also contains high-resolution images taken from multiple viewpoints, as well as reconstructed sparse point clouds and 2D keypoints. OnePose is primarily used to evaluate the accuracy and robustness of the algorithms when dealing with a single object and is suitable for the task of fine-grained estimation of single-object pose.

The MOPED dataset is designed specifically for multi-object pose estimation tasks, offering a wealth of RGB and depth image data that covers various backgrounds, lighting conditions, and object poses. It includes detailed annotation information such as object location, pose, and category labels. This dataset aims to test and evaluate the ability of algorithms to handle multiple objects in complex environments, particularly in scenarios involving object interactions, occlusions, and overlaps. The MOPED dataset provides an important benchmark and challenge for research in the field of multi-object pose estimation.

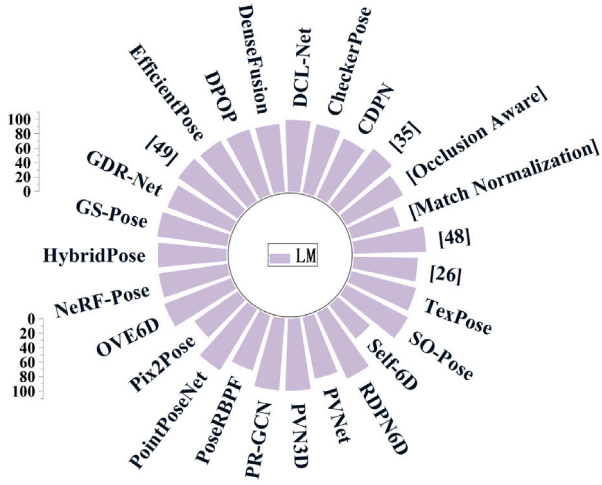

Fig. 16. Performance analysis of the LM dataset.

The SAPIEN dataset [190] is a comprehensive dataset specifically designed for 6D object pose estimation and scene understanding. It includes a wealth of synthetic 3D models and image data and integrates multimodal information such as video, audio, and textual descriptions. This dataset is suitable for developing and evaluating object pose estimation and scene understanding algorithms in complex 3D scenes, providing significant support for research in robotics and autonomous driving systems.

5.2. Metrics

The criteria for evaluating the performance of an algorithm in a 6D pose estimation task are multidimensional, including the algorithm's accuracy, robustness, and generalization ability. Commonly used evaluation metrics include average distance (ADD) [191], translation error [192], rotation error [193], and intersection over union (IOU) [194]. Instance-level approaches tend to use ADD metrics to measure the accuracy of the algorithm in predicting poses across the entire 6D space, whereas category-level approaches focus more on evaluating translational and angular errors. These evaluation metrics provide researchers with a comprehensive set of benchmarks for evaluating and comparing different pose estimation algorithms and guiding the direction of algorithm design to better suit specific application requirements. Fig. 16, 17, 18, 19, 20, 21, and 22 show the pose estimation performance of various algorithms on different datasets (LM, REAL275, CAMERA25, O-LM, YCB-Video) and evaluation metrics (IOU, n° ncm). Each sector represents an algorithm, with the length from the center to the outer edge indicating the performance score. A higher score indicates a better performance of the algorithm at that distance.

ADD metrics are a common method used to assess the accuracy of symmetry-free rigid body objects in object pose estimation tasks. Given the true pose of the object (R, T) and the estimated pose $(\tilde{R}, \tilde{T})$, ADD calculates the average Euclidean distance between the estimated and the true positions for each 3D model point x, as shown in Eq. (1). Where M represents the set of points in the target 3D model, and m represents the number of points. If the average distance is less than some predefined threshold (e.g., 10% of the object diameter), the estimate is considered correct. This metric is simple and intuitive and is suitable for pose estimation tasks that require high accuracy.

$$\mathrm{ADD} = \frac{1}{m} \sum_{x \in M} \left\| (R_x + T) - (\tilde{R}_x + \tilde{T}) \right\| \tag{1}$$

However, the ADD metric is prone to the problem of multiple solutions when dealing with symmetric objects, which leads to inaccurate error calculation. Therefore, researchers propose to use the ADD-S metric to

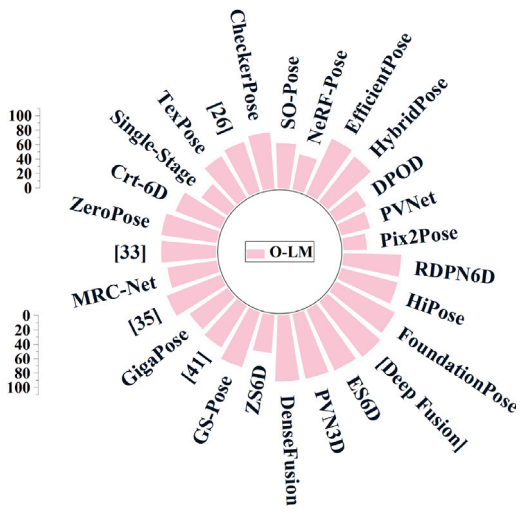

Fig. 17. Performance analysis of the O-LM dataset.

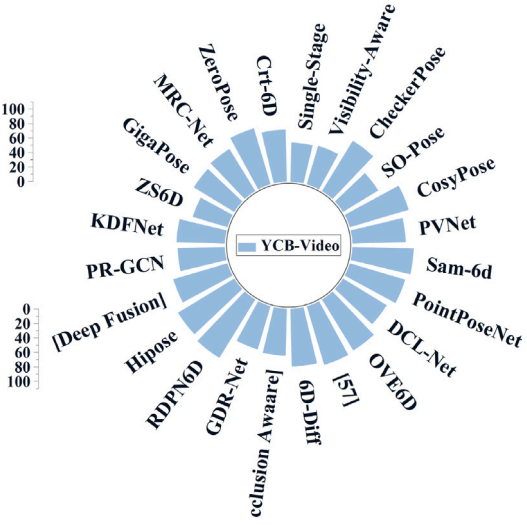

Fig. 18. Performance analysis of the YCB-Video dataset.

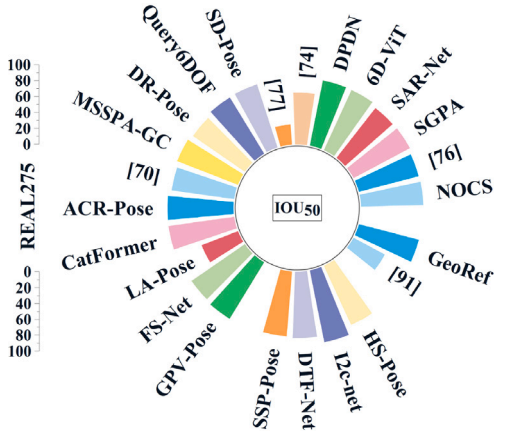

Fig. 19. Performance analysis of the Real275 dataset (IOU_{50}).

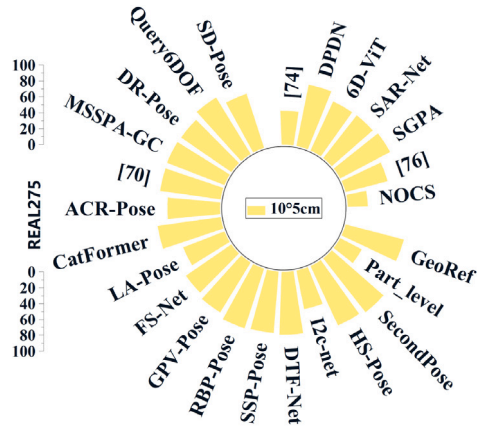

Fig. 20. Performance analysis of the Real275 dataset (IOU_{50}).

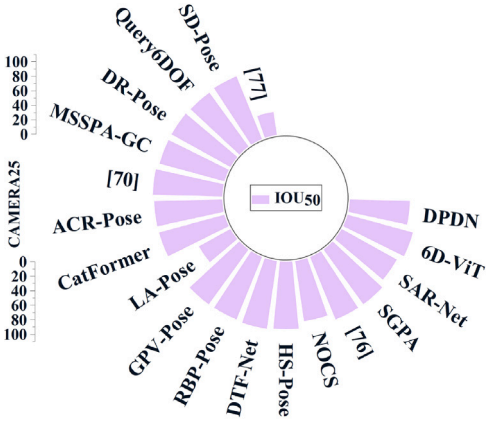

Fig. 21. Performance analysis of the CAMERA25 dataset (IOU_{50}).

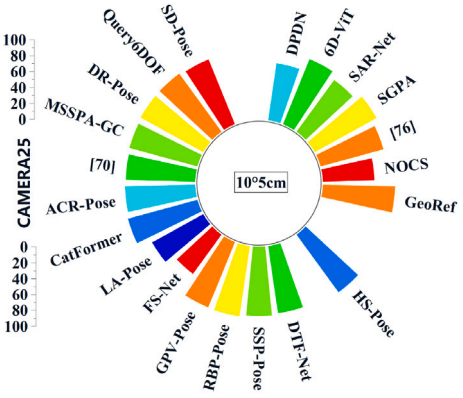

Fig. 22. Performance analysis of the CAMERA25 dataset (10° 5cm).

measure the performance of symmetric objects. Different from the ADD metric, ADD-S evaluates the pose estimation accuracy by minimizing the distance between the estimated and the model points of the object in the real pose, as shown in Eq. (2). If this distance is less than a set threshold, then the pose estimation is correct.

$$\text{ADD-S} = \frac{1}{m}\sum_{x_1 \in M} \min_{x_2 \in M} \left\| (R_x + T) - (\tilde{R}_x + \tilde{T}) \right\| \tag{2}$$

Where j denotes the index of the point in the true pose.

Unlike the point clouds generated by LiDAR, the point cloud data obtained from depth map conversion often contains a certain degree of

noise, so many methods propose a point cloud reconstruction process. The main advantage of the chamfer distance as a measure of the similarity between two point clouds is that it takes into account the mutual distance [195] between the two point clouds, enabling a better measure of the similarity between the reconstruction results and the target in pose estimation tasks. Given two point clouds, P and Q, point cloud P contains the point set $\{P_i\}_{i=1}^{m}$, and point cloud Q contains the point set $\{Q_j\}_{j=1}^{n}$. Then the minimum distance from point cloud P to point cloud Q and the minimum distance from point cloud Q to point cloud P can be expressed as:

$$d(p_i, Q) = min_{q_j \in Q} \left\| p_i - q_j \right\| \tag{3}$$

$$d(q_j, P) = min_{p_i \in P} \left\| q_j - p_i \right\| \tag{4}$$

The chamfer distance is the sum of the average of all the minimum distances in the above two steps:

$$CD(P, Q) = \frac{1}{m} \sum_{i=1}^{m} d(p_i, Q) + \frac{1}{n} \sum_{j=1}^{n} d(q_j, P) \tag{5}$$

Translation and rotation errors are important metrics for category-level object pose estimation tasks to measure the difference between the predicted pose and the true pose. The translation error reflects the difference in distance between the estimated and actual position of the object center, usually expressed as a Euclidean distance (e.g., centimeters), with smaller errors indicating higher spatial localization accuracy. Rotational error, on the other hand, measures the accuracy of the object's orientation during pose estimation, usually expressed in terms of angle, with smaller errors indicating more accurate orientation estimates. Common translation and rotation errors include 5° 2cm, 5° 5cm, 10° 2cm, 10° 5cm, etc. These error metrics help evaluate the accuracy and robustness of the algorithms in different poses to ensure accuracy and stability in real-world applications.

Intersection over Union (IOU) is also one of the commonly used evaluation metrics for category-level object pose estimation, used to measure the overlap between the predicted 3D bounding box (B_{pred}) and the ground truth 3D bounding box (B_{true}). The value of IOU ranges from 0 to 1, with higher values indicating a greater overlap between the predicted and ground truth bounding boxes, meaning better prediction accuracy. Its calculation formula is shown in Eq. (6).

$$IOU = \frac{(B_{true} \cap B_{pred})}{(B_{true} \cup B_{pred})} \tag{6}$$

Here, the symbols ∩ and ∪ represent the intersection and union, respectively. Commonly used IOU metrics include IOU_{25}, IOU_{50}, and IOU_{75}. IOU_{25} indicates that the overlap area is at least 25% of the total area, making it suitable for preliminary evaluations of the model's capability, especially when detecting larger targets. IOU_{50} requires at least 50% overlap and is used to assess the model's basic accuracy in real-world applications. IOU_{75} is designed for tasks that demand high-precision detection. By using different IOU thresholds, researchers can gain a comprehensive understanding of the model's performance at low, medium, and high accuracy, optimize it for specific application scenarios, and evaluate its robustness in handling complex situations, occlusions, and varied object poses.

The BOP metric provides a unified and standardized evaluation framework for object pose estimation algorithms, allowing for direct comparison of the performance of different algorithms. First, the Visible Surface Discrepancy (VSD) measures the difference between the visible surface of the estimated object pose and the real object's surface, typically used to reflect whether the estimated object's surface matches the visible part of the real object. It can be expressed as Eq. (7).

$$e_{\text{VSD}}(\tilde{D}, D, \tilde{V}, V, \tau) = \text{avg} \begin{cases} 0, & p \in V \wedge |\tilde{D}(p) - D(p)| < \tau \\ 1, & \text{otherwise} \end{cases} \tag{7}$$

Where $\tilde{D}$ represents the distance from the camera center to the 3D projection point after estimating the object model, and D represents the true distance. $\tilde{V}$ and V are the visibility masks obtained by comparing the estimated distance with the true distance. τ is the maximum allowable difference in camera distance for overlapping pixels. From Eq. (7), it can be seen that if the surface discrepancy of the same object in two different poses is consistent, meaning the object looks exactly the same in both poses, then e_{VSD} will be 0.

Then, the Maximum Symmetry-Aware Surface Distance (MSSD) takes into account the importance of object symmetry, especially when the object has significant symmetry. This metric quantifies the maximum distance between the estimated pose's surface and the real surface, while utilizing symmetry for optimization, as shown in Eq. (8).

$$e_{\text{MSSD}}(\tilde{P}, P, S_M, V_M) = \min_{S \in S_M} \max_{V \in V_M} \left\| \tilde{P}_x - PS_x \right\|_2 \tag{8}$$

Where $\tilde{P}$ is the predicted transformation matrix; P is the true transformation matrix; S_M is the set of global transformation matrices of the object model; and V_M is the set of vertices of the object model. The smaller the value of e_{MSSD}, the closer the predicted pose is to the true pose.

The Maximum Symmetry-Aware Projection Distance (MSPD) measures the maximum difference between the estimated object's pose and the real object in the projection plane, especially when considering the symmetry of the object. This metric is particularly useful for handling projection differences of objects in the visual plane, and its formula is shown in Eq. (9).

$$e_{MSPD}(\tilde{P}, P, S_M, V_M) = min_{S \epsilon S_M} max_{x \epsilon V_M} \left\| proj(\tilde{P}_x) - proj(PS_x) \right\|_2 \tag{9}$$

Finally, the Average Recall (AR) is a composite result of VSD, MSSD, and MSPD, used to evaluate the overall performance of pose estimation.

5.3. Data quality optimization and pose estimation

In object pose estimation tasks, the quality of input data has a decisive impact on model performance. Whether for instance-level or category-level methods, high-quality input data is a key factor in improving model performance. Currently, instance-level object pose estimation methods mainly rely on RGB images and CAD models of objects, while category-level methods more often use RGB images and point cloud information.

For RGB images, object detection and segmentation algorithms like Mask R-CNN have been widely applied in the data preprocessing stage. Mask R-CNN, through instance segmentation technology, can accurately extract the mask information of the target object, thus effectively separating the object from a complex background and providing more accurate data input for pose estimation. Additionally, some studies have adopted multi-view feature fusion strategies, where image features from different viewpoints are combined and weighted to enhance the geometric feature representation of the target object. [196] used a Generative Adversarial Network (GAN) to handle occluded objects, where the network can restore missing information in the occluded parts, improving the accuracy and robustness of pose estimation. Pose estimation in low-light conditions is an important challenge in object pose estimation. Low light typically leads to increased image noise, decreased contrast, and loss of detail, making it difficult for the model to extract accurate feature information from the image. To address this problem, [197] proposed a low-light preprocessing module that uses a restoration branch network and skip connections to transform and refine the extracted features, restoring a clearer final image. [198] used a visual appearance enhancement algorithm, employing a U-Net-like structure [199], and applied feature map addition and multiplication fusion methods to merge color and texture information, revealing clearer detailed information.

Unlike point cloud data captured by LiDAR, category-level methods typically generate point cloud data from depth maps and camera intrinsic/extrinsic parameters. However, this data often contains a

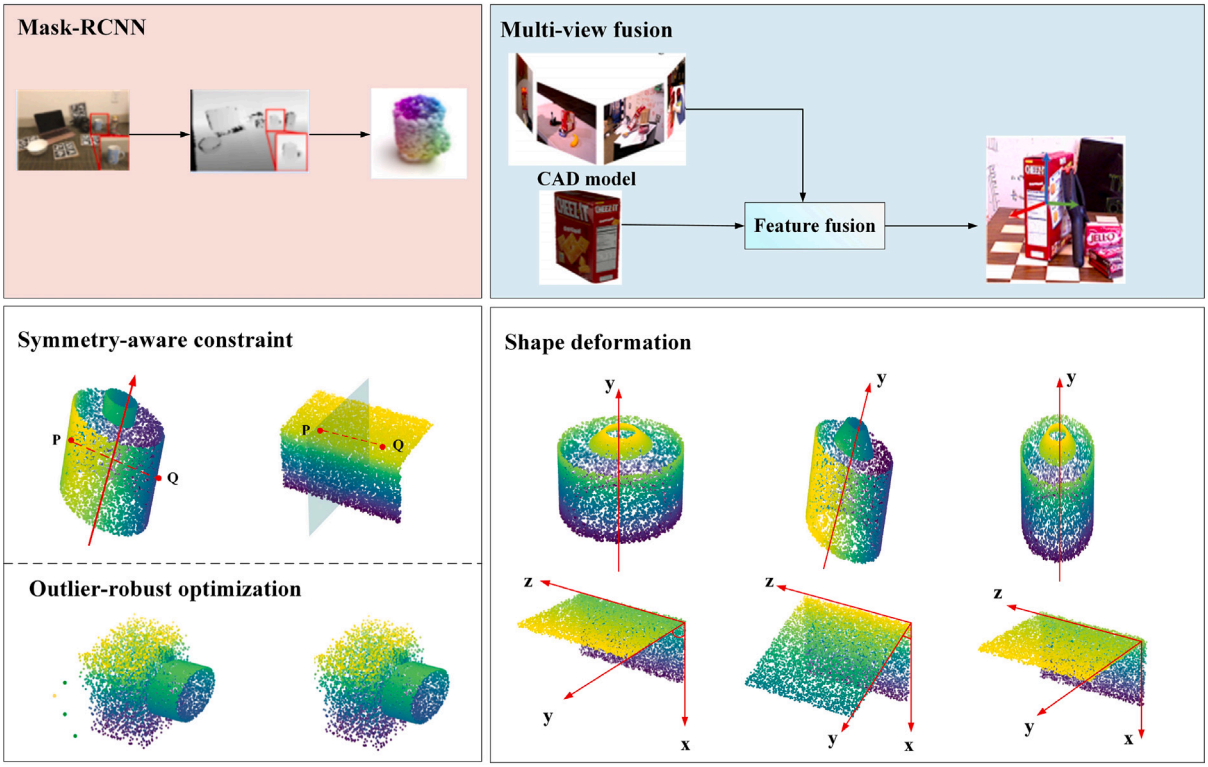

Fig. 23. Data Quality Optimization and Robustness Enhancement Methods for Object Pose Estimation. The figure illustrates important data preprocessing and data augmentation methods in object pose estimation. For example: instance segmentation, feature fusion, point cloud shape deformation, and physical geometric constraints.

significant amount of noise, which may adversely affect the stability and pose estimation accuracy of the model. To reduce the impact of noise, many methods use 3D point cloud reconstruction techniques to rebuild and optimize the input data. For instance, [200] designed an outlier-robust optimizer in the point cloud reconstruction process, effectively removing the interference of outliers and providing cleaner input for pose estimation. [201] considered the symmetry of many objects in real life and introduced object symmetry geometric perception constraints when reconstructing point clouds, ensuring that the reconstructed object maintains inter-class similarity. Furthermore, category-level methods aim to estimate the poses of different instances within the same category, but objects within the same category have varying sizes, shapes, and appearances. This makes it impossible for models to rely on simple shape templates or generic features to estimate poses, as feature variations across different instances may lead to pose estimation biases. To solve this problem, FS-Net applied a linear shape deformation method to proportionally stretch and compress the input point cloud to simulate instances with different shapes. SSP-Pose proposed a nonlinear shape deformation module that performs nonlinear transformations along the object's symmetry axis to generate new instances with different shapes, while using the object's symmetry properties to guide the geometric deformation process, ensuring the deformed object retains symmetry.

In summary, current technologies have made significant progress in addressing the data quality issues in object pose estimation, including advanced data preprocessing techniques, feature fusion and enhancement strategies, noise handling methods, and shape deformation approaches, all of which significantly improve model performance and robustness. The Fig. 23 below presents an overall framework, summarizing how different data quality optimization methods are implemented in the task.

6. Existing challenges and future prospects

In recent years, more and more researchers have applied deep learning techniques to the task of object pose estimation and significantly advanced the field through strategies such as innovative network structures, multi-loss function optimization, integration of attention mechanisms, and multimodal data fusion. These techniques have proved their extensive practical value in several application scenarios. However, with the expansion of application scenarios, how to effectively combine the pose estimation techniques with other perception techniques, such as visual recognition and sensor fusion, to achieve a more comprehensive environment understanding and stronger decision-making capabilities has become a key direction for future research. In this paper, we will further delve into the current challenges and future trends in the field of object pose estimation.

6.1. Main challenges

Despite significant progress in object pose estimation technologies and their extensive practical value, several challenges remain. First, occlusion is a major obstacle, with both partial and complete occlusions greatly increasing the difficulty of estimation. Second, textureless objects often lack distinctive feature points or edge information in images, making it challenging for local feature-based detection and matching algorithms to identify and track these objects. Additionally, symmetric objects frequently exhibit geometric invariance, which often leads to ambiguity issues with traditional feature-matching methods. Effectively combining geometric constraints and texture information and developing more robust algorithms to accurately discriminate and estimate the pose of symmetric objects represents a key challenge in current research. Finally, real-time performance and computational efficiency are critical bottlenecks in practical applications. High-precision algorithms perform well in controlled environments but struggle to meet real-time requirements in resource-constrained scenarios.

6.2. Future outlook

By comparing the object pose estimation methods in recent years and the existing challenges, we can foresee a series of emerging trends

and development directions that will drive the field to realize new breakthroughs and play a more important role in multiple industries. Future research should include the following areas.

(1) Training using synthetic data. BB8 [202] and YOLO6D [203] train by combining target detection and pose estimation into a unified network framework using real data. The trained results tend to be better when the real images are sufficiently large and fully cover the different bit positions of the object. However, obtaining large-scale real labeled data is costly and time-consuming. Synthetic data, on the other hand, can be computer-generated, which saves the time and cost of data collection and labeling and covers a variety of scenarios, angles, lighting, and occlusions to improve the generalization ability and robustness of the model. In addition, synthetic data are accurately labeled without labeling errors, which is especially important for pose estimation tasks that rely on high-precision labeling.

(2) Perform category-level object pose estimation. The ultimate goal of the object pose estimation domain is to deploy to the real world, so a higher generalizability of the model is required. Unlike instance-level methods, category-level pose estimation methods can be trained for a given category and do not require a CAD model, resulting in higher generalizability to new objects of the same class. Researchers have continuously improved the performance of category-level object pose estimation methods by introducing attention mechanisms, establishing a canonical coordinate space, and using more robust network architectures. In addition, self-supervised or unsupervised learning-based methods have demonstrated strong competitiveness.

(3) Pose estimation for transparent objects and articulated bodies. Current object pose estimation mainly focuses on ordinary objects such as rigid bodies, while less research has been done on transparent and articulated bodies. Transparent objects are prone to distortion of image information due to optical properties, such as refraction and reflection, while articulated bodies require pose estimation systems to have high-dimensional data processing capability and real-time performance due to their complex motion patterns and multi-degree-of-freedom joints. Although there have been some research attempts to address these problems, overall, research on these two types of objects is still in the exploratory stage, and more innovative approaches and practical validation are needed to improve performance.

(4) Training using multi-view and multi-modal data. The fusion of multi-view images and point cloud data can provide more spatial information in complex or occluded environments and significantly improve the object pose estimation accuracy. The fusion of multimodal data, such as visual, depth, or semantic information, can compensate for the limitations of a single sensor and improve the robustness of the system. Meanwhile, the approach of combining natural language modeling with object pose estimation enhances the understanding of object pose by converting natural language descriptions into visual commands and fusing linguistic and visual features in a multimodal learning framework.

(5) Developing more lightweight network models. Traditional complex models have high computational and storage demands, making real-time processing difficult. Therefore, it is necessary to design smaller network structures and lightweight models with low computational complexity using techniques such as model compression, pruning, quantization, and knowledge distillation. In addition, the optimized feature extraction algorithms and convolution operations can significantly reduce computational costs, thereby expanding the potential applications of lightweight models in object pose estimation.

7. Conclusion

This paper reviews the latest research advancements in deep learning-based object pose estimation, covering a wide range of methods from instance-level to category-level, as well as approaches for unseen objects and articulated objects. Despite challenges such as occlusion, weak texture, and pose ambiguity in symmetric objects, researchers have significantly improved pose estimation accuracy and robustness through innovative network architectures, multi-modal data fusion, and self-supervised learning strategies. In the future, with the widespread use of synthetic data, enhanced generalization capabilities at the category level, and further research into special objects such as transparent and articulated objects, object pose estimation technology is expected to achieve broader applications. At the same time, the development of lightweight network models will further drive the deployment of this technology on resource-constrained devices, laying the foundation for more efficient and practical object pose estimation systems.

CRediT authorship contribution statement

Jing Wang: Writing – original draft, Supervision, Investigation, Conceptualization. **Guohan Liu:** Writing – original draft, Supervision, Methodology, Investigation. **Wenxin Ding:** Visualization, Supervision, Methodology, Conceptualization. **Yuying Li:** Conceptualization, Supervision, Methodology, Investigation. **Wanying Song:** Resources, Supervision, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the Natural Science Foundation of China (61901358) and the Natural Science Basis Research Plan in Shaanxi Province of China (2025JC-YBMS-672).

Data availability

No data was used for the research described in the article.

References

[1] H. Zhou, J. Liu, Z. Liu, Y. Liu, X. Wang, Rotate-and-render: Unsupervised photorealistic face rotation from single-view images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5911–5920.

[2] K. Park, T. Patten, J. Prankl, M. Vincze, Multi-task template matching for object detection, segmentation and pose estimation using depth images, in: 2019 International Conference on Robotics and Automation, ICRA, IEEE, 2019, pp. 7207–7213.

[3] J. Cheng, P. Liu, Q. Zhang, H. Ma, F. Wang, J. Zhang, Real-time and efficient 6-D pose estimation from a single RGB image, IEEE Trans. Instrum. Meas. 70 (2021) 1–14.

[4] Y. Li, H. Wang, L.M. Dang, T.N. Nguyen, D. Han, A. Lee, I. Jang, H. Moon, A deep learning-based hybrid framework for object detection and recognition in autonomous driving, IEEE Access 8 (2020) 194228–194239.

[5] Y. Su, J. Rambach, N. Minaskan, P. Lesur, A. Pagani, D. Stricker, Deep multi-state object pose estimation for augmented reality assembly, in: 2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct, ISMAR-Adjunct, IEEE, 2019, pp. 222–227.

[6] G. Niu, Q. Yang, Y. Gao, M.-O. Pun, Vision-based autonomous landing for unmanned aerial and ground vehicles cooperative systems, IEEE Robot. Autom. Lett. 7 (3) (2021) 6234–6241.

[7] T. Yin, X. Zhou, P. Krahenbuhl, Center-based 3d object detection and tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11784–11793.

[8] G. Terzakis, M. Lourakis, A consistently fast and globally optimal solution to the perspective-n-point problem, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, Springer, 2020, pp. 478–494.

[9] T. Cao, F. Luo, Y. Fu, W. Zhang, S. Zheng, C. Xiao, DGECN: A depth-guided edge convolutional network for end-to-end 6D pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3783–3792.

[10] C. Liu, F. Chen, L. Deng, R. Yi, L. Zheng, C. Zhu, J. Wang, K. Xu, 6DOF pose estimation of a 3D rigid object based on edge-enhanced point pair features, Comput. Vis. Media 10 (1) (2024) 61–77.

[11] G. Zhai, X. Min, Perceptual image quality assessment: a survey, Sci. China Inf. Sci. 63 (2020) 1–52.

[12] X. Min, H. Duan, W. Sun, Y. Zhu, G. Zhai, Perceptual video quality assessment: A survey, Sci. China Inf. Sci. 67 (11) (2024) 211301.

[13] X. Min, K. Gu, G. Zhai, X. Yang, W. Zhang, P. Le Callet, C.W. Chen, Screen content quality assessment: Overview, benchmark, and beyond, ACM Comput. Surv. 54 (9) (2021) 1–36.

[14] Q. Zheng, Y. Fan, L. Huang, T. Zhu, J. Liu, Z. Hao, S. Xing, C.-J. Chen, X. Min, A.C. Bovik, et al., Video quality assessment: A comprehensive survey, 2024, arXiv preprint arXiv:2412.04508.

[15] A. Dhillon, G.K. Verma, Convolutional neural network: a review of models, methodologies and applications to object detection, Prog. Artif. Intell. 9 (2) (2020) 85–112.

[16] B. Koonce, B.E. Koonce, Convolutional neural networks with swift for tensorflow: Image recognition and dataset categorization, Springer, 2021.

[17] C. Gao, J. Zhu, F. Zhang, Z. Wang, X. Li, A novel representation learning for dynamic graphs based on graph convolutional networks, IEEE Trans. Cybern. 53 (6) (2022) 3599–3612.

[18] S. Mei, Y. Geng, J. Hou, Q. Du, Learning hyperspectral images from RGB images via a coarse-to-fine CNN, Sci. China Inf. Sci. 65 (2022) 1–14.

[19] K. Guo, H. Ye, X. Gao, H. Chen, An accurate and robust method for absolute pose estimation with uav using RANSAC, Sensors 22 (15) (2022) 5925.

[20] K. Park, T. Patten, M. Vincze, Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7668–7677.

[21] X. Zhao, Q. Li, Y. Chao, Q. Wang, Z. He, D. Liang, RT-less: a multi-scene RGB dataset for 6D pose estimation of reflective texture-less objects, Vis. Comput. 40 (8) (2024) 5187–5200.

[22] D. Zhang, A. Barbot, F. Seichepine, F.P.-W. Lo, W. Bai, G.-Z. Yang, B. Lo, Micro-object pose estimation with sim-to-real transfer learning using small dataset, Commun. Phys. 5 (1) (2022) 80.

[23] J. Wang, X. Chen, W. Jiang, L. Hua, J. Liu, H. Sui, Pvnet: A novel semantic segmentation model for extracting high-quality photovoltaic panels in large-scale systems from high-resolution remote sensing imagery, Int. J. Appl. Earth Obs. Geoinf. 119 (2023) 103309.

[24] C. Song, J. Song, Q. Huang, Hybridpose: 6d object pose estimation under hybrid representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 431–440.

[25] F. Li, S.R. Vutukur, H. Yu, I. Shugurov, B. Busam, S. Yang, S. Ilic, Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2123–2133.

[26] A. Kanezaki, Y. Matsushita, Y. Nishida, Rotationnet for joint object categorization and unsupervised pose estimation from multi-view images, IEEE Trans. Pattern Anal. Mach. Intell. 43 (1) (2019) 269–283.

[27] Y. Di, F. Manhardt, G. Wang, X. Ji, N. Navab, F. Tombari, So-pose: Exploiting self-occlusion for direct 6d pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 12396–12405.

[28] R.L. Haugaard, T.M. Iversen, Multi-view object pose estimation from correspondence distributions and epipolar geometry, in: 2023 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2023, pp. 1786–1792.

[29] Y. Sun, S. Dai, J. Dang, J. Yong, A 6D object pose estimation method combining self-attention mechanism, in: 2024 5th International Conference on Computer Engineering and Application, ICCEA, IEEE, 2024, pp. 1315–1319.

[30] D. Groos, H. Ramampiaro, E.A. Ihlen, EfficientPose: Scalable single-person pose estimation, Appl. Intell. 51 (4) (2021) 2518–2533.

[31] R. Ye, Y. Ren, X. Zhu, Y. Wang, M. Liu, L. Wang, An efficient pose estimation algorithm for non-cooperative space objects based on dual-channel transformer, Remote. Sens. 15 (22) (2023) 5278.

[32] Y. Labbé, J. Carpentier, M. Aubry, J. Sivic, Cosypose: Consistent multi-view multi-object 6d pose estimation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16, Springer, 2020, pp. 574–591.

[33] S. Zakharov, I. Shugurov, S. Ilic, Dpod: 6d pose object detector and refiner, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1941–1950.

[34] J. Hu, P.R. Pagilla, View planning for object pose estimation using point clouds: An active robot perception approach, IEEE Robot. Autom. Lett. 7 (4) (2022) 9248–9255.

[35] D. Wang, G. Zhou, Y. Yan, H. Chen, Q. Chen, GeoPose: Dense reconstruction guided 6D object pose estimation with geometric consistency, IEEE Trans. Multimed. 24 (2021) 4394–4408.

[36] R. Lian, H. Ling, Checkerpose: Progressive dense keypoint localization for object pose estimation with graph neural network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14022–14033.

[37] J. Mei, X. Jiang, H. Ding, Spatial feature mapping for 6dof object pose estimation, Pattern Recognit. 131 (2022) 108835.

[38] Y. Lin, Y. Su, S. Inuganti, Y. Di, N. Ajilforoushan, H. Yang, Y. Zhang, J. Rambach, Resolving symmetry ambiguity in correspondence-based methods for instance-level object pose estimation, 2024, arXiv preprint arXiv:2405.10557.

[39] W. Chen, X. Jia, H.J. Chang, J. Duan, A. Leonardis, 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, 2020, pp. 4233–4242.

[40] R. Lian, H. Ling, Visibility-aware keypoint localization for 6dof object pose estimation, 2024, arXiv preprint arXiv:2403.14559.

[41] H.-Y. Peng, J.-P. Zhang, M.-H. Guo, Y.-P. Cao, S.-M. Hu, Charactergen: Efficient 3d character generation from single images with multi-view pose canonicalization, ACM Trans. Graph. 43 (4) (2024) 1–13.

[42] Z. Li, G. Wang, X. Ji, 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, 2019, pp. 7678–7687.

[43] S. Yu, D.-H. Zhai, Y. Xia, Synthetic depth image-based category-level object pose estimation with effective pose decoupling and shape optimization, IEEE Trans. Instrum. Meas. (2024).

[44] H. Chen, F. Manhardt, N. Navab, B. Busam, Texpose: Neural texture learning for self-supervised 6d object pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4841–4852.

[45] S.R. Vutukur, H. Brock, B. Busam, T. Birdal, A. Hutter, S. Ilic, Nerf-feat: 6D object pose estimation using feature rendering, in: 2024 International Conference on 3D Vision (3DV), IEEE, 2024, pp. 1146–1155.

[46] P. Castro, T.-K. Kim, Crt-6d: Fast 6d object pose estimation with cascaded refinement transformers, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 5746–5755.

[47] Q. Huang, Z. Jie, L. Ma, L. Shen, S. Lai, A pyramid fusion MLP for dense prediction, IEEE Trans. Image Process. (2025).

[48] H. Chen, P. Wang, F. Wang, W. Tian, L. Xiong, H. Li, Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2781–2790.

[49] S.R. Vutukur, M. Ba, B. Busam, M. Kayser, G. Singh, SABER-6D: Shape representation based implicit object pose estimation, 2024, arXiv preprint arXiv:2408.05867.

[50] Y. Li, Y. Mao, R. Bala, S. Hadap, Mrc-net: 6-dof pose estimation with multiscale residual correlation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 10476–10486.

[51] J. Chen, M. Sun, T. Bao, R. Zhao, L. Wu, Z. He, Zeropose: Cad-model-based zero-shot pose estimation, 2023, arXiv preprint arXiv:2305.17934.

[52] Y. Lin, Y. Su, S. Inuganti, Y. Di, N. Ajilforoushan, H. Yang, Y. Zhang, J. Rambach, Resolving symmetry ambiguity in correspondence-based methods for instance-level object pose estimation, 2024, arXiv preprint arXiv:2405.10557.

[53] Y. Lin, Y. Su, S. Inuganti, Y. Di, N. Ajilforoushan, H. Yang, Y. Zhang, J. Rambach, Resolving symmetry ambiguity in correspondence-based methods for instance-level object pose estimation, 2024, arXiv preprint arXiv:2405.10557.

[54] Y. Hu, P. Fua, W. Wang, M. Salzmann, Single-stage 6d object pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2930–2939.

[55] D. Maji, S. Nagori, M. Mathew, D. Poddar, YOLO-6D-pose: Enhancing YOLO for single-stage monocular multi-object 6d pose estimation, in: 2024 International Conference on 3D Vision (3DV), IEEE, 2024, pp. 1616–1625.

[56] Y. Wu, M. Greenspan, Learning better keypoints for multi-object 6dof pose estimation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 564–574.

[57] X. Liu, Y. Zou, T. Che, P. Ding, P. Jia, J. You, B. Kumar, Conservative wasserstein training for pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8262–8272.

[58] V.N. Nguyen, Y. Hu, Y. Xiao, M. Salzmann, V. Lepetit, 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, 2022, pp. 6771–6780.

[59] P. Ausserlechner, D. Haberger, S. Thalhammer, J.-B. Weibel, M. Vincze, Zs6d: Zero-shot 6d object pose estimation using vision transformers, in: 2024 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2024, pp. 463–469.

[60] V.N. Nguyen, T. Groueix, M. Salzmann, V. Lepetit, Gigapose: Fast and robust novel object pose estimation via one correspondence, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 9903–9913.

[61] I. Shugurov, F. Li, B. Busam, S. Ilic, Osop: A multi-stage one shot object pose estimation framework, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 6835–6844.

[62] X. Deng, A. Mousavian, Y. Xiang, F. Xia, T. Bretl, D. Fox, PoseRBPF: A rao–blackwellized particle filter for 6-D object pose tracking, IEEE Trans. Robot. 37 (5) (2021) 1328–1342.

[63] G. Lee, J.-S. Kim, S. Kim, K. Kim, 6D object pose estimation using a particle filter with better initialization, IEEE Access 11 (2023) 11451–11462.

[64] Q. Luo, T.-B. Xu, F. Liu, T. Li, Z. Wei, Learning shared template representation with augmented feature for multi-object pose estimation, Neural Netw. 176 (2024) 106352.

[65] D. Cai, J. Heikkilä, E. Rahtu, Gs-pose: Cascaded framework for generalizable segmentation-based 6d object pose estimation, 2024, arXiv preprint arXiv:2403.10683.

[66] W. Li, H. Xu, J. Huang, H. Jung, P.K. Yu, N. Navab, B. Busam, GCE-pose: Global context enhancement for category-level object pose estimation, 2025, arXiv preprint arXiv:2502.04293.

[67] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, J. Sun, 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, 2020, pp. 11632–11641.

[68] D.-C. Hoang, J.A. Stork, T. Stoyanov, Voting and attention-based pose relation learning for object pose estimation from 3D point clouds, IEEE Robot. Autom. Lett. 7 (4) (2022) 8980–8987.

[69] C. Yang, H. Guo, [Retracted] a method of image semantic segmentation based on PSPNet, Math. Probl. Eng. 2022 (1) (2022) 8958154.

[70] G. Zhou, Y. Yan, D. Wang, Q. Chen, A novel depth and color feature fusion framework for 6d object pose estimation, IEEE Trans. Multimed. 23 (2020) 1630–1639.

[71] X. Liu, S. Iwase, K.M. Kitani, Kdfnet: Learning keypoint distance field for 6d object pose estimation, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2021, pp. 4631–4638.

[72] T. Zhang, Y. Yang, Y. Zeng, Y. Zhao, Cognitive template-clustering improved linemod for efficient multi-object pose estimation, Cogn. Comput. 12 (4) (2020) 834–843.

[73] N. Mo, W. Gan, N. Yokoya, S. Chen, Es6d: A computation efficient and symmetry-aware 6d pose regression framework, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 6718–6727.

[74] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, S. Savarese, Densefusion: 6d object pose estimation by iterative dense fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3343–3352.

[75] F. Gao, Q. Li, Q. Sun, Improving 6D object pose estimation based on semantic segmentation, in: 2023 IEEE International Conference on Systems, Man, and Cybernetics, SMC, IEEE, 2023, pp. 3763–3769.

[76] Y. An, D. Yang, M. Song, HFT6D: Multimodal 6D object pose estimation based on hierarchical feature transformer, Measurement 224 (2024) 113848.

[77] A. Kashefi, PointNet with KAN versus PointNet with MLP for 3D classification and segmentation of point sets, 2024, arXiv preprint arXiv:2410.10084.

[78] J. Zhou, K. Chen, L. Xu, Q. Dou, J. Qin, 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, 2023, pp. 13967–13977.

[79] G. Zhou, H. Wang, J. Chen, D. Huang, 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, 2021, pp. 2793–2802.

[80] B. Wen, W. Yang, J. Kautz, S. Birchfield, Foundationpose: Unified 6d pose estimation and tracking of novel objects, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 17868–17879.

[81] J. Park, N.I. Cho, Dprost: Dynamic projective spatial transformer network for 6d pose estimation, in: European Conference on Computer Vision, Springer, 2022, pp. 363–379.

[82] Y. Lin, Y. Su, P. Nathan, S. Inuganti, Y. Di, M. Sundermeyer, F. Manhardt, D. Stricker, J. Rambach, Y. Zhang, Hipose: Hierarchical binary surface encoding and correspondence pruning for rgb-d 6dof object pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 10148–10158.

[83] Y. Lu, M. Jiang, Z. Liu, X. Mu, Dual-branch adaptive attention transformer for occluded person re-identification, Image Vis. Comput. 131 (2023) 104633.

[84] Z.-W. Hong, Y.-Y. Hung, C.-S. Chen, RDPN6D: Residual-based dense point-wise network for 6dof object pose estimation based on RGB-D images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 5251–5260.

[85] L. Xu, H. Qu, Y. Cai, J. Liu, 6D-diff: A keypoint diffusion framework for 6d object pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 9676–9686.

[86] G. Wang, F. Manhardt, F. Tombari, X. Ji, 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, 2021, pp. 16611–16621.

[87] J. Richter-Klug, P. Mania, G. Kazhoyan, M. Beetz, U. Frese, Improving object pose estimation by fusion with a multimodal prior–utilizing uncertainty-based CNN pipelines for robotics, IEEE Robot. Autom. Lett. 7 (2) (2022) 2282–2288.

[88] G. Wang, F. Manhardt, J. Shao, X. Ji, N. Navab, F. Tombari, Self6d: Self-supervised monocular 6d object pose estimation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, Springer, 2020, pp. 108–125.

[89] G. Wang, F. Manhardt, X. Liu, X. Ji, F. Tombari, Occlusion-aware self-supervised monocular 6D object pose estimation, IEEE Trans. Pattern Anal. Mach. Intell. 46 (3) (2021) 1788–1803.

[90] Z. Zhang, Q. Ma, Y. Zhang, Z. Chen, J. Chen, H. Zheng, InterTeach: A novel approach for semi-supervised medical image segmentation using cooperative teacher-student networks, IEEE Trans. Circuits Syst. Video Technol. (2024).

[91] X. Jiang, D. Li, H. Chen, Y. Zheng, R. Zhao, L. Wu, 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, 2022, pp. 11174–11184.

[92] T. Pöllabauer, J. Li, V. Knauthe, S. Berkei, A. Kuijper, End-to-end probabilistic geometry-guided regression for 6dof object pose estimation, 2024, arXiv preprint arXiv:2409.11819.

[93] T. Pöllabauer, A. Pramod, V. Knauthe, M. Wahl, FAST GDRNPP: Improving the speed of state-of-the-art 6D object pose estimation, 2024, arXiv preprint arXiv:2409.12720.

[94] F. Liu, Y. Hu, M. Salzmann, Linear-covariance loss for end-to-end learning of 6d pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14107–14117.

[95] F. Duffhauss, S. Koch, H. Ziesche, N.A. Vien, G. Neumann, Symfm6d: Symmetry-aware multi-directional fusion for multi-view 6d object pose estimation, IEEE Robot. Autom. Lett. (2023).

[96] J. Chen, G. Zhang, D. Li, Y. He, Two-stage keypoint detection network for robust 6D pose estimation, in: 2024 36th Chinese Control and Decision Conference, CCDC, IEEE, 2024, pp. 5292–5297.

[97] H. Li, J. Lin, K. Jia, Dcl-net: Deep correspondence learning network for 6d pose estimation, in: European Conference on Computer Vision, Springer, 2022, pp. 369–385.

[98] W. Chen, J. Duan, H. Basevi, H.J. Chang, A. Leonardis, PointPoseNet: Point pose network for robust 6D object pose estimation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 2824–2833.

[99] H. Jiang, M. Salzmann, Z. Dang, J. Xie, J. Yang, Se (3) diffusion model-based point cloud registration for robust 6d object pose estimation, Adv. Neural Inf. Process. Syst. 36 (2024).

[100] Z. Dang, L. Wang, Y. Guo, M. Salzmann, Match normalization: Learning-based point cloud registration for 6d object pose estimation in the real world, IEEE Trans. Pattern Anal. Mach. Intell. (2024).

[101] B. Matteo, T. Tsesmelis, S. James, F. Poiesi, A. Del Bue, 6Dgs: 6d pose estimation from a single image and a 3d gaussian splatting model, in: European Conference on Computer Vision, Springer, 2024, pp. 420–436.

[102] G. Zuo, S. Yu, S. Yu, H. Liu, M. Zhao, Sca-pose: category-level 6D pose estimation with adaptive shape prior based on CNN and graph convolution, Intell. Serv. Robot. (2025) 1–11.

[103] J. Wang, L. Luo, W. Liang, Z.-X. Yang, OA-pose: Occlusion-aware monocular 6-dof object pose estimation under geometry alignment for robot manipulation, Pattern Recognit. 154 (2024) 110576.

[104] J. Sun, L. Chen, Y. Xie, S. Zhang, Q. Jiang, X. Zhou, H. Bao, Disp r-cnn: Stereo 3d object detection via shape prior guided instance disparity estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10548–10557.

[105] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, L.J. Guibas, 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, 2019, pp. 2642–2651.

[106] B. Yuan, W. Yao, P. Jing, J. Zhang, K.F. Tsang, S. Wang, Context-aware focal alignment network for micro-video multi-label classification, Pattern Anal. Appl. 27 (4) (2024) 150.

[107] H. Lin, Z. Liu, C. Cheang, Y. Fu, G. Guo, X. Xue, 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, 2022, pp. 6707–6717.

[108] Y. Zhan, X. Wang, L. Nie, Y. Zhao, T. Yang, Q. Ruan, TG-pose: Delving into topology and geometry for category-level object pose estimation, IEEE Trans. Multimed. (2024).

[109] Z. Li, Y. Hu, M. Salzmann, X. Ji, SD-pose: Semantic decomposition for cross-domain 6D object pose estimation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 2020–2028.

[110] L. Zhou, Z. Liu, R. Gan, H. Wang, M.H. Ang, DR-pose: A two-stage deformation-and-registration pipeline for category-level 6D object pose estimation, in: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2023, pp. 1192–1199.

[111] G. Li, Y. Li, Z. Ye, Q. Zhang, T. Kong, Z. Cui, G. Zhang, Generative category-level shape and pose estimation with semantic primitives, in: Conference on Robot Learning, PMLR, 2023, pp. 1390–1400.

[112] K. Chen, Q. Dou, Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2773–2782.

[113] T.G. Jantos, M.A. Hamdad, W. Granig, S. Weiss, J. Steinbrener, Poet: Pose estimation transformer for single-view, multi-object 6D pose estimation, in: Conference on Robot Learning, PMLR, 2023, pp. 1060–1070.

[114] L. Zou, Z. Huang, N. Gu, G. Wang, 6D-vit: Category-level 6d object pose estimation via transformer-based instance representation learning, IEEE Trans. Image Process. 31 (2022) 6907–6921.

[115] J.Q. Davis, A. Gu, K. Choromanski, T. Dao, C. Re, C. Finn, P. Liang, Catformer: Designing stable transformers via sensitivity analysis, in: International Conference on Machine Learning, PMLR, 2021, pp. 2489–2499.

[116] S. Yu, D.-H. Zhai, Y. Zhan, W. Wang, Y. Guan, Y. Xia, 6-d object pose estimation based on point pair matching for robotic grasp detection, IEEE Trans. Neural Networks Learn. Syst. (2024).

[117] R. Wang, X. Wang, T. Li, R. Yang, M. Wan, W. Liu, Query6dof: Learning sparse queries as implicit shape prior for category-level 6dof pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14055–14064.

[118] T. Cao, F. Luo, Y. Fu, W. Zhang, S. Zheng, C. Xiao, DGECN: A depth-guided edge convolutional network for end-to-end 6D pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3783–3792.

[119] Z. Fan, Z. Song, Z. Wang, J. Xu, K. Wu, H. Liu, J. He, Acr-pose: Adversarial canonical representation reconstruction network for category level 6d object pose estimation, in: Proceedings of the 2024 International Conference on Multimedia Retrieval, 2024, pp. 55–63.

[120] B. Sun, H. Kang, L. Guan, H. Li, P. Mordohai, G. Hua, Glissando-net: Deep single view category level pose estimation ANd 3D reconstruction, IEEE Trans. Pattern Anal. Mach. Intell. (2024).

[121] J. Lin, Z. Wei, C. Ding, K. Jia, Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks, in: European Conference on Computer Vision, Springer, 2022, pp. 19–34.

[122] M. Tian, M.H. Ang, G.H. Lee, Shape prior deformation for categorical 6d object pose and size estimation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16, Springer, 2020, pp. 530–546.

[123] J. Yang, Y. Chen, X. Meng, C. Yan, M. Li, R. Cheng, L. Liu, T. Sun, L. Kneip, MV-ROPE: Multi-view constraints for robust category-level object pose and size estimation, in: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2024, pp. 7588–7595.

[124] Y. He, H. Fan, H. Huang, Q. Chen, J. Sun, Towards self-supervised category-level object pose and size estimation, 2022, arXiv preprint arXiv:2203.02884.

[125] R. She, Q. Kang, S. Wang, W.P. Tay, K. Zhao, Y. Song, T. Geng, Y. Xu, D.N. Navarro, A. Hartmannsgruber, PointDifformer: Robust point cloud registration with neural diffusion and transformer, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–15.

[126] X. Ning, Z. Yu, L. Li, W. Li, P. Tiwari, DILF: Differentiable rendering-based multi-view image–language fusion for zero-shot 3D shape understanding, Inf. Fusion 102 (2024) 102033.

[127] Z. Fan, Z. Song, J. Xu, Z. Wang, K. Wu, H. Liu, J. He, Object level depth reconstruction for category level 6d object pose estimation from monocular rgb image, in: European Conference on Computer Vision, Springer, 2022, pp. 220–236.

[128] R. Zhang, Z. Huang, G. Wang, C. Zhang, Y. Di, X. Zuo, J. Tang, X. Ji, Lapose: Laplacian mixture shape modeling for RGB-based category-level object pose estimation, 2024, arXiv preprint arXiv:2409.15727.

[129] Y. Yin, J. Lyu, Y. Wang, H. Liu, H. Wang, B. Chen, Towards robust probabilistic modeling on SO (3) via rotation laplace distribution, IEEE Trans. Pattern Anal. Mach. Intell. (2025).

[130] L. Zou, Z. Huang, N. Gu, G. Wang, MSSPA-GC: Multi-scale shape prior adaptation with 3D graph convolutions for category-level object pose estimation, Neural Netw. 166 (2023) 609–621.

[131] W. Chen, X. Jia, H.J. Chang, J. Duan, L. Shen, A. Leonardis, Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1581–1590.

[132] Y. Di, R. Zhang, Z. Lou, F. Manhardt, X. Ji, N. Navab, F. Tombari, Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 6781–6791.

[133] J. Chen, M. Sun, Y. Zheng, T. Bao, Z. He, D. Li, G. Jin, Z. Rui, L. Wu, X. Jiang, Geo6D: Geometric-constraints-guided direct object 6D pose estimation network, IEEE Trans. Multimed. (2025).

[134] R. Zhang, Y. Di, Z. Lou, F. Manhardt, F. Tombari, X. Ji, Rbp-pose: Residual bounding box projection for category-level pose estimation, in: European Conference on Computer Vision, Springer, 2022, pp. 655–672.

[135] L. Zheng, C. Wang, Y. Sun, E. Dasgupta, H. Chen, A. Leonardis, W. Zhang, H.J. Chang, Hs-pose: Hybrid scope feature extraction for category-level object pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 17163–17173.

[136] G. Ponimatkin, Y. Labbé, B. Russell, M. Aubry, J. Sivic, Focal length and object pose estimation via render and compare, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3825–3834.

[137] R. Zhang, Y. Di, F. Manhardt, F. Tombari, X. Ji, Ssp-pose: Symmetry-aware shape prior deformation for direct category-level object pose estimation, in: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2022, pp. 7452–7459.

[138] Y. Chen, Y. Di, G. Zhai, F. Manhardt, C. Zhang, R. Zhang, F. Tombari, N. Navab, B. Busam, Secondpose: Se (3)-consistent dual-stream feature fusion for category-level pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 9959–9969.

[139] H. Wang, Z. Fan, Z. Zhao, Z. Che, Z. Xu, D. Liu, F. Feng, Y. Huang, X. Qiao, J. Tang, Dtf-net: Category-level pose estimation and shape reconstruction via deformable template field, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 3676–3685.

[140] A. Remus, S. D'Avella, F. Di Felice, P. Tripicchio, C.A. Avizzano, I2c-net: using instance-level neural networks for monocular category-level 6d pose estimation, IEEE Robot. Autom. Lett. 8 (3) (2023) 1515–1522.

[141] C. Ruosch, M. Birem, A. Bey-Temsamani, Bench marking of industrial object pose estimation algorithm based on vision & deep learning, 2024, p. 177, Communications for Industry 4. 0/ 5. 0 (ARCI 2024).

[142] L. Zheng, T.H.E. Tse, C. Wang, Y. Sun, H. Chen, A. Leonardis, W. Zhang, H.J. Chang, GeoReF: Geometric alignment across shape variation for category-level object pose refinement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 10693–10703.

[143] X. Shi, T. Tang, J. Chen, S. Lv, Y. Liu, Precise depth estimation by calculating affine transformation parameters, IEEE Trans. Geosci. Remote Sens. (2024).

[144] Z. Zhang, J. Yu, L. Cui, Q. Ling, et al., Part-level reconstruction for self-supervised category-level 6D object pose estimation with coarse-to-fine correspondence optimization, in: ACM Multimedia 2024, 2024, pp. 1–9.

[145] X. Deng, J. Geng, T. Bretl, Y. Xiang, D. Fox, iCaps: Iterative category-level object pose and shape estimation, IEEE Robot. Autom. Lett. 7 (2) (2022) 1784–1791.

[146] B. Wan, Y. Shi, K. Xu, Socs: Semantically-aware object coordinate space for category-level 6d object pose estimation under large shape variations, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14065–14074.

[147] J. Liu, Y. Chen, X. Ye, X. Qi, Ist-net: Prior-free category-level pose estimation with implicit space transformation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 13978–13988.

[148] X. Li, Y. Weng, L. Yi, L.J. Guibas, A. Abbott, S. Song, H. Wang, Leveraging se (3) equivariance for self-supervised category-level object pose estimation from point clouds, Adv. Neural Inf. Process. Syst. 34 (2021) 15370–15381.

[149] X. Lin, W. Yang, Y. Gao, T. Zhang, Instance-adaptive and geometric-aware keypoint learning for category-level 6d object pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 21040–21049.

[150] L. Zou, Z. Huang, N. Gu, G. Wang, Learning geometric consistency and discrepancy for category-level 6D object pose estimation from point clouds, Pattern Recognit. 145 (2024) 109896.

[151] Y. Zhang, J.J. Leonard, Shapeicp: Iterative category-level object pose and shape estimation from depth, 2024, arXiv preprint arXiv:2408.13147.

[152] J. Zhang, M. Wu, H. Dong, Generative category-level object pose estimation via diffusion models, Adv. Neural Inf. Process. Syst. 36 (2024).

[153] J. Cai, Y. He, W. Yuan, S. Zhu, Z. Dong, L. Bo, Q. Chen, Ov9d: Open-vocabulary category-level 9d object pose and size estimation, 2024, arXiv preprint arXiv:2403.12396.

[154] X. Lin, M. Zhu, R. Dang, G. Zhou, S. Shu, F. Lin, C. Liu, Q. Chen, Clipose: Category-level object pose estimation with pre-trained vision-language knowledge, IEEE Trans. Circuits Syst. Video Technol. (2024).

[155] J. Lin, Z. Wei, Z. Li, S. Xu, K. Jia, Y. Li, 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, 2021, pp. 3560–3569.

[156] J. Lin, Z. Wei, Y. Zhang, K. Jia, Vi-net: Boosting category-level 6d object pose estimation via learning decoupled rotations on the spherical representations, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14001–14011.

[157] T. Lee, B.-U. Lee, I. Shin, J. Choe, U. Shin, I.S. Kweon, K.-J. Yoon, UDA-COPE: Unsupervised domain adaptation for category-level object pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14891–14900.

[158] W. Ma, A. Wang, A. Yuille, A. Kortylewski, Robust category-level 6d pose estimation with coarse-to-fine rendering of neural features, in: European Conference on Computer Vision, Springer, 2022, pp. 492–508.

[159] D. Chen, J. Li, Z. Wang, K. Xu, Learning canonical shape space for category-level 6d object pose and size estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11973–11982.

[160] Y. You, W. He, J. Liu, H. Xiong, W. Wang, C. Lu, Cppf++: Uncertainty-aware sim2real object pose estimation by vote aggregation, IEEE Trans. Pattern Anal. Mach. Intell. (2024).

[161] C. Zhang, Y. Ling, M. Lu, M. Qin, H. Wang, Category-level object detection, pose estimation and reconstruction from stereo images, 2024, arXiv preprint arXiv:2407.06984.

[162] J. Sun, Z. Wang, S. Zhang, X. He, H. Zhao, G. Zhang, X. Zhou, Onepose: One-shot object pose estimation without cad models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 6825–6834.
[163] J. Wang, H. Yu, X. Lin, Z. Li, W. Sun, N. Akhtar, EFRNet-VL: An end-to-end feature refinement network for monocular visual localization in dynamic environments, Expert Syst. Appl. 243 (2024) 122755.
[164] X. He, J. Sun, Y. Wang, D. Huang, H. Bao, X. Zhou, Onepose++: Keypoint-free one-shot object pose estimation without CAD models, Adv. Neural Inf. Process. Syst. 35 (2022) 35103–35115.
[165] C. Rockwell, N. Kulkarni, L. Jin, J.J. Park, J. Johnson, D.F. Fouhey, FAR: Flexible accurate and robust 6dof relative camera pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 19854–19864.
[166] T. Lee, J. Tremblay, V. Blukis, B. Wen, B.-U. Lee, I. Shin, S. Birchfield, I.S. Kweon, K.-J. Yoon, Tta-cope: Test-time adaptation for category-level object pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 21285–21295.
[167] J. Liu, W. Sun, C. Liu, H. Yang, X. Zhang, A. Mian, Mh6d: multi-hypothesis consistency learning for category-level 6-d object pose estimation, IEEE Trans. Neural Networks Learn. Syst. (2024).
[168] C. Wang, R. Martín-Martín, D. Xu, J. Lv, C. Lu, L. Fei-Fei, S. Savarese, Y. Zhu, 6-pack: Category-level 6d pose tracker with anchor-based keypoints, in: 2020 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2020, pp. 10059–10066.
[169] J. Zhang, W. Huang, B. Peng, M. Wu, F. Hu, Z. Chen, B. Zhao, H. Dong, Omni6DPose: A benchmark and model for universal 6D object pose estimation and tracking, 2024, arXiv preprint arXiv:2406.04316.
[170] Y. Guo, F. Wang, H. Chu, S. Wen, Cross-modal attention and geometric contextual aggregation network for 6dof object pose estimation, Neurocomputing 617 (2025) 128891.
[171] H. Zhao, S. Wei, D. Shi, W. Tan, Z. Li, Y. Ren, X. Wei, Y. Yang, S. Pu, Learning symmetry-aware geometry correspondences for 6d object pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14045–14054.
[172] K. Park, A. Mousavian, Y. Xiang, D. Fox, 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, 2020, pp. 10710–10719.
[173] J. Huang, H. Yu, K.-T. Yu, N. Navab, S. Ilic, B. Busam, Matchu: Matching unseen objects for 6d pose estimation from rgb-d images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 10095–10105.
[174] Y. Wen, X. Li, H. Pan, L. Yang, Z. Wang, T. Komura, W. Wang, Disp6d: Disentangled implicit shape and pose learning for scalable 6d pose estimation, in: European Conference on Computer Vision, Springer, 2022, pp. 404–421.
[175] Y. Xu, K.-Y. Lin, G. Zhang, X. Wang, H. Li, Rnnpose: 6-dof object pose estimation via recurrent correspondence field estimation and pose optimization, IEEE Trans. Pattern Anal. Mach. Intell. 46 (7) (2024) 4669–4683.
[176] C. Zhao, Y. Hu, M. Salzmann, Fusing local similarities for retrieval-based 3d orientation estimation of unseen objects, in: European Conference on Computer Vision, Springer, 2022, pp. 106–122.
[177] T. Wang, G. Hu, H. Wang, Object pose estimation via the aggregation of diffusion features, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 10238–10247.
[178] J. Lin, L. Liu, D. Lu, K. Jia, Sam-6d: Segment anything model meets zero-shot 6d object pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 27906–27916.
[179] J. Corsetti, D. Boscaini, F. Giuliari, C. Oh, A. Cavallaro, F. Poiesi, High-resolution open-vocabulary object 6D pose estimation, 2024, arXiv preprint arXiv:2406.16384.
[180] D. Cai, J. Heikkilä, E. Rahtu, Ove6d: Object viewpoint encoding for depth-based 6d object pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 6803–6813.
[181] X. Li, H. Wang, L. Yi, L.J. Guibas, A.L. Abbott, S. Song, Category-level articulated object pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3706–3715.
[182] L. Liu, H. Xue, W. Xu, H. Fu, C. Lu, Toward real-world category-level articulation pose estimation, IEEE Trans. Image Process. 31 (2022) 1072–1083.
[183] N. Heppert, M.Z. Irshad, S. Zakharov, K. Liu, R.A. Ambrus, J. Bohg, A. Valada, T. Kollar, Carto: Category and joint agnostic reconstruction of articulated objects, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 21201–21210.
[184] Y. Che, R. Furukawa, A. Kanezaki, OP-align: Object-level and part-level alignment for self-supervised category-level articulated object pose estimation, 2024, arXiv preprint arXiv:2408.16547.
[185] H. Xue, L. Liu, W. Xu, H. Fu, C. Lu, Omad: Object model with articulated deformations for pose estimation and retrieval, 2021, arXiv preprint arXiv:2112.07334.
[186] Y. Weng, H. Wang, Q. Zhou, Y. Qin, Y. Duan, Q. Fan, B. Chen, H. Su, L.J. Guibas, Captra: Category-level pose tracking for rigid and articulated objects from point clouds, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13209–13218.
[187] L. Liu, A. Huang, Q. Wu, D. Guo, X. Yang, M. Wang, KPA-tracker: Towards robust and real-time category-level articulated object 6D pose tracking, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, 2024, pp. 3684–3692.
[188] Y. Xiang, T. Schmidt, V. Narayanan, D. Fox, Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes, 2017, arXiv preprint arXiv:1711.00199.
[189] J. Wang, K. Chen, Q. Dou, Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2021, pp. 4807–4814.
[190] K. Mo, Y. Qin, F. Xiang, H. Su, L. Guibas, O2o-afford: Annotation-free large-scale object-object affordance learning, in: Conference on Robot Learning, PMLR, 2022, pp. 1666–1677.
[191] Y. Liu, J. Yang, X. Gu, Y. Chen, Y. Guo, G.-Z. Yang, Egofish3d: Egocentric 3d pose estimation from a fisheye camera via self-supervised learning, IEEE Trans. Multimed. 25 (2023) 8880–8891.
[192] Y. Hai, R. Song, J. Li, D. Ferstl, Y. Hu, Pseudo flow consistency for self-supervised 6d object pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14075–14085.
[193] Y. Hai, R. Song, J. Li, Y. Hu, Shape-constraint recurrent flow for 6d object pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4831–4840.
[194] A. Bangunharcana, A. Magd, K.-S. Kim, DualRefine: Self-supervised depth and pose estimation through iterative epipolar sampling and refinement toward equilibrium, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 726–738.
[195] H. Fan, H. Su, L.J. Guibas, A point set generation network for 3d object reconstruction from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 605–613.
[196] J. Zhou, K. Chen, M. Wei, X.-P. Zhang, Q. Dou, J. Qin, Canonical shape reconstruction with SE (3) equivariance learning for weakly-supervised object pose estimation, IEEE Trans. Circuits Syst. Video Technol. (2025).
[197] F. Xu, Z. Zhu, C. Feng, J. Leng, P. Zhang, X. Yu, C. Wang, X. Chen, An object planar grasping pose detection algorithm in low-light scenes, Multimedia Tools Appl. (2024) 1–22.
[198] Y. Gao, L. Chen, J. Liu, Vaepose: 6D pose estimation with visual appearance enhancement for low-light conditions, in: 2024 IEEE International Conference on Industrial Technology, ICIT, IEEE, 2024, pp. 1–7.
[199] V.N. Nguyen, T. Groueix, G. Ponimatkin, Y. Hu, R. Marlet, M. Salzmann, V. Lepetit, Nope: Novel object pose estimation from a single image, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 17923–17932.
[200] J. Wu, X. Ru, S. Lu, R. Xiong, Y. Wang, Recurrent volume-based 3D feature fusion for real-time multi-view object pose estimation, IEEE Trans. Instrum. Meas. (2024).
[201] D.-T. Huang, E.-T. Lin, L. Chen, L.-F. Liu, L. Zeng, SD-net: Symmetric-aware keypoint prediction and domain adaptation for 6D pose estimation in bin-picking scenarios, in: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2024, pp. 2747–2754.
[202] M. Rad, V. Lepetit, 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, 2017, pp. 3828–3836.
[203] B. Tekin, S.N. Sinha, P. Fua, Real-time seamless single shot 6d object pose prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 292–301.