The Visual Computer (2025) 41:8009–8023
https://doi.org/10.1007/s00371-025-03852-6

RESEARCH

GRPoseNet: a generalizable and robust 6D object pose estimation network using sparse RGB views

Wubin Shi[1,2] · Shaoyan Gai[1,2] · Feipeng Da[1,2] · Zeyu Cai[1,2] · Jiaoling Wang[3,4]

Accepted: 17 February 2025 / Published online: 11 March 2025

Abstract

Six-degree-of-freedom object pose estimation plays a crucial role in various computer vision and robotics tasks. Existing methods often rely heavily on CAD models and substantial prior information, limiting their generalization to unseen objects in open scenes. To address this limitation, we propose GRPoseNet, a generalizable and robust 6D object pose estimation network that can predict the pose of unseen objects using only sparse RGB images with reference poses. GRPoseNet comprises an open-world detector, a viewpoint selector, and an adaptive multi-scale refiner. The open-world detector leverages pre-trained large models for zero-shot segmentation and feature extraction, overcoming detection and matching errors with unseen objects. The viewpoint selector uses our designed similarity network to select the most similar reference view for initial pose estimation. The adaptive multi-scale refiner further refines the pose by iteratively updating rotation and translation residuals based on multi-scale features and adaptive weights. Extensive experiments on benchmark datasets and our robust test dataset, RBMOP, demonstrate that GRPoseNet achieves state-of-the-art performance, showing excellent generalization and robustness to unseen objects and sparse views. The codes and datasets are available at: https://github.com/KierSaS/GRPoseNet.

Keywords Object pose estimation · Deep learning · Generalizable · Zero shot

1 Introduction

6D object pose estimation is a crucial step in various computer vision and robotics tasks, such as augmented reality [1, 2], autonomous driving [3, 4], and robot manipulation [5, 6]. Currently, there are three main approaches about pose estimation: instance-level pose estimation, category-level pose estimation, and the recently emerged generalizable objects pose estimation. For instance-level 6D object pose estimation, in which the training set and the test set are the same objects, great progress has been made in recent years due to the development of deep learning. However, the excellent performance of this method is mainly due to its high-fidelity CAD model of the object and the pre-training of the target object.

We hope to find a solution for the open visual scene of intelligent robots in the future to realize the pose estimation on unseen scenes and objects. There is no doubt that the instance-level approach does not seem suited to this challenge. The category-level pose estimation method proposed in recent years turns to the scene of open vision. By training on the same category of object, this method [7–10] can estimate the 6D pose of the object that has not been seen in this

✉ Shaoyan Gai
qxxymm@163.com

Wubin Shi
wubinshi@seu.edu.cn

Feipeng Da
dafp@seu.edu.cn

Zeyu Cai
zeyucai@seu.edu.cn

Jiaoling Wang
kclwjl@126.com

1 School of Automation, Southeast University, Nanjing, China

2 Key Laboratory of Measurement and Control of Complex Systems of Engineering, Nanjing, China

3 Zhejiang Provincial Key Laboratory of Agricultural Intelligent Equipment and Robotics/College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou, China

4 Nanjing Institute of Agricultural Mechanization, Ministry of Agriculture and Rural Affairs, Nanjing, China

category. But they are still limited to pre-defined category-level assumptions.

The generalizable objects pose estimator was developed mainly to overcome the above shortcomings and realize the pose estimation of unseen objects in open vision. In contrast to generalized pose estimation approaches based on reference template matching [11] or rendering [12, 13], we focus on the model-free generalized pose estimator [14–17]. Some model-free methods such as OnePose++ [15] and FS6D [18] use a large amount of reference data to model objects using structure from motion (SFM). Then feature point matching is used to establish 2D and 3D correspondences and solve poses. With the recent development of reconstruction techniques such as Nerf and 3D Gaussian splatting, some methods [19–22] use Nerf and 3D Gaussian to reconstruct high-quality models and perform pose estimation. However, we maintain that relying on multi-view ($\geq$ 128) reconstruction for all previously unseen objects is not suitable for addressing the open-world visual perception scenarios encountered by robots. When faced with a variety of scenes and objects, it is too expensive to collect and reconstruct a high-precision 3D model for each object.

In this paper, we propose GRPoseNet, as shown in Fig. 1. The method aims to efficiently estimate model-free objects and also has good generalization and robustness when detecting and estimating unseen objects in cluttered and sparse views. In the detection stage, previous methods that directly extract and detect features using the entire image will produce errors in unseen objects and cluttered backgrounds. In order to reduce its impact on the accuracy of target detection and matching, we use segmentation and visual large models (SAM2 and DINOV2) to implement zero-shot segmentation and feature extraction tasks. Then, we use the designed semantic and geometric matching scores to establish the correspondence between the same object in different scenes. In the pose estimation stage, we follow a coarse-to-fine strategy. For the view selector, we consider the in-plane rotation and scale in the matching process and design a similarity network with multi-head self-attention mechanism to evaluate the similarity between the target view and the reference view. Finally, the reference view pose with the highest similarity is used as the initial pose. In the subsequent refinement network, we use multi-scale fusion features to fully utilize similar view information to encourage the network to update rotation and translation and adopt an adaptive weight strategy to weaken the pose gap caused by sparse views.

In order to evaluate and verify the robustness of our method under background changes, we created a new synthetic dataset called Robust Background Model-free Object Pose Dataset (RBMOP). Our experiments on LineMOD [23], GenMOP [14], and the synthetic dataset RBMOP show that our method consistently outperforms other RGB-based generalization methods. We summarize the contributions as follows:

- We alleviate the challenges of 6D pose estimation caused by sparse views and unseen objects in complex backgrounds, achieving robust results using only RGB sparse views.
- We propose an open-world detector that leverages large models and a matching design based on semantic and geometric information to achieve zero-shot segmentation and matching tasks. The selector and refiner fully utilizes similar view information through multi-scale information and adaptive weights to alleviate the sparse view problem.
- We construct a novel synthetic dataset named RBMOP, which can be used to evaluate the robustness of the pose estimation model to background and environmental changes.
- Comprehensive experiments demonstrate that the proposed method achieves high performance across various datasets. Extensive ablation studies confirm the effectiveness of key components of the proposed network.

2 Related works

2.1 Instance-level 6D object pose estimation

Instance-level pose estimation refers to using model information for training and testing. Based on the different approaches to utilizing models, instance-level pose estimation can be broadly categorized into three types: correspondence-based methods, voting-based methods and template-based methods. Correspondence-based methods primarily establish correspondences between the input data and existing models, by establishing 2D–3D [24, 25] or 3D–3D [26, 27] point correspondences. The pose is then solved using PnP or ICP methods. Some of these correspondences are generated through pixel voting [28, 29], which defines the voting-based approach. These methods [30, 31] have been proved to achieve good prediction performance in partially occluded scenes. In contrast, template matching methods [32–35] align templates with images or depth maps through feature descriptors, where object templates come from a template library generated by rendering object models.

2.2 Category-level 6D object pose estimation

Due to the great prospects of open vision [36], researchers have begun to work on the pose estimation problem of unseen objects. Category-level pose estimation methods [37, 38] solve the pose estimation problem of different instance objects under the same category. In order to deal with intra-class differences, He Wang [9] introduced the concept of

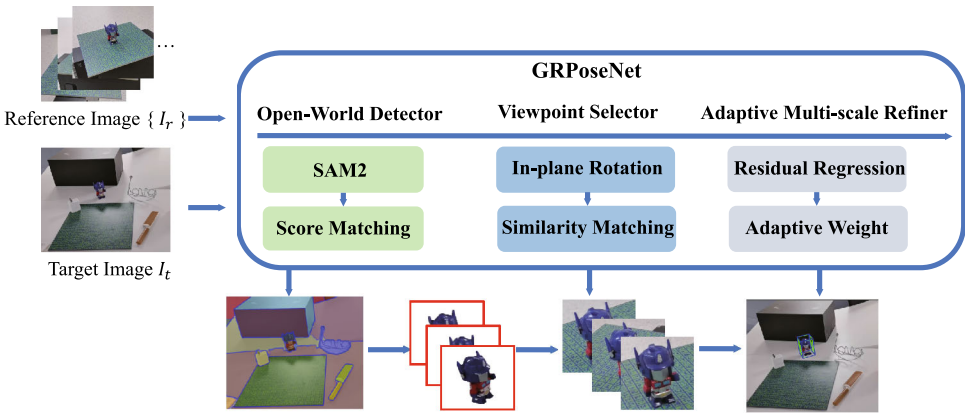

Fig. 1 Overall pipeline of the proposed GRPoseNet framework. It mainly consists of an open-world detector, a viewpoint selector, and an adaptive multi-scale refiner

normalized object coordinate space (NOCS) based on RGBD images. In this approach, different objects are mapped to NOCS space, and the 6D object pose is calculated based on 3D–3D correspondences. Many subsequent methods [39, 40] have adopted and improved upon this normalized space concept. Another approach [41–43] involves shape deformation strategy, where a 3D object model is reconstructed through explicit deformation based on learned categorical shape priors. The network infers the dense correspondence between the deep observation and the reconstructed 3D model, and then estimates the 6D pose. However, these category-level methods struggle when dealing with objects from novel categories.

2.3 Generalizable 6D object pose estimation

In recent years, there have been some model-free pose estimation methods that can generalize to new objects. Unlike methods [44–46] that rely on depth maps and object masks, recent approaches such as Gen6D [14], OnePose [47], and OnePose++ [15] only require a set of reference images to estimate the object pose and can be extended to unseen objects. Specifically, Gen6D first detects the object boxes in the reference image and the query image, then selects the reference images pose that is most similar to the query image as the initialization pose according to the similarity network, and finally performs pose refinement. OnePose++ reconstructs the sparse point cloud and matches the query images to determine the pose of the object. With the recent development of 3D reconstruction, such as Nerf [48] and 3D Gaussian Splatting [49], some work [50–52] also achieved pose estimation of unseen objects based on Nerf and 3DGS for model reconstruction and pose regression.

2.4 Large-scale pre-trained foundation models

Recently, many works [53, 54] have employed foundational models as backbone networks across a wide range of tasks, demonstrating strong generalization capabilities. For example, DINO [55] and DINOv2 [56], after self-supervised learning, have shown robust extraction of visual features in image representation [57], object recognition [58, 59], and tracking [17, 60]. The Segment Anything Model (SAM) [61] is a foundational model focused on prompt-based segmentation tasks, supporting interactive segmentation through points, boxes, text, and masks. Notably, it exhibits outstanding zero-shot segmentation capabilities, performing excellently in various visual scenarios [62, 63]. The zero-shot generalization and segmentation capabilities of these foundational models open new possibilities for detection and pose estimation in open-world scenarios.

3 Methods

3.1 Problem formulation

In this section, we formally propose the challenging setting of generalizable and robust 6D object pose estimation with sparse RGB views. The 6D object pose means a translation $\mathbf{T}$ and a rotation $\mathbf{R}$ that transform the object coordinate X_{obj} to the camera coordinate $X_{\text{cam}} = \mathbf{R}X_{\text{obj}} + \mathbf{T}$. In our challenging setting, we assume that N RGB reference views I_q^i with poses $\{I_q^i, p^i\}_i^N$ can be accessed. At this time, we receive an image of the object in the new scene. Our goal is to identify the target object and predict its 6D pose.

Unlike the conventional closed set problem, we focus on locating unseen objects, which means that the objects have not been pre-trained by the network. This is a challenge for the design of both object detectors and pose estimators. The detector network may not be able to accurately determine the feature correspondence between different views of the same object, which may be caused by unseen objects and diverse backgrounds. For the pose estimator, sparse reference views, unlike dense support views, may have a large pose gap with the target view. The lack of similar reference views deepens the difficulty of pose refinement.

3.2 Open-world detector

In generalizable pose estimation, one of the most common failure modes is that the network cannot find the feature correspondences between different views of the same object. The query image and the reference image are usually very large, but the object only occupies a small area on the query image. And due to the challenging setting we proposed, the new objects are not pre-trained by the network. The significant variations in the background and environment of the object affect the extracted image features. As a result, conventional detection networks do not perform well in locating the target objects. Our solution is to use a large model trained on a large amount of data to achieve zero-shot segmentation and detection tasks, to detect the correspondence between any image and any object in an open-world detector.

Our design of the open-world detector is shown in Fig. 2. For the input target image I_{t} and reference image I_{r}, we utilize SAM2 for zero-shot segmentation. Specifically, SAM2 consists of an image encoder Ψ_{image}, a prompt encoder Ψ_{prompt}, and a mask decoder Φ_{mask}, as described by the following:

$$\{M, S\} = \Phi_{\mathrm{mask}}\left\{\Psi_{\mathrm{image}}(I), \Psi_{\mathrm{prompt}}(P)\right\}, \tag{1}$$

where I is the input image and P is various types of prompts such as points, boxes, and masks. The outputs M and S are the segmented mask proposals and the corresponding confidence scores. We generate all potential mask proposals by uniformly sampling a dense 2D grid as prompts, producing all possible mask predictions. Additionally, we identify and select stable masks, where a mask is considered stable if its thresholded probability map falls between $0.5-$ and $0.5+$ and is consistent with similar masks. Meanwhile, for the retained proposals, we measure the overlap by calculating the image IoU between the current proposal and the remaining ones, and we apply non-maximal suppression (NMS) to reduce redundancy and duplicates generated by SAM2 during segmentation.

Through instance segmentation of the target and reference images, we obtained all mask proposals $\left\{M_{\mathrm{crop}}\right\}_1^k$ and $\left\{N_{\mathrm{crop}}\right\}_1^l$. For all instances $m \in M_{\mathrm{crop}}$ and $n \in N_{\mathrm{crop}}$, we compute the matching scores S_m between all masks. Since the prompt image primarily contains the target object O, we only need to identify the highest similarity value in the matrix. The row and column indices of this value indicate the mask label of the target object. However, unlike the traditional method [64, 65] that only uses semantic information as the only criterion for measuring similarity, we also integrate geometric information to comprehensively evaluate the proposal similarity.

Specifically, we use DINOV2 to extract the embedded CLS tokens and CLS patch tokens for all proposals M and N. The semantic similarity matching score S_s mainly calculated by:

$$\begin{aligned} S_s &= \mathrm{cosine}_{\mathrm{sim}}(\mathrm{CLS}_m \cdot CLS_n) \\ &= \frac{\mathrm{CLS}_m \cdot \mathrm{CLS}_n}{\sqrt{\sum_{i=1}^{l} \mathrm{CLS}_m[i]^2} \cdot \sqrt{\sum_{i=1}^{l} \mathrm{CLS}_n[i]^2}}, \end{aligned} \tag{2}$$

where CLS represents the image semantic information extracted by DINOV2 in vector form. S_s is calculated by the cosine similarity between proposal m and proposal n.

Secondly, we utilize the CLS patch token of the images to compute the average similarity of the geometric appearance. We calculate the average cosine similarity score between each patch in proposal m and the most similar corresponding patch in proposal n, using S_a for shape matching.

$$S_a = \frac{1}{Pa_m} \sum_{j=1}^{Pa_m} \max_{k=1..Pa_n} \left[\mathrm{cosine}_{\mathrm{sim}}(\mathrm{CLS}_m^j \cdot \mathrm{CLS}_n^k)\right], \tag{3}$$

where Pa_m and Pa_n represent the number of patches decomposed from proposals m and n. CLS_m^j denotes the jth patch feature embedding of proposal m. The calculation of $\mathrm{cosine}_{\mathrm{sim}}$ follows Eq. 2.

The final matching score S_m for the two proposals is calculated as follows:

$$S_m = \frac{w_1 S_s + w_2 S_a}{w_1 + w_2}. \tag{4}$$

Because S_s and S_a are within the range $[0, 1]$, $w1$ and $w2$ are the weights balancing the similarity, which are set to 1 empirically. $S_m(i, j)$ represents the similarity between the ith mask proposal in M and the jth mask proposal in N.

By identifying the highest score in S_m, we can extract the segmentation masks of the target object O from both images. Based on the detected mask and scale, we compute the center point to estimate the initial 3D translation $\mathbf{T}$ and crop the object region. Additionally, by finding the maximum similarity value in each row of the score matrix S_m, we can easily extend the approach to multi-object instance detection and matching between two images.

3.3 Viewpoint selector

In the case of detecting the target object, our pose estimation approach involves first selecting similar viewpoints and then refining the pose. The viewpoint selector aims to find a reference image with the pose most similar to the object in the target image using a similarity prediction network. Considering in-plane rotation, we use the reference image's viewpoint as the query image's viewpoint to estimate the object's initial rotation. Figure 3 illustrates the design of our viewpoint

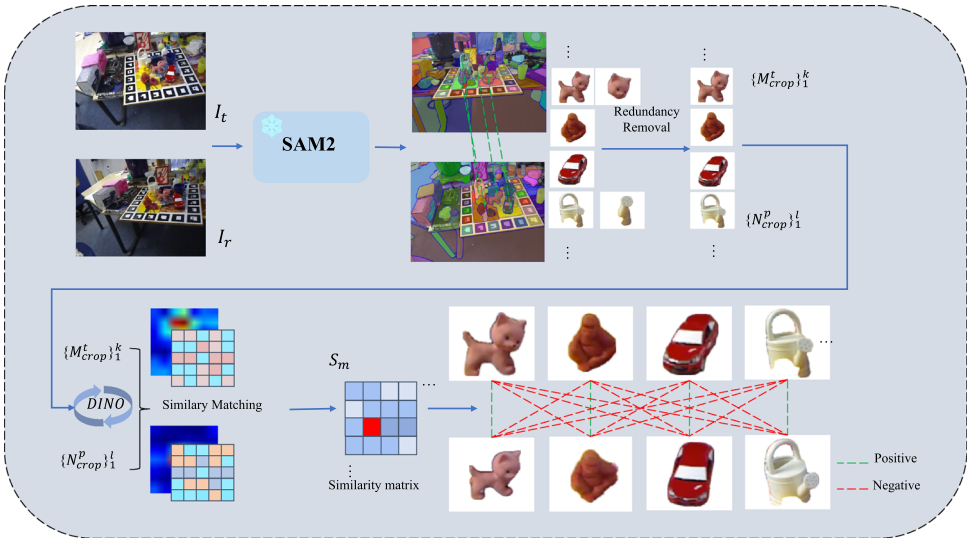

Fig. 2 Framework of the open-world detector. We utilize the pre-trained segmentation model SAM2 for zero-shot segmentation and design a similarity evaluation system to establish instance-level correspondence

selector. The viewpoint selector mainly constructs a similarity prediction network to find a reference image whose viewpoint is most similar to the target image, and uses its rotation angle as the initial rotation angle.

Specifically, considering untrained objects, to enhance the robustness of the feature extraction module, we use DINOV2 for feature map extraction. To fully account for object scale variations, we extract multi-scale information and perform pixel-wise multiplication at the same scale to generate correlations. The feature maps are resized through convolutional modules and concatenated to capture information at both coarse and fine granularity. To compress the data and retain representative features, we fuse the multi-scale correlated features using convolution. Positional encoding is added to strengthen spatial information representation before feeding the features into the multi-head self-attention module. Compared to the local receptive field mechanism in standard convolution, the multi-head self-attention mechanism establishes non-local dependencies on the feature map, enabling the network to better handle rotations, scale variations, or viewpoint changes. The output features pass through a linear layer to predict the rotation angle. Following pooling and viewpoint encoding, the self-attention mechanism is used again to enhance feature interactions and calculate the similarity score between the two images. The rotation pose R of the most similar reference image is used as the initial rotation pose.

3.4 Adaptive multi-scale refiner

Unlike [66], which mitigates the sparsity of views using cascaded iterative operations based on a multi-scale feature pyramid, we propose a multi-scale adaptive refiner, as illustrated in Fig. 4. This refiner integrates multi-scale information and employs a regression network guided by a loss function to predict pose residuals. By combining the residual predictions of the regression network with adaptive weighting, the refiner iteratively predicts pose residuals using multiple similar views and applies adaptive weighting, effectively addressing the pose gaps caused by sparse views and improving the accuracy of pose estimation.

The network uses a pre-trained DINOV2 model to extract features at different scales. High-resolution features typically capture texture and edge details, while low-resolution features are better at capturing global structures or the overall object outline. When these multi-scale features are concatenated and fed into convolutional blocks with residual connections, the network becomes more robust in handling complex scenes and object recognition. The residual modules also mitigate the risk of overfitting caused by the limited number of sparse views. The Transformer encoder structure with a multi-head self-attention mechanism is introduced to capture the global image information and enhance the feature expression. By adding markers with position encoding, we process them separately through Transformer encoders and linearly project them to the output dimension. More specifically, the network encourages predicting the rotation residual ΔR and translation residual ΔT between the view poses, as shown in Eq. 5.

$$\begin{cases} \Delta R = \text{Linear}(\Psi_{\text{tranformer}}(F^f_{\text{target}}, F^f_{\text{ref}})) \\ \Delta T = \text{Linear}(\Psi_{\text{tranformer}}(F^f_{\text{target}}, F^f_{\text{ref}})) \end{cases}. \tag{5}$$

Inspired by the idea of iterative updates from multiple views, we adopt a Top-3 strategy, selecting the three reference images with the highest similarity scores from the selector. Residual predictions are generated iteratively, and the pose residuals are computed by adaptively adjusting the weights based on the similarity scores, updating the final predicted pose $[R, T]$.

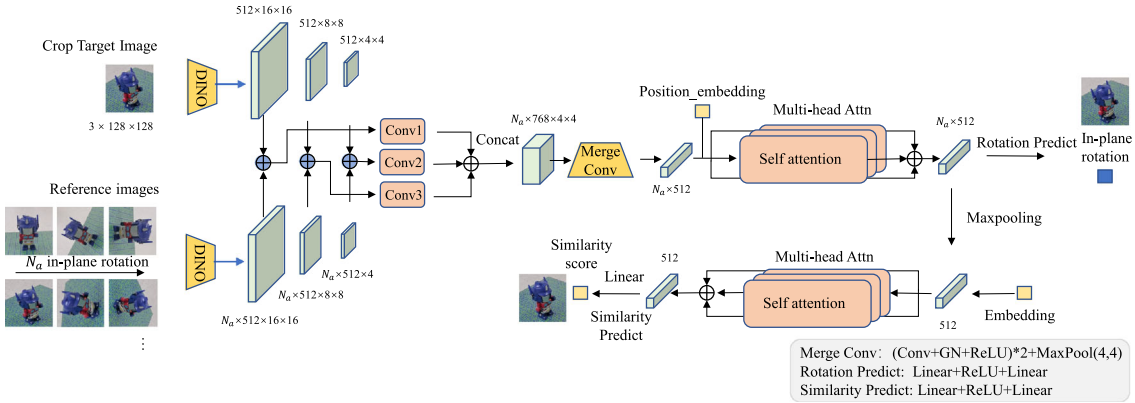

Fig. 3 Architecture of the viewpoint selector. The network considers the in-plane rotation of objects in the image, and by inputting the target image I_t^c and multiple sets of reference images, it outputs the in-plane rotation angle and the similarity scores with different reference images

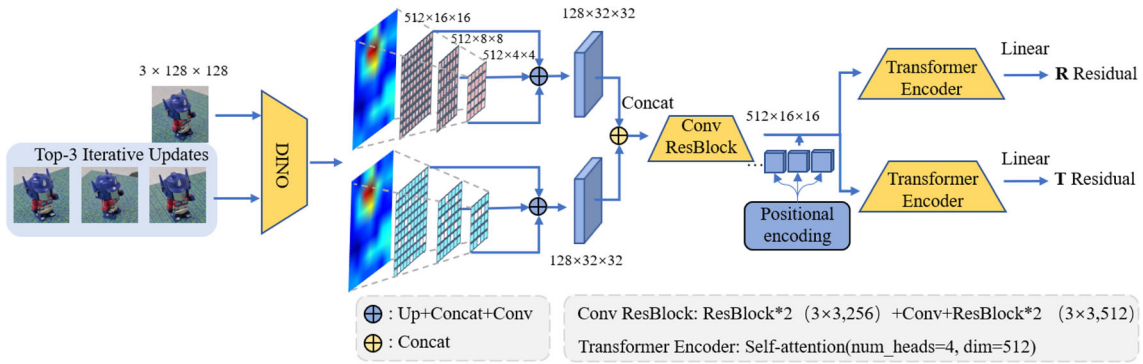

Fig. 4 Architecture of adaptive multi-scale refiner. We first fuse multi-scale image features to leverage information from different levels. To better capture global image information, the processed features are fed into a multi-head self-attention Transformer encoder. This encoder evaluates the alignment quality between the target and reference images and updates the pose residual

$$\begin{cases} \Delta R_{\text{final}} = \sum_{k=1}^{K} w_k \Delta R_k \\ \Delta T_{\text{final}} = \sum_{k=1}^{K} w_k \Delta T_k \end{cases}, \tag{6}$$

where k represents the kth reference view of the iteration, w_k represents an adaptive weight, which can be calculated by the similarity score.

$$w_k = \frac{\text{Sim}(I_{\text{target}}, I_{ref}^k)}{\sum_{k=1}^{K} \text{Sim}(I_{\text{target}}, I_{\text{ref}}^k)}. \tag{7}$$

3.5 Supervised training

The model mainly consists of three modules. Each module can be run and trained independently. In the detector part, SAM2 and DINOV2 are both pre-trained models, and the final feature calculations extracted also have fixed formulas, such as cosine similarity. Therefore, the network does not need to be trained.

For the viewpoint selector, we mainly focus on in-plane rotation prediction and similarity prediction. To train the model to predict the in-plane rotation angle, we calculate the in-plane rotation between the query image and the reference image using the ground truth and supervise the prediction with the following loss function:

$$l_{\text{angle}} = \left\| \theta_i - \theta_{gt} \right\|_2^2, \tag{8}$$

where θ_i represents the predicted value and θ_{gt} represents the true value of the in-plane rotation.

For similarity prediction, we treat it as a classification problem. The model aims to maximize the similarity of correctly matched pairs S_j, while minimizing the similarity of incorrectly matched pairs. We use l_{sim} to supervise the train-

ing of similarity prediction.

$$l_{\text{sim}} = -\frac{s_j}{\log \sum_1^i \exp s_i}. \tag{9}$$

During the training phase of the refiner, our main goal is to make the predicted residuals between the two images close to the true residuals. By incorporating this loss function into the training pipeline, the refiner learns to iteratively predict pose residuals with greater accuracy. The network training is supervised by L2 loss:

$$\begin{aligned} L_{\text{refine}} &= L_R + L_T \\ &= \left\| \Delta R_{\text{gt}} - \Delta R_{\text{pre}} \right\|_2 + \left\| \Delta T_{\text{gt}} - \Delta T_{\text{pre}} \right\|_2 . \end{aligned} \tag{10}$$

4 Experiments

4.1 Datasets

LineMOD Dataset [23]. The LineMOD dataset is a widely used dataset for object pose estimation. It consists of 13 different objects with significant shape changes. Each object has about 1000 test images.

GenMOP Dataset [14]. GenMOP captured 10 objects in different video sequences, each of which was divided into training and testing parts, each of which contained about 200 images.

RBMOP Dataset. RBMOP is a synthetic dataset we created for 6D object pose estimation with different backgrounds, lighting, and noise. In order to fully explore the robustness of the network to background and noise changes, some example images are shown in Fig. 5.

4.2 Implementation details and metrics

Metrics. We adopted the average distance (**ADD**) and projection error as metrics to evaluate the performance of all methods. For an object **O** consists of vertices v, the ADD of asymmetric objects with the ground-truth pose R^*, T^* and the predicted pose R, T is calculated by:

$$\text{ADD} = \frac{1}{m} \sum_{v \in O} \left\| (Rv + T) - \left(R^* v + T^*\right) \right\| . \tag{11}$$

On the ADD, we compute the recall rate with 10% of the object diameter (ADD-0.1d). For the projection error, we computed the recall rate at 5 pixels (**Prj-5**).

Implementation details. All our experiments are deployed on a server equipped with two Intel Xeon Gold 6226R CPUs and two GTX 3090Ti GPUs. The deep learning framework environment is set to PyTorch 1.10 and CUDA 11.3. In order to achieve a fair comparison, we completely follow the same dataset settings of Gen6D on Linemod and GenMOP datasets. Our test dataset come from five objects in the GenMOP dataset and five targets in the LineMOD dataset. We use the Adam optimizer [67] to train our network for 300,000 iterations, the batch size is 8, the learning rate is 10^{-4}, and the learning rate decays to half of the original after every 100,000 iterations.

4.3 Results on LineMOD

We evaluate our proposed GRPoseNet by comparing it against five state-of-the-art methods on their respective baselines. Specifically, PvNet [30] represents instance-level pose estimation methods. Gen6D [14] and Cas6D [66] are used as baselines for model-free pose estimation approaches, while OnePose++ [15] and GS-Pose [68] serve as baselines for multi-view reconstruction-based pose estimation methods.

In the comparison on the Linemod dataset, PvNet uses real CAD models for instance-level testing. For the four generalized pose estimation methods, we follow the same experimental baseline setup [14], using the full reference views from the Linemod dataset and taking the eight objects in Linemod as unseen objects during testing. Table 1 presents the quantitative evaluation results of each method on the Linemod dataset, while Fig. 6 shows the qualitative comparison results of the main baseline methods.

The results in Table 1 show that our method has some shortcomings compared with the instance-level method with detailed CAD model. The main reason is that our method does not have a detailed and accurate model, and the object has not been trained in advance. Without training on objects, the generalized model will not have targeted parameters to perceive the details of the features, resulting in poor performance. Notably, although our method may not achieve the highest accuracy for certain objects, it shows superior generalization and robustness across different object categories compared with Gen6D and Cas6D. This is mainly attributed to the fact that we introduce a detector with zero-shot detection capability to handle unseen objects. For some objects where our method performs poorly, we think it is mainly due to the low-texture objects and the occlusion of some views. These challenges lead to difficulties in viewpoint selection and optimization, resulting in performance degradation for Gen6D, Cas6D, and our approach. In contrast, model reconstruction-based methods like OnePose++ and GS-Pose perform exceptionally well on almost all objects when sufficient reference views($\approx$ 180) are available. However, the performance of model reconstruction-based pose estimation methods heavily relies on the quantity and quality of reference images. Our method demonstrates superior performance under conditions of sparse views and reference views with complex backgrounds. Detailed results can be

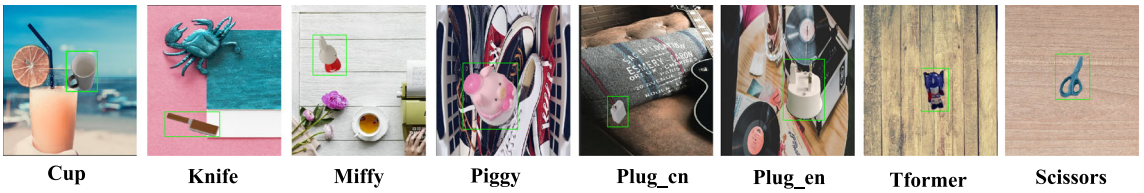

Fig. 5 Objects in the RBMOP dataset. The dataset we created contains a total of 3200 images (2.5G) with pose truth labels, including 8 categories of objects, each of which has 400 images

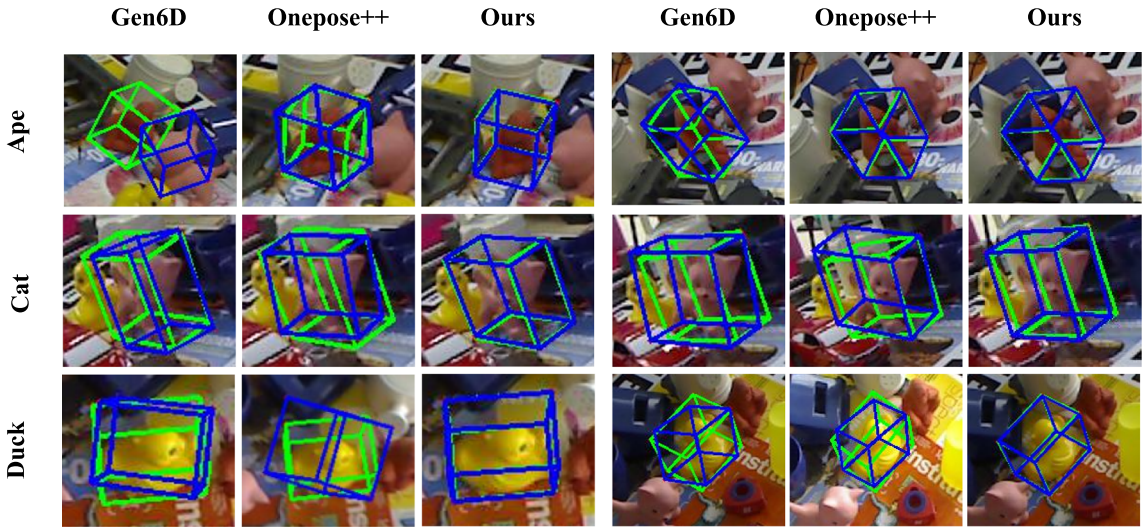

Fig. 6 Qualitative comparison results of the main baseline methods on the Linemod dataset. The ground-truth poses are drawn in green, while estimated poses are drawn in blue

found in the GenMOP dataset testing and ablation study sections.

4.4 Results on GenMOP and RBMOP

As GenMOP is a model-free dataset, the instance-level method lacks model information for reference. To approximate the performance of the instance-level method under our challenging settings, we use the 3D bounding box as key points for PvNet voting. The quantitative results are shown in Table 2.

Compared with the instance-level method PvNet [30] which lacks sufficient prior information, the five generalized pose estimation methods show excellent performance on objects that have not been trained by the network. This also reflects the rapid expansion of the other five methods, which can perform more effective pose estimation on unseen objects without pre-training.

Our method achieves excellent pose estimation results on most objects. However, the performance decline observed on the chair object can be attributed to its unique geometric structure and shape, which differ from common chairs. This distinction poses a challenge to SAM-based detectors, especially in distinguishing the relationship between the overall chair and its individual components. Both OnePose++ and GS-Pose experience some performance degradation on this dataset, likely due to the GenMOP reference images used for model reconstruction. These images, in addition to containing the target objects, also include cluttered environments, which reduce the quality of the reconstructed point clouds and ultimately affect the overall accuracy. The qualitative results in Fig. 7 highlight these key observations.

In order to explore the robustness of the pose estimation model to complex backgrounds and noise in images, we synthesized a new dataset by background changes based on GenMOP and created RBMOP using data enhancement methods including background blur, color and illumination changes, and multi-channel noise injection. Table 3 reports the evaluation results of our method compared with four other approaches. The performance of our method is always better than previous work. This further proves that our method has better robustness to various background changes and noises. For some scenes where the background of the picture changes, Gen6D and Cas6D may not be able to locate the center of the object and accurately obtain the 2D detection box. For OnePose++ and GS-Pose, we think that the cluttered background and noise have important effects on its reconstruction process. Specifically, the decline in the qual-

Table 1 Quantitative comparison of LineMOD datasets

Metrics	Method	Object name								Avg.
		Cat	Duck	Bvise	Cam	Driller	Lamp	Eggbox	Glue	
ADD-0.1d	PvNet	**87.85**	**80.24**	**97.72**	**91.49**	**97.91**	**98.12**	98.74	**95.66**	**93.46**
	Gen6D	58.08	41.22	78.68	66.37	66.90	89.75	70.33	52.70	65.50
	Cas6D	60.58	51.27	86.72	70.10	84.84	93.38	98.78	88.51	79.27
	OnePose++	69.72	42.54	97.30	88.04	92.45	97.80	**99.34**	49.13	79.54
	GS-Pose	84.53	67.89	96.32	88.04	87.12	94.00	99.20	88.90	88.25
	Ours	62.48	58.84	93.82	78.15	85.91	92.31	97.02	74.8	80.41
Prj-5	PvNet	**99.40**	**99.12**	98.31	99.04	**98.24**	97.14	**99.28**	98.82	**98.67**
	Gen6D	95.81	80.85	81.69	90.49	72.94	91.17	97.93	96.04	88.36
	Cas6D	99.00	93.50	93.41	96.27	94.95	96.93	98.31	98.84	96.40
	OnePose++	95.56	97.70	**99.57**	**99.63**	89.81	98.49	98.64	78.68	94.76
	GS-Pose	99.00	97.60	98.50	99.00	91.90	**99.00**	96.90	96.80	96.21
	Ours	98.82	92.21	91.57	98.16	95.85	97.68	98.85	**98.91**	96.51

The best and second-best performing methods are highlighted in bold and underlined, respectively

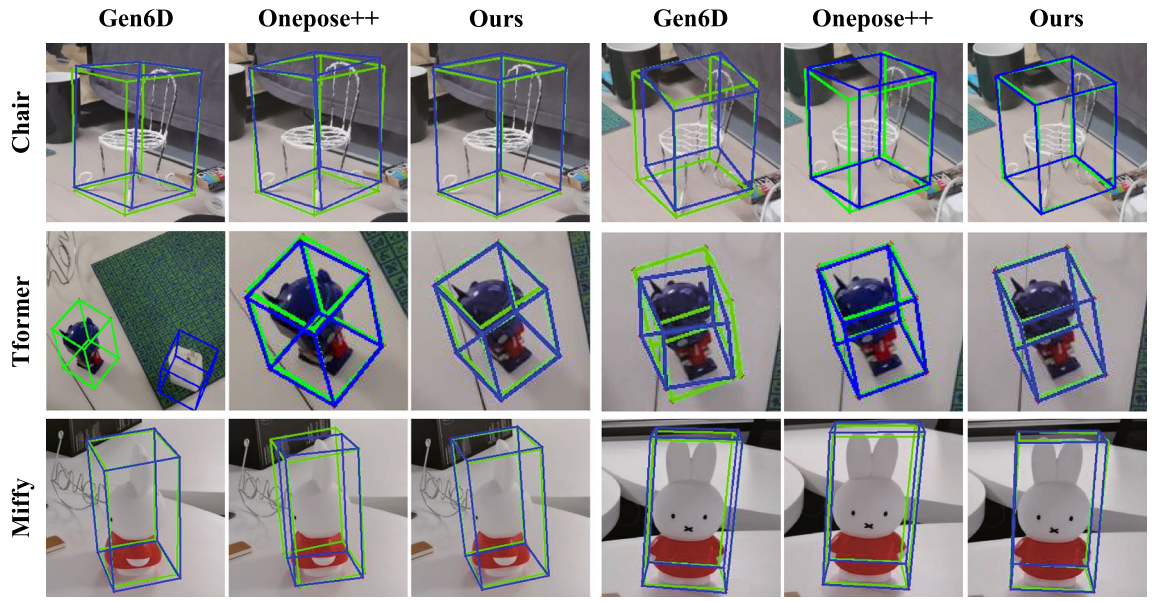

Fig. 7 Qualitative results on the GenMOP dataset. The ground-truth poses are drawn in green, while estimated poses are drawn in blue

ity and accuracy of model reconstruction adversely affects the pose estimation results.

4.5 Ablation studies

4.5.1 Effect of open-world object detector

One of our important improvements is to design an open-world detector using a large model. The detector uses a segmentation model to segment image instances and identifies and locates the target object O through the design of matching scores. We conducted an ablation study on the detector on the GenMOP dataset, and the results are shown in Table 4.

Compared to directly using full-image features for object detection and localization in Gen6D, the design of the open-world detector significantly improves object localization accuracy and enhances pose estimation precision.

4.5.2 Effect of selector and refiner

To investigate the effectiveness of the selector and adaptive multi-scale refiner designs, the ablation study results on GenMOP are presented in Table 5. Compared to the Gen6D selector, our selector shows an improvement of approximately 8%. This improvement is mainly attributed to DINO's robust feature extraction and the global dependencies established by the multi-head self-attention mechanism, allowing

Table 2 Quantitative results on GenMOP dataset

Metrics	Method	Object name					Avg.
		Chair	PlugEN	Piggy	Scissors	TFormer	
ADD-0.1d	PvNet	48.20	2.18	65.43	**43.27**	16.35	35.08
	Gen6D	60.24	18.89	72.86	32.33	63.41	49.54
	Cas6D	**64.00**	**23.36**	**77.87**	36.88	65.47	**53.52**
	OnePose++	57.66	4.85	69.42	32.48	58.6	44.60
	GS-Pose	61.47	20.87	71.54	33.41	61.74	49.81
	Ours	58.74	23.18	76.94	37.57	**66.52**	52.59
Prj-5	PvNet	14.57	32.42	78.89	85.61	59.32	54.16
	Gen6D	54.00	71.93	95.98	93.23	98.74	82.77
	Cas6D	62.50	**78.67**	**99.50**	**97.43**	**99.50**	**87.52**
	OnePose++	60.49	72.54	95.32	80.45	90.57	79.87
	GS-Pose	**63.87**	75.40	97.74	82.77	94.79	82.91
	Ours	58.74	77.28	98.27	96.42	99.17	85.98

We present the comparative results between three generalization methods and one instance-level method The best and second-best performing methods are highlighted in bold and underlined, respectively

Table 3 Quantitative comparison of RBMOP datasets in 32-shot

Method	Object name					Avg.
	Chair	PlugEN	Piggy	Scissors	TFormer	
ADD-0.1d						
Gen6D	39.74	12.63	65.68	29.74	42.35	38.03
Cas6D	43.81	12.24	66.74	31.22	43.89	39.58
OnePose++	30.40	4.27	62.21	27.57	27.37	30.36
GS-Pose	34.73	14.79	64.47	27.88	29.15	34.20
Ours	**47.17**	**30.66**	**72.92**	**36.90**	**46.18**	**46.77**
Prj-5						
Gen6D	37.74	62.42	**83.57**	81.56	87.41	70.54
Cas6D	38.41	69.10	81.74	81.99	88.27	71.90
OnePose++	34.89	14.36	53.47	71.34	78.54	50.52
GS-Pose	36.44	68.83	60.54	72.11	79.84	63.55
Ours	**42.89**	**70.26**	82.18	**82.47**	**89.94**	**73.55**

The best and second-best performing methods are highlighted in bold and underlined, respectively

Table 4 Ablation studies about detector

Method	Localization ACC	Pose ACC	
	mAP.avg	ADD-0.1d.avg	Prj-5.avg
OSOP [69]	40.35	Null	Null
Gen6D	76.42	49.54	82.77
Ours(w/S_a)	81.25	52.17	84.28
Ours(all)	83.64	52.59	85.97

The mAP@ [.5 : .95] (%) is used as localization accuracy metric. (w/S_a) means we only use semantic scores for similarity matching

the network to handle unseen objects and variations in rotation and viewpoint effectively.

Building on the selector, we introduced Gen6D's volume-based refiner and our proposed adaptive multi-scale refiner. Data analysis shows that the refiner plays a very important role in optimizing the pose. Gen6D's volume-based refiner significantly improves the results, mainly due to 3D volume construction and 3D CNN regression. However, our adaptive pose refiner based on multi-scale and multi-similar image information achieves better optimization results.

4.5.3 Effects of number of reference images and running time

To investigate the robustness of GRPoseNet against sparse view challenges, we progressively reduced the full reference views in Linemod ($\approx$ 180) to as few as 8 and evaluated the average ADD-0.1d metric. To better analyze the trends, Fig. 8 illustrates the impact of the number of reference views on various methods.

Table 5 Ablation studies about selector and refiner on GenMOP

Metrics	Method	Object name					Avg.
		Chair	PlugEN	Piggy	Scissors	TFormer	
ADD-0.1d	ObjDesc [70]	3.50	5.14	14.07	1.25	7.54	8.55
	Gen-Selector	14.00	7.48	39.70	16.81	11.51	17.90
	+Vol Ref	**60.24**	18.89	72.86	32.33	63.41	49.54
	Our-Selector	16.21	7.74	41.13	16.78	14.15	19.20
	+Adaptive Ref	58.74	**23.18**	**76.94**	**37.57**	**66.52**	**52.59**
Prj-5	ObjDesc [70]	4.00	10.75	4.52	18.53	8.33	9.23
	Gen-Selector	11.50	40.65	33.17	34.05	64.29	36.73
	+Vol Ref	54.00	71.93	95.98	93.23	98.74	82.77
	Our-Selector	17.18	41.11	42.28	34.42	69.84	40.96
	+Adaptive Ref	**58.74**	**77.28**	**98.27**	**96.42**	**99.17**	**85.97**

The best and second-best performing methods are highlighted in bold and underlined, respectively

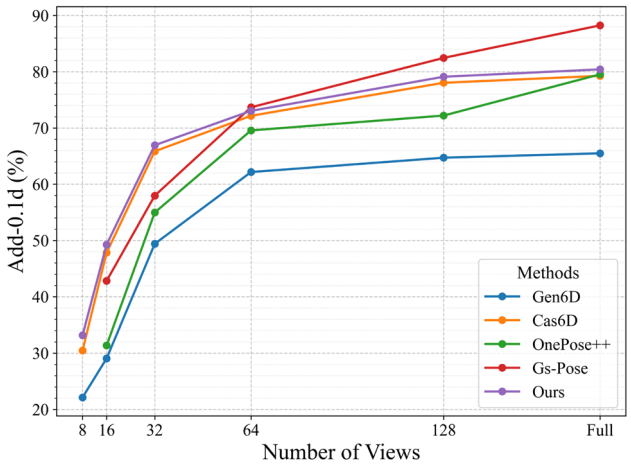

Fig. 8 Performance comparison on the Linemod dataset with different numbers of reference images

Specifically, the model reconstruction-based method GS-Pose achieves optimal performance with 64 reference views, likely due to the independent reference images available for each object in Linemod, which facilitate high-quality model reconstruction. However, our method achieves its best performance when the number of sparse views is reduced to 32. At this stage, significant accuracy degradation is observed for model reconstruction-based methods. Leveraging the cascaded optimization of the multi-scale feature pyramid, Cas6D also exhibits excellent robustness to sparse views. When background-induced variations in detection accuracy are not considered, our method achieves an average improvement of approximately 3% over Cas6D and 24% over Gen6D.

In addition, to investigate the runtime efficiency of our method, we conducted detailed tests on a server equipped with two Intel Xeon Gold 6226R CPUs and two GTX 3090Ti GPUs. Since most of the compared methods are based on the three stages of detection, selection, and refinement, we report the time consumption of each method at each stage for a more comprehensive comparison. The detailed comparison is shown in Fig. 9.

Although our refiner abandons the 3D volumetric feature refinement approach used in Gen6D, the introduction of multi-scale features and Transformer modules does not reduce computation time. However, it is more efficient than the multilayer feature pyramid and cascaded optimization method employed by Cas6D. Our SAM2-based detector accounts for 63% of the total time consumption. Despite incorporating various techniques to reduce redundant mask generation, SAM2's zero-shot full-image segmentation remains relatively time-consuming. Considering that the pre-trained segmentation model can be replaced, we experimented with substituting FastSAM [71] as the pre-trained large segmentation model. This replacement enabled faster pose estimation at the cost of approximately a 6.8% average accuracy loss.

5 Conclusion

In this paper, we propose GRPoseNet, a generalizable and robust 6D object pose estimation method based on sparse RGB images, composed of an open-world detector, a viewpoint selector, and an adaptive multi-scale refiner. This model has good generalization ability for estimating the pose of unseen objects. The detector introduces segmentation and large visual models (SAM2 and DINOV2) to perform zero-shot segmentation and feature extraction, effectively improving the localization and detection accuracy of unseen objects and enhancing the generalization ability of the model. Viewpoint selector and refiner apply multi-scale fusion and multi-head self-attention to fully utilize and enhance sparse RGB image features. By iteratively refining the top 3 most similar reference images and applying adaptive weights, we mitigate the impact of sparse views and estimate the

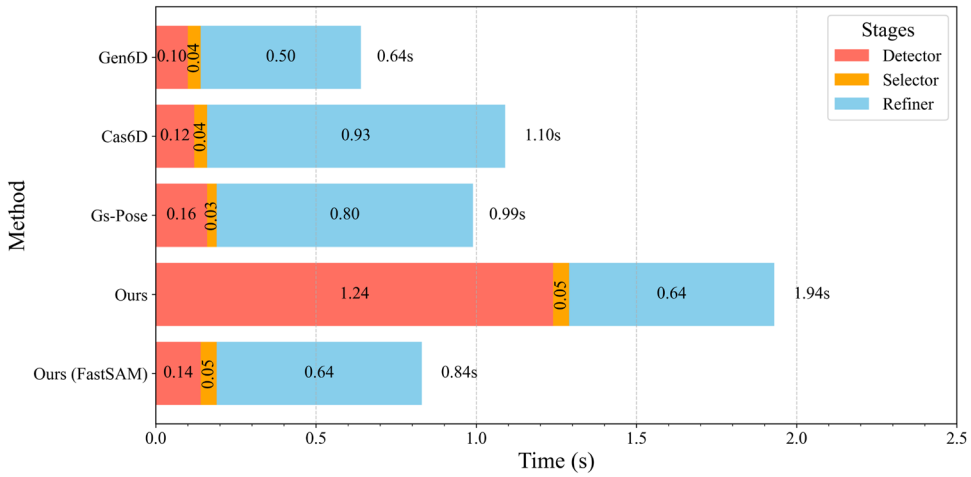

Fig. 9 Average time consumption comparison on Linemod

final object pose. Experiments on LineMOD, GenMOP, and our RBMOP datasets demonstrate that our method achieves state-of-the-art performance, showing excellent generalization and robustness to unseen objects and sparse views.

Acknowledgements This paper is supported by the National Natural Science Foundation of China (No. 62073245) and the Shanghai Science and Technology Innovation Action Plan (22511104900).

Declarations

Conflict of interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Limitation and discussions We rely on reference images with pose labels to achieve pose estimation for unseen objects, but the accuracy is still lower than instance-level methods. Additionally, our detector, selector, and refiner models are independently trained and executed. In the future, we aim to develop an end-to-end version and further reduce the reliance on reference view labels.

References

1. Ali, S.G., Wang, X., Li, P., Jung, Y., Bi, L., Kim, J., Chen, Y., Feng, D.D., Magnenat Thalmann, N., Wang, J., et al.: A systematic review: virtual-reality-based techniques for human exercises and health improvement. Front. Public Health **11**, 1143947 (2023)
2. Boutsi, A.-M., Bakalos, N., Ioannidis, C.: Pose estimation through mask-r cnn and vslam in large-scale outdoors augmented reality. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. **4**, 197–204 (2022)
3. Chen, H., Wang, P., Wang, F., Tian, W., Xiong, L., Li, H.: 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, pp. 2781–2790 (2022)
4. Hoque, S., Xu, S., Maiti, A., Wei, Y., Arafat, M.Y.: Deep learning for 6d pose estimation of objects-a case study for autonomous driving. Expert Syst. Appl. **223**, 119838 (2023)
5. Wang, C., Martín-Martín, R., Xu, D., Lv, J., Lu, C., Fei-Fei, L., Savarese, S., Zhu, Y.: 6-pack: category-level 6d pose tracker with anchor-based keypoints. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 10059–10066. IEEE (2020)
6. Jeon, M.-H., Kim, J., Ryu, J.-H., Kim, A.: Ambiguity-aware multi-object pose optimization for visually-assisted robot manipulation. IEEE Robot. Autom. Lett. **8**(1), 137–144 (2022)
7. Ma, W., Wang, A., Yuille, A., Kortylewski, A.: Robust category-level 6d pose estimation with coarse-to-fine rendering of neural features. In: European Conference on Computer Vision, pp. 492–508. Springer (2022)
8. Zou, L., Huang, Z., Gu, N., Wang, G.: 6d-vit: category-level 6d object pose estimation via transformer-based instance representation learning. IEEE Trans. Image Process. **31**, 6907–6921 (2022)
9. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2642–2651 (2019)
10. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE Trans. Pattern Anal. Mach. Intell. **13**(04), 376–380 (1991)
11. Nguyen, V.N., Hu, Y., Xiao, Y., Salzmann, M., Lepetit, V.: Templates for 3d object pose estimation revisited: generalization to new objects and robustness to occlusions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6771–6780 (2022)
12. Sundermeyer, M., Durner, M., Puang, E.Y., Marton, Z.-C., Vaskevicius, N., Arras, K.O., Triebel, R.: Multi-path learning for object pose estimation across domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13916–13925 (2020)
13. Li, Y., Wang, G., Ji, X., Xiang, Y., Fox, D.: Deepim: deep iterative matching for 6d pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 683–698 (2018)
14. Liu, Y., Wen, Y., Peng, S., Lin, C., Long, X., Komura, T., Wang, W.: Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. In: European Conference on Computer Vision, pp. 298–315. Springer (2022)
15. He, X., Sun, J., Wang, Y., Huang, D., Bao, H., Zhou, X.: Onepose++: keypoint-free one-shot object pose estimation without cad models. Adv. Neural. Inf. Process. Syst. **35**, 35103–35115 (2022)
16. Lin, J., Liu, L., Lu, D., Jia, K.: 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, pp. 27906–27916 (2024)

17. Wen, B., Yang, W., Kautz, J., Birchfield, S.: Foundationpose: unified 6d pose estimation and tracking of novel objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17868–17879 (2024)
18. He, Y., Wang, Y., Fan, H., Sun, J., Chen, Q.: Fs6d: few-shot 6d pose estimation of novel objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6814–6824 (2022)
19. Yen-Chen, L., Florence, P., Barron, J.T., Rodriguez, A., Isola, P., Lin, T.-Y.: inerf: inverting neural radiance fields for pose estimation. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1323–1330. IEEE (2021)
20. Li, F., Vutukur, S.R., Yu, H., Shugurov, I., Busam, B., Yang, S., Ilic, S.: 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, pp. 2123–2133 (2023)
21. Sun, Y., Wang, X., Zhang, Y., Zhang, J., Jiang, C., Guo, Y., Wang, F.: icomma: inverting 3d gaussians splatting for camera pose estimation via comparing and matching. arXiv preprint arXiv:2312.09031 (2023)
22. Cai, D., Heikkilä, J., Rahtu, E.: Gs-pose: Cascaded framework for generalizable segmentation-based 6d object pose estimation. arXiv preprint arXiv:2403.10683 (2024)
23. Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N.: Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In: Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5–9, 2012, Revised Selected Papers, Part I 11, pp. 548–562. Springer (2013)
24. Rad, M., Lepetit, V.: Bb8: a scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3828–3836 (2017)
25. Wei, L., Xie, F., Sun, L., Chen, J., Zhang, Z.: A modal fusion network with dual attention mechanism for 6d pose estimation. Vis. Comput. **40**(10), 7411–7425 (2024)
26. Chen, W., Duan, J., Basevi, H., Chang, H.J., Leonardis, A.: Pointposenet: point pose network for robust 6d object pose estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2824–2833 (2020)
27. Liu, S., Xu, F., Wu, C., Chi, J., Yu, X., Wei, L., Leng, C.: Cmt-6d: a lightweight iterative 6dof pose estimation network based on cross-modal transformer. Vis. Comput. **41**, 2011–2027 (2025)
28. Glasner, D., Galun, M., Alpert, S., Basri, R., Shakhnarovich, G.: Aware object detection and pose estimation. In: 2011 International Conference on Computer Vision, pp. 1275–1282. IEEE (2011)
29. Ullah, F., Wei, W., Fan, Z., Yu, Q.: 6d object pose estimation based on dense convolutional object center voting with improved accuracy and efficiency. Vis. Comput. **40**(8), 5421–5434 (2024)
30. Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H.: Pvnet: pixel-wise voting network for 6dof pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4561–4570 (2019)
31. Wang, C., Xu, D., Zhu, Y., Martín-Martín, R., Lu, C., Fei-Fei, L., Savarese, S.: Densefusion: 6d object pose estimation by iterative dense fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3343–3352 (2019)
32. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: Posecnn: a convolutional neural network for 6d object pose estimation in cluttered scenes. arXiv preprint arXiv:1711.00199 (2017)
33. Li, Z., Wang, G., Ji, X.: Cdpn: coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7678–7687 (2019)
34. Zeng, A., Yu, K.-T., Song, S., Suo, D., Walker, E., Rodriguez, A., Xiao, J.: Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 1386–1383. IEEE (2017)
35. Hinterstoisser, S., Holzer, S., Cagniart, C., Ilic, S., Konolige, K., Navab, N., Lepetit, V.: Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In: 2011 International Conference on Computer Vision, pp. 858–865. IEEE (2011)
36. Thalhammer, S., Hönig, P., Weibel, J.-B., Vincze, M.: Open challenges for monocular single-shot 6d object pose estimation. arXiv preprint arXiv:2302.11827 (2023)
37. Peng, W., Yan, J., Wen, H., Sun, Y.: Self-supervised category-level 6d object pose estimation with deep implicit shape representation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 2082–2090 (2022)
38. Liu, J., Sun, W., Liu, C., Zhang, X., Fu, Q.: Robotic continuous grasping system by shape transformer-guided multi-object category-level 6d pose estimation. IEEE Trans. Ind. Inform. **19**(11), 11171–11181 (2023)
39. Chen, K., James, S., Sui, C., Liu, Y.-H., Abbeel, P., Dou, Q.: Stereopose: category-level 6d transparent object pose estimation from stereo images via back-view nocs. In: 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 2855–2861. IEEE (2023)
40. Zhang, H., Opipari, A., Chen, X., Zhu, J., Yu, Z., Jenkins, O.C.: Transnet: category-level transparent object pose estimation. In: European Conference on Computer Vision, pp. 148–164. Springer (2022)
41. He, Y., Fan, H., Huang, H., Chen, Q., Sun, J.: Towards self-supervised category-level object pose and size estimation. arXiv preprint arXiv:2203.02884 (2022)
42. Tian, M., Ang, M.H., Lee, G.H.: 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, pp. 530–546. Springer (2020)
43. Chen, W., Jia, X., Chang, H.J., Duan, J., Shen, L., Leonardis, A.: 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, pp. 1581–1590 (2021)
44. Park, K., Mousavian, A., Xiang, Y., Fox, D.: Latentfusion: end-to-end differentiable reconstruction and rendering for unseen object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10710–10719 (2020)
45. Zhang, B., Sheng, B., Li, P., Lee, T.-Y.: Depth of field rendering using multilayer-neighborhood optimization. IEEE Trans. Vis. Comput. Graphics **26**(8), 2546–2559 (2019)
46. Kamel, A., Sheng, B., Yang, P., Li, P., Shen, R., Feng, D.D.: Deep convolutional neural networks for human action recognition using depth maps and postures. IEEE Trans. Syst. Man Cybern. Syst. **49**(9), 1806–1819 (2018)
47. Sun, J., Wang, Z., Zhang, S., He, X., Zhao, H., Zhang, G., Zhou, X.: Onepose: one-shot object pose estimation without cad models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6825–6834 (2022)
48. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: representing scenes as neural radiance fields for view synthesis. Commun. ACM **65**(1), 99–106 (2021)
49. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graphics **42**(4), 139:1–139:14 (2023)

50. Yen-Chen, L., Florence, P., Barron, J.T., Rodriguez, A., Isola, P., Lin, T.-Y.: inerf: inverting neural radiance fields for pose estimation. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1323–1330. IEEE (2021)
51. Li, F., Vutukur, S.R., Yu, H., Shugurov, I., Busam, B., Yang, S., Ilic, S.: 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, pp. 2123–2133 (2023)
52. Lin, Y., Müller, T., Tremblay, J., Wen, B., Tyree, S., Evans, A., Vela, P.A., Birchfield, S.: Parallel inversion of neural radiance fields for robust pose estimation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 9377–9384. IEEE (2023)
53. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: Deep vit features as dense visual descriptors. **2**(3), 4 (2021). arXiv:2112.05814
54. Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., Yi, K.M.: Unsupervised semantic correspondence using stable diffusion. Adv. Neural Inform. Process. Syst. **36**, 15 (2024)
55. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9650–9660 (2021)
56. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
57. Pan, P., Fan, Z., Feng, B.Y., Wang, P., Li, C., Wang, Z.: Learning to estimate 6dof pose from limited data: a few-shot, generalizable approach using rgb images. In: 2024 International Conference on 3D Vision (3DV), pp. 1059–1071. IEEE (2024)
58. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
59. Jiang, H., Karpur, A., Cao, B., Huang, Q., Araujo, A.: Omniglue: generalizable feature matching with foundation model guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19865–19875 (2024)
60. Tumanyan, N., Singer, A., Bagon, S., Dekel, T.: Dino-tracker: Taming dino for self-supervised point tracking in a single video. arXiv preprint arXiv:2403.14548 (2024)
61. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)
62. Xie, Z., Guan, B., Jiang, W., Yi, M., Ding, Y., Lu, H., Zhang, L.: Pa-sam: Prompt adapter sam for high-quality image segmentation. arXiv preprint arXiv:2401.13051 (2024)
63. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot. arXiv preprint arXiv:2305.03048 (2023)
64. Fan, Z., Pan, P., Wang, P., Jiang, Y., Xu, D., Wang, Z.: Pope: 6-dof promptable pose estimation of any object in any scene with one reference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7771–7781 (2024)
65. Nguyen, V.N., Groueix, T., Ponimatkin, G., Lepetit, V., Hodan, T.: Cnos: a strong baseline for cad-based novel object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2134–2140 (2023)
66. Pan, P., Fan, Z., Feng, B.Y., Wang, P., Li, C., Wang, Z.: Learning to estimate 6dof pose from limited data: a few-shot, generalizable approach using rgb images. In: 2024 International Conference on 3D Vision (3DV), pp. 1059–1071 (2024). IEEE
67. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
68. Cai, D., Heikkilä, J., Rahtu, E.: Gs-pose: cascaded framework for generalizable segmentation-based 6d object pose estimation. arXiv preprint arXiv:2403.10683 (2024)
69. Shugurov, I., Li, F., Busam, B., Ilic, S.: Osop: a multi-stage one shot object pose estimation framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6835–6844 (2022)
70. Wohlhart, P., Lepetit, V.: Learning descriptors for object recognition and 3d pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3109–3118 (2015)
71. Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., Wang, J.: Fast segment anything. arXiv preprint arXiv:2306.12156 (2023)

Wubin Shi received the bachelor's degree from Zhengzhou University in 2019 and successfully gained a master's degree from University of Chinese Academy of Sciences in 2022. Currently, he is pursuing Ph.D. degree at Southeast University and dedication to the deep research in the field of robot vision.

Shaoyan Gai received the Ph.D. degree from Southeast University, Nanjing, China, in 2008. He is currently an Associate Professor and a Ph.D. Advisor with Southeast University. His main research interests include 3-D measurement and 3-D face recognition.

Feipeng Da received the Ph.D. degree from the School of Automation, Southeast University, Nanjing, China, in 1998. He is currently a Professor with the School of Automation, Southeast University. He has published an academic monograph and authored or coauthored over 150 high quality articles, of which are retrieved by Science Citation Index (SCI), Engineering Index (EI), and Index to Scientific and Technical Proceedings (ISTP) more than 100 times. He has 40 authorized invention patents, one authorized patent for utility models, four software copyrights, and three international invention patents Patent Cooperation Treaty (PCT) applied. Professor Da also serves as a Reviewer for the journals from different areas, such as IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS-I: REGULAR PAPERS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS-II: EXPRESS BRIEFS, Physics Letters A, Neural Networks, Pattern Recognition, Optics Express, Optics Letters, and Optics and Lasers in Engineering.

Zeyu Cai received his B.Sc. degree from Northwestern Polytechnical University in 2014 and the M.Sc. degree from the Chinese Academy of Agricultural Sciences in 2021. He is currently pursuing the Ph.D. degree in Southeast University of Nanjing, China, with a focus on computational imaging and deep learning in computer vision.

Jiaoling Wang received the Ph.D. degree from the Chinese Academy of Agricultural Sciences (CAAS) and completed postdoctoral research at Northwest A&F University. He currently serves as an Associate Researcher in Nanjing Institute of Agricultural Mechanization, Ministry of Agriculture and Rural Affairs. His research focuses on intelligent harvesting technologies for edible fungi and the development of smart processing equipment for agricultural products, specializing in the integration of intelligent systems with agricultural mechanization solutions.