The Visual Computer (2024) 40:5187–5200
https://doi.org/10.1007/s00371-023-03097-1

ORIGINAL ARTICLE

RT-less: a multi-scene RGB dataset for 6D pose estimation of reflective texture-less objects

Xinyue Zhao[1] · Quanzhi Li[1] · Yue Chao[1] · Quanyou Wang[1] · Zaixing He[1] · Dong Liang[2]

Accepted: 9 September 2023 / Published online: 18 October 2023

Abstract

The 6D (6 Degree of freedom) pose estimation (or pose measurement) of machined reflective texture-less objects, which are common in industry, is a significant but challenging technique. It has attracted increasing attention in academia and industry. However, it is difficult to obtain suitable public datasets of such objects, which makes relevant studies inconvenient. Thus, we proposed the Reflective Texture-Less (RT-Less) object dataset, which is a new public dataset of reflective texture-less metal parts for pose estimation research. The dataset contains 38 machined texture-less reflective metal parts in total. Different parts demonstrate the symmetry and similarity of shape and size. The dataset contains 289 K RGB images and the same number of masks, including 25,080 real images, 250,800 synthetic images in the training set, and 13,312 real images captured in 32 different scenes in the test set. The dataset also provides accurate ground truth poses, bounding-box annotations and masks for these images, which makes RT-Less suitable for object detection and instance segmentation. To improve the accuracy of the ground truth, an iterative pose optimization method using only RGB images is proposed. Baselines of the state-of-the-art pose estimation methods are provided for further comparative studies. The dataset and results of baselines are available at: http://www.zju-rtl.cn/RT-Less/.

Keywords Pose estimation · Pose measurement · Object detection · Instance segmentation · Reflective · Texture-less · Machine vision

1 Introduction

6D (Degree of freedom) pose estimation has been widely used in automation, auto-driving, unmanned air vehicle (UAV) reconnaissance, intelligent medical and service robots and other applications [1–3]. The current techniques perform well for rich textured, obviously colored and rough-surfaced objects [4–9]. However, many objects such as metal parts are usually strongly reflective and texture-less, and mechanical parts of the same batch often do not have an obvious difference in color. It is far from easy for current technology to detect and estimate poses for these objects. The strong reflectivity of machined metal parts makes it difficult to acquire reliable depth information via 3D vision. The lack of reliable texture and color information on the surface of metal parts makes it difficult to detect feature points on them. However, texture-less and reflective objects are the most common type in industry. Accurate pose estimation of these objects is necessary for automatic and intelligent production. Thus, it is crucial to study the pose estimation of these objects. However, there are only some small and undisclosed datasets [5, 6] of such objects, which is not conducive for researchers to conduct in-depth research and comparison and hinders the technology progress.

Therefore, in this paper, a new public dataset called the Reflective Texture-Less (RT-Less) object dataset is proposed to facilitate research on 6D pose estimation. Previously, for the non-reflective objects, many methods used RGB-D images to estimate poses, and many RGB-D public datasets were issued. However, since it is difficult to acquire reliable depth information of reflective objects, a mainstream way is to use RGB images for pose estimation. Therefore, the generated RT-Less is an RGB dataset. It is divided into a training set and a testing set. The training set contains images and

✉ Zaixing He
zaixinghe@zju.edu.cn

[1] The School of Mechanical Engineering, The State Key Lab of Fluid Power and Mechatronic Systems, Zhejiang University, Hangzhou 310027, China

[2] The College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

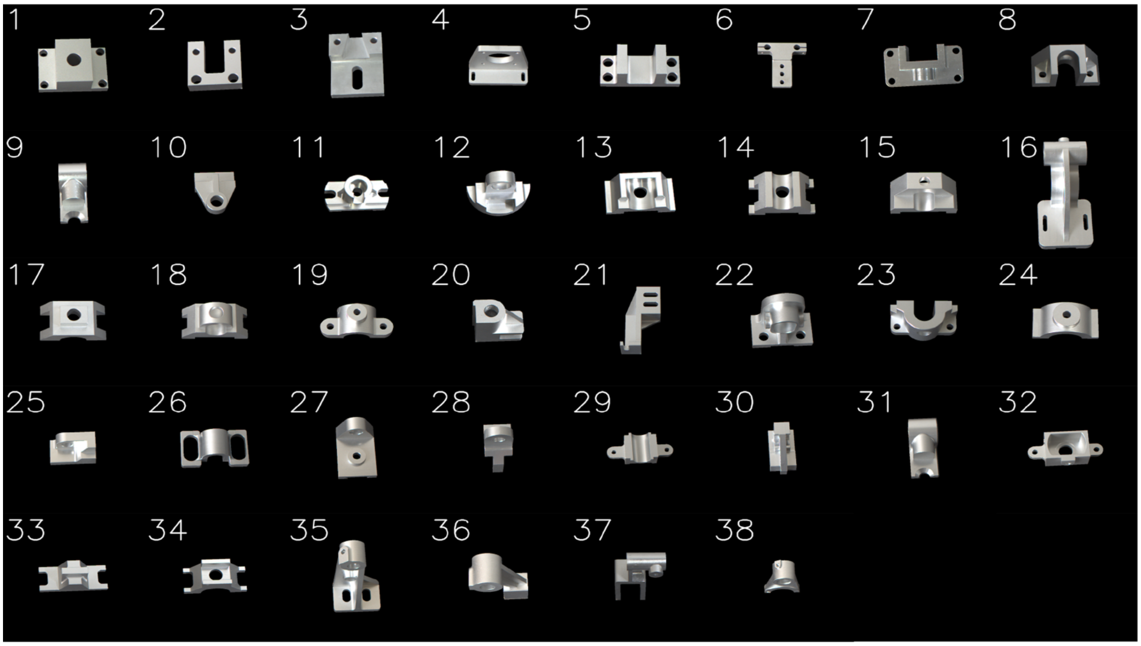


Fig. 1 Sample training images of 38 objects

3D models of 38 metal parts that cover typical features, e.g., large planes, surfaces, bevels, and circular holes. Figure 1 shows the samples of training images of 38 objects. All these industrial parts have strong reflectivity and no regular texture.

RT-Less is a dataset for the industrial domain. To ensure the authenticity of the industrial attributes of objects, the design and processing of all objects in the dataset are from the machining factories of metal parts. During the machining process, some objects were chamfered, and a more complex structure was designed, such as parts No. 22, No. 37, and No. 11. To simulate the small difference in shape among different parts in the actual industry, several parts have high similarity, e.g., parts No. 9 and No. 31 are similar in shape.

All images in RT-Less are almost obtained and annotated automatically. For each training and testing image, there is a corresponding mask image and a 2D bounding box so that this dataset can be adapted to object detection and pose estimation methods that require masks. In addition to RGB images, to facilitate most of the current studies that combine RGB images with CAD models, the dataset also provides three types of CAD model files.

All objects in our dataset come from the real production line and are processed by standard processes. In addition, we imitate the actual industrial scenarios in terms of parts placement, parts types and shape design, lighting setup and background. Since the features of machined parts are relatively simple, the detection of similar parts in the actual industry is common, and the testing set introduces multiple similar parts with identical features in some scenes to better mimic the real situation. By combining different situations, there were 32 scenes in the testing set, as shown in Fig. 2. Every target object was overlaid with colored 3D object models at the ground truth poses. There are also some parts that are not overlaid, which are similarly looking distractor parts and are set up to enhance the difficulty of some scenes, in line with the actual industrial production situation. We have divided the testing set into three levels of difficulty according to lighting conditions, part complexity, and background difficulty, indicated by the color tag in the upper left corner of each scene in Fig. 2. Green represents difficulty level 1, orange represents difficulty level 2, and red represents difficulty level 3. The higher the difficulty level, the greater the challenge of the scene. In terms of illumination, there are three scenes: bright natural illumination, dark illumination and bright artificial illumination, as shown in Fig. 3, which correspond to sunny environments, cloudy environments and artificial illumination at night. Finally, in terms of the background of images, the training set has only one pure black background. In the testing set, there are four backgrounds with different difficulties to verify the robustness of different methods for background changes, as shown in Fig. 4.

Based on the design of objects and scenes, RT-Less is close to industrial reality, which can encourage researchers to research in related fields. Similar to the real world, different scenes have different difficulties; therefore, testing images in different scenes also has different difficulties.

In summary, the motivation of this paper is for the convenience of scholars in the field to address the pose estimation problem of reflective texture-less parts, which is of significant important in industrial manufacturing. The main contributions include:

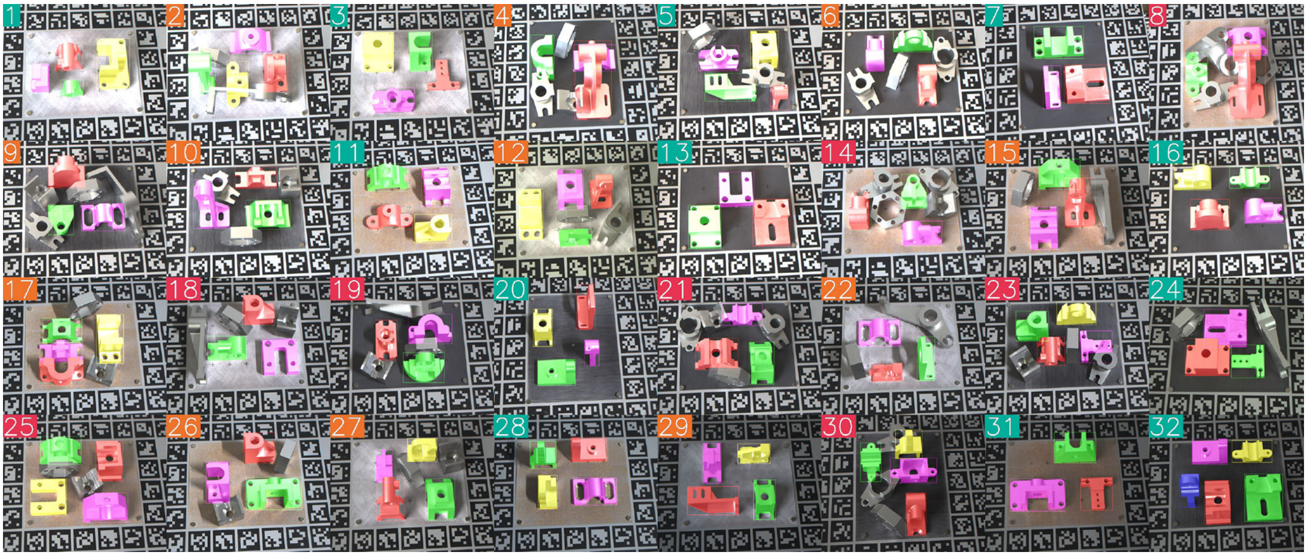


Fig. 2 Sample images of 32 scenes overlaid with colored 3D models with the ground truth poses (only target objects were overlaid). The number of each scene is shown in the upper left corner, and the color represents the difficulty level, with green for level 1, orange for level 2, and red for level 3

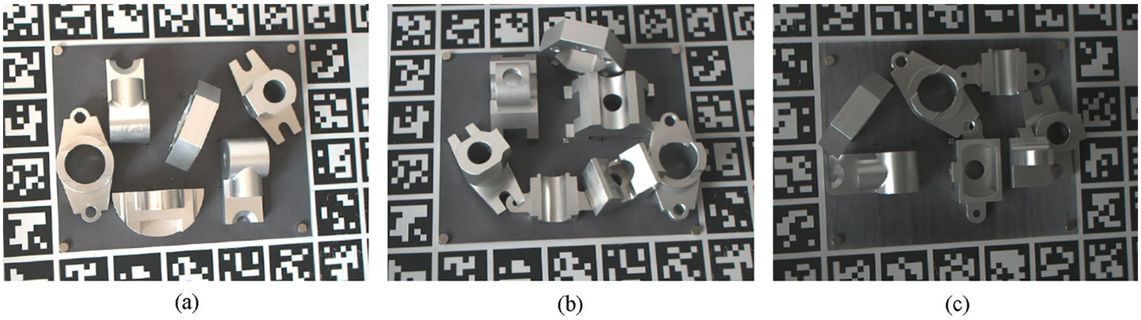


Fig. 3 Different lighting conditions in the testing set. **a** Bright artificial illumination. **b** Bright natural illumination. **c** Dark illumination

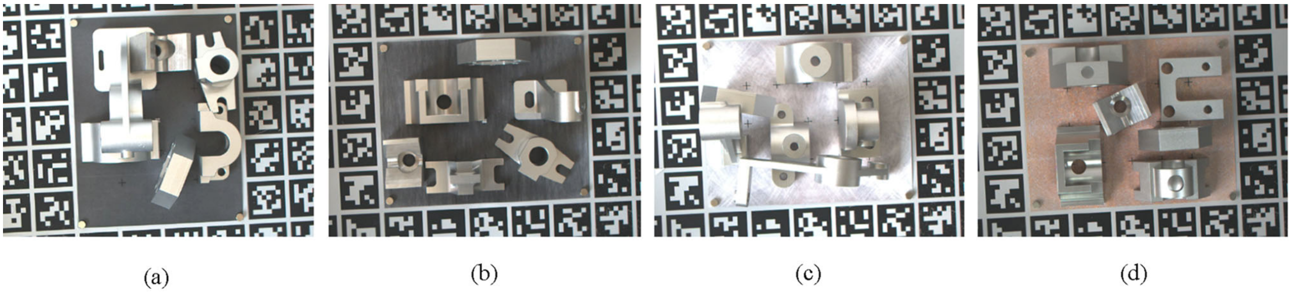


Fig. 4 Different backgrounds in the testing set. **a** Black background. **b** Black textured background. **c** Reflective metal background. **d** Rust background

1. A novel large-scale public dataset is constructed. The objects are rich in various types of reflective metal parts, different from household objects, in various scenes. Such a database is meaningful to advancing the development of pose estimation techniques.
2. An iterative pose optimization method is proposed to improve the accuracy of pose annotations (the ground truth) using only RGB images. Accurate ground truth pose and bounding-box are proposed. It is difficult to obtain an accurate pose using only RGB information. We propose an iterative pose optimization method based on ArUco marker and RANSAC to get the accurate poses.
3. Baselines of the state-of-the-art pose estimation methods are provided for further comparative studies.

The remainder of this article is organized as follows. Section 2 reviews the relevant datasets. Section 3 describes the details of our methods of obtaining and post-processing the dataset. Section 4 introduces suggestions for usage. Section 5 shows a simple experiment on the dataset. Section 6 summarizes the full text.

2 Related datasets

In recent years, with the development of computer vision and deep learning, increasingly many pose estimation datasets have been proposed, which can be divided into non-industrial objects and industrial objects. Although there has also been attention paid to industrial scene objects in recent years, the focus on datasets for reflective metal objects is still insufficient [9]. The following are some popular pose estimation datasets constructed in recent years, which we categorize into non-industrial and industrial datasets.

2.1 Non-industrial datasets

Ru-APC [10]**:** This dataset contains 25 APC [11] objects. The dataset labels provide the ground truth poses of 25 APC objects and 10,368 images in total. Each image has a corresponding depth map. The dataset is oriented to warehouse logistics capture. Most of the objects in the dataset are daily necessities with rich textures.

Linemod/Linemod-Occluded [12, 13]**:** Linemod is a popular dataset in the field of pose estimation that contains 15 textureless household objects with identifiable colors, shapes and sizes, and obvious clutter interference. Linemod-Occluded is the upgrade of Linemod, which introduces more challenging images with various occlusion difficulty levels. The dataset contains 10 k images of 20 objects under three different lighting conditions. However, Linemod and Linemod-Occluded contain only objects and scenes for the family environment, which are different from those in industrial production.

YCB-Video [14]**:** YCB-Video was recently proposed with the POSECNN detector. The dataset contains 21 objects provided by the YCB [15], 133,827 testing images, and 92 video frames. Large perspective changes and a large number of images containing clutter and occlusion are the main characteristics of the dataset. However, all objects and scenes in this dataset are from daily life. The objects do not have the reflective and texture-less properties of typical metal parts.

Other Dataset: ESKO6d [16] provided the object of glass and ceramic storage containers in the kitchen scene. Most of the objects have texture-less, glossy or transparent glass properties. The scene includes cabinets, drawers or dishwashers, and the testing images are largely occluded [6] and RAPID-LR [5] are RGB datasets, which contain fewer objects, each testing image has only one object, and there are less background changes. HomebrewedDB [17] includes 33 objects, including 17 toys, 8 household common objects and 8 industrial common objects. For the testing set, 13 scenes with different difficulties are taken. Objects include texture and non-texture, occlusion and lighting at different levels, and changes in appearance of objects. IC-MI [18] contains six objects, two of which are texture-less and four are textured, and 700 real testing images. Each testing image contains multiple objects with clutter and slight occlusion. FAT [19] contains 21 household objects from the daily life environment of the YCB dataset [15] and 60 K images in total, which is a synthetic dataset and large in scale. However, it is mainly for daily life objects, and the object attributes are notably different from mechanical parts. TUD-L and TYO-L [20] set different lighting levels for household objects in different environments. The testing image contains multiple objects, and there are also background changes. RobotP [21] contains not only images and ground truth but also 3D models, object masks and bounding boxes, and synthetic images, but it deals with domestic scenes.

The above-mentioned datasets are all from non-industrial scenarios, mostly consisting of common household objects. These datasets do not meet the requirements of today's industrial applications.

2.2 Industrial datasets

ROBI [22, 23]**:** ROBI is a public dataset for 6D object pose estimation and multi-view depth fusion in robotic bin-picking scenarios. It includes 7 small reflective metallic connection accessories, and it has a total of 8 K images.

MVTec ITODD: MVTec ITODD [24] contains 28 objects with different characteristics. Among them, two objects, "clamp big" and "clamp small", out of the 28 are machined metal part with high reflectivity. The objects with realistic industrial settings contain more than 800 images, during which a few of them are public.

T-Less [25]**:** T-less is a popular dataset with 30 industry-related objects and has no obvious texture or identifiable color. The objects exhibit symmetry and similarity in shape and size, and a few are composed of other objects. It includes images from three different sensors and two types of 3D object models. The T-Less dataset is very rich in objects and control settings, but it focuses on non-reflective objects.

Siléane Dataset [26]: This dataset contains 2601 annotated scenes depicting various numbers of object instances in bulk, consisting of both synthetic and real data. The synthetic data is generated to simulate the inside of a bin where a random number of objects have been dropped (between 0 and 11), while the real data consists of scenes with objects lying on different surfaces at varying distances from the camera. The dataset also includes a synthetic dataset for validation and an additional dataset targeting object detection and pose estimation in cluttered environments. This dataset is primarily designed for evaluating object detection and pose estimation methods based on depth or RGBD data.

Other Dataset: Both Fraunhofer IPA Bin-Picking [27] and IC-BIN [28] are for bin-picking, whose object repeatedly appears in the testing image and is seriously occluded, but they all contain fewer objects. S2R-Pick [29] proposed a method for generating realistic synthetic images using 3D model rendering, which can quickly create large amounts of rendering datasets. By generating semi-synthetic images through copy and paste strategy [30], it can effectively reduce the domain gap between real images and rendered images. Using the generated dataset for training, it achieved good results in the bin picking task for reflective objects. The ILS dataset [31] provides 6 different shapes and sizes of industrial objects with reflective properties, including T-screw, Wrench, Tri-sleeve, L-connecter, Valve and Handle. However, the dataset is used for performing a bin picking task and hence does not provide detailed pose information for the target objects.

The datasets mentioned above were designed for industrial objects, but primarily for non-reflective ones. While [29] and [31] proposed datasets for reflective industrial objects, these were mainly focused on bin picking tasks. However, RT-Less differs from them in several aspects. The unique characteristics of RT-Less will be introduced in detail in Sect. 3.

3 RT-less dataset

RT-Less has the following characteristics:

- It contains a large number of training and testing images of reflective texture-less metal parts.
- In the image acquisition phase, camera placement uses the eye-in-hand method to simulate a real industrial perspective.
- The testing scenes are set to simulate the real situation in many aspects, such as lighting, background, object design, and placement.
- Accurate ground truth poses and bounding box annotations for each object are provided. Every image has a corresponding mask image.
- In addition to RGB images, three types of formats of CAD models were provided to assist training.
- Testing images are provided with varying levels of difficulty. The testing images contain many different situations to evaluate the method in different aspects.

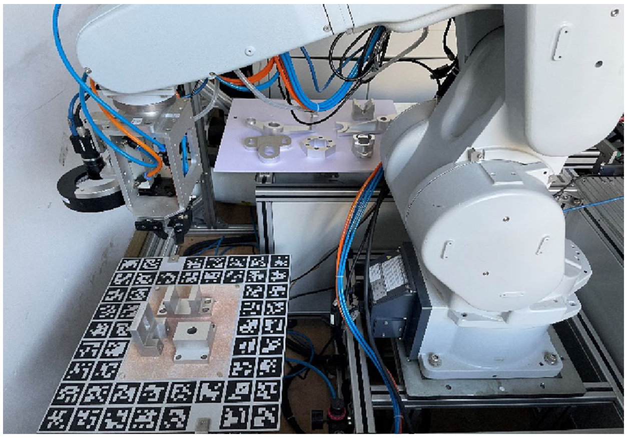


Fig. 5 Image acquisition equipment for RT-Less

The rest describes the dataset preparation process, including the image acquisition system, camera calibration, 3D models and ground truth poses, and bounding-box annotation generation.

3.1 Image acquisition system

Using the device in Fig. 5, training and testing images were acquired. It consists of an MV-CA050-11UC industrial camera with five million pixels, MELFA RV13FD 6-DoF manipulator, turntable and pose calibration board fixed above it. The camera is fixed on the manipulator in eye-in-hand form to obtain the most realistic image with the motion of the manipulator. Below the manipulator is a pose calibration board, which can rotate 360 degrees driven by the turntable under it, and parts are placed on the board.

To ensure that all target parts are in the camera field of vision, the manipulator was controlled to ensure that the camera center moved at a quarter of the space sphere whose center is the rotating center of the turntable. There are three space spheres: the sphere diameters of the space spheres are 750 mm, 800 mm and 850 mm, and the spherical angle is 130°. Through these settings, the camera is controlled to always shoot in the direction of the parts, as shown in Fig. 6. Simultaneously, to ensure the coordination of the hardware of the entire shooting system, a communication control system was constructed to transmit signals among different hardware to complete controllable automatic photography.

3.2 Calibration of camera

The intrinsic and distortion parameters of the camera were calibrated using the calibration method by Zhang [32]. Each calibration takes at least 28 standard black and white checkerboard photographs. The reprojection error of each calibration is controlled below 0.15 pixels.

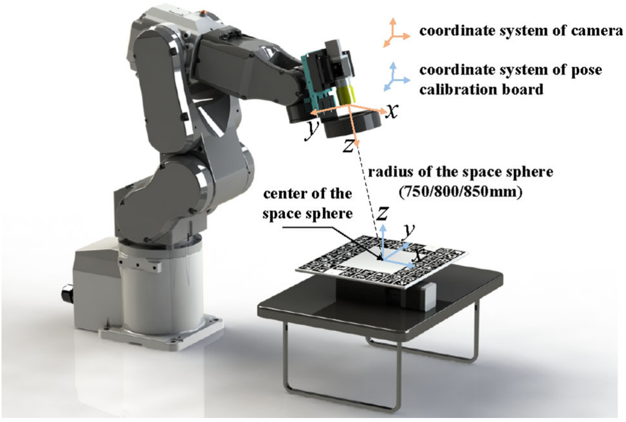

coordinate system of camera
coordinate system of pose calibration board
x
y
z
radius of the space sphere (750/800/850mm)
center of the space sphere
z
y
x

Fig. 6 Keeping the z-axis of the camera coordinate system passing through the center of the calibration plate (also the center of the sampling space sphere) during the acquisition

External parameter calibration of the camera is realized by a pose calibration board with ArUco [33] markers. The ArUco markers are drawn on the aluminum plate by a laser light drawing process to ensure higher accuracy and clarity. The aluminum plate is bonded to the glass substrate after oxidation. The glass substrate can ensure high overall flatness, and the aluminum panel has non-reflective and opaque physical properties to ensure higher calibration accuracy. In the external parameter calibration process, by detecting the image coordinate information of four corners of each ArUco marker on the panel, the external parameter matrix can be obtained through PnP (Perspective-n-Point).

In the automatic calibration process of external camera parameters, we have improved Garrido [33] to obtain pose precisely only using RGB images and ArUco markers. There are two main optimizations which called iterative pose optimization:

1. Corner extraction optimization: First, obtain the set of ArUco marker corner p_{ori}^{mn} through Garrido to generate the ROI (Regions of Interest) o_{mn} (where m is ArUcoId and n = 1, 2, 3, 4, similarly hereinafter, where p_1 and p_2 represent two points on an edge, the subscripts x, y represent the horizontal and vertical coordinates). Then, the edge l_{mn} is obtained within o_{mn} with less interference, respectively. Finally, the intersection point p_{end}^{mn} is obtained by extending l_{mn}, and we can match with the corresponding 3d points to form 2D-3D point pairs $(2d, 3d)_{ori}$ because we already know the 3D coordinates of each ArUco marker.
2. Iteration and fusion of poses: First, based on ransac to filter out the point pairs with large errors in $(2d, 3d)_{ori}$ to get $(2d, 3d)_{\mathrm{end}}$. Then, group the $(2d, 3d)_{\mathrm{end}}$ and calculate the pose using PnP for each group. Finally, fuse the results of each group to get the final camera external reference $\mathrm{T}_{\mathrm{c}}^{b}$, which is the relative poses of the camera and the calibration plate. In these processes, we can extract more accurate key points and resist the interference of bad points to improve the results.

As shown in Fig. 7, better results can be obtained compared with not using these two optimizations. And these processes are summarized in Algorithm 1.

3.3 Images and 3D models

The training and testing images are the core of our dataset. In the shooting process of the testing set, there are 104 points in the motion space of the manipulator. These points are distributed on different layers of 1/4 space spheres. By rotating the bottom turntable with the center of the sphere as the rotation center, the 1/4 space spheres are complemented, and the 360° full-view image is obtained. The turntable position has four changes; the position of camera changes with the manipulator in 104 changes; there are 32 scenes and $104 \times 4 \times 32 = 13{,}312$ images in total. In the shooting process of the training set, 11 points in the motion space of the manipulator are distributed on a circle, and the 360° full-view images are captured by rotating the bottom turntable. The turntable

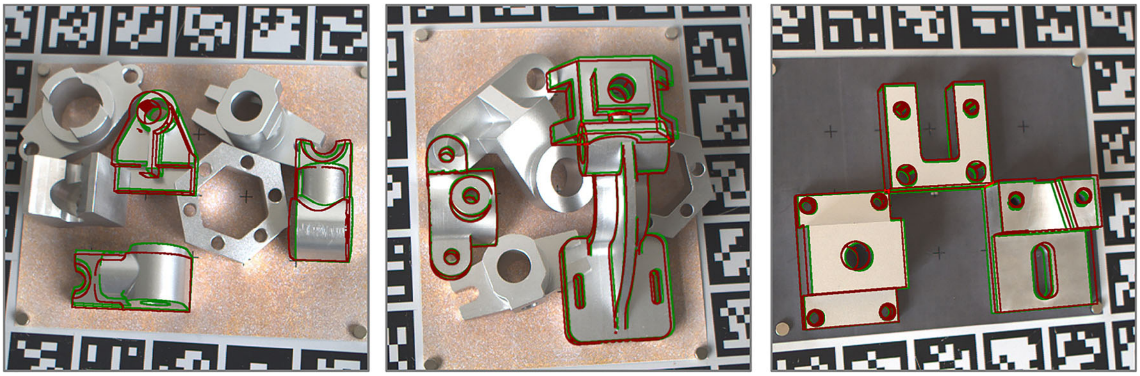
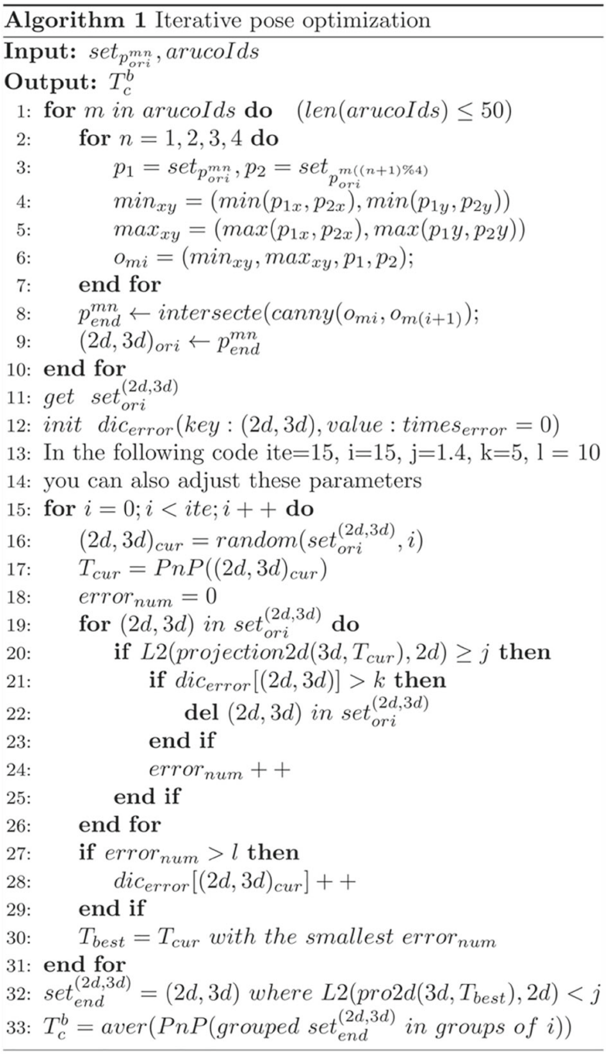


Fig. 7 Ground truth comparison before (red) and after (green) optimization

position has 60 types of changes, and the camera position has 11 changes. The training set contains 38 models with 60 × 11 × 38 = 25,080 images in total.



```
Algorithm 1 Iterative pose optimization
Input: $set_{p_{ori}^{mn}}, arucoIds$
Output: $T_c^b$
for $m$ in $arucoIds$ do   $(len(arucoIds) \leq 50)$
    for $n = 1, 2, 3, 4$ do
        $p_1 = set_{p_{ori}^{mn}}, p_2 = set_{p_{ori}^{m((n+1)\%4)}}$
        $min_{xy} = (min(p_{1x}, p_{2x}), min(p_{1y}, p_{2y}))$
        $max_{xy} = (max(p_{1x}, p_{2x}), max(p_1y, p_2y))$
        $o_{mi} = (min_{xy}, max_{xy}, p_1, p_2);$
    end for
    $p_{end}^{mn} \leftarrow intersecte(canny(o_{mi}, o_{m(i+1)});$
    $(2d, 3d)_{ori} \leftarrow p_{end}^{mn}$
end for
$get\ \ set_{ori}^{(2d,3d)}$
$init\ \ dic_{error}(key : (2d, 3d), value : times_{error} = 0)$
In the following code ite=15, i=15, j=1.4, k=5, l = 10
you can also adjust these parameters
for $i = 0; i < ite; i++$ do
    $(2d, 3d)_{cur} = random(set_{ori}^{(2d,3d)}, i)$
    $T_{cur} = PnP((2d, 3d)_{cur})$
    $error_{num} = 0$
    for $(2d, 3d)$ in $set_{ori}^{(2d,3d)}$ do
        if $L2(projection2d(3d, T_{cur}), 2d) \geq j$ then
            if $dic_{error}[(2d, 3d)] > k$ then
                del $(2d, 3d)$ in $set_{ori}^{(2d,3d)}$
            end if
            $error_{num}++$
        end if
    end for
    if $error_{num} > l$ then
        $dic_{error}[(2d, 3d)_{cur}]++$
    end if
    $T_{best} = T_{cur}$ with the smallest $error_{num}$
end for
$set_{end}^{(2d,3d)} = (2d, 3d)$ where $L2(pro2d(3d, T_{best}), 2d) < j$
$T_c^b = aver(PnP(grouped\ set_{end}^{(2d,3d)}$ in groups of $i))$
```

The shooting of the testing set is more complicated than the shooting of the training set. The range of motion of the manipulator is extended to a spherical surface instead of only a plane to simulate the spatial motion of the manipulator in a more realistic industrial scene, which can get perspectives that are consistent with industrial reality and consequently shoot more realistic images. For example, the automatic loading and unloading, as well as assembly and processing operations commonly found in industry use robotic arms to view the part being operated from spherical surface perspectives.

For synthetic images, we used Blender to generate images with realistic reflections. We distributed our camera viewpoints in a Fibonacci arrangement, covering all possible poses. Then, we randomly arranged three light sources on a sampling sphere and changed the intensity of the light sources randomly to simulate the reflection of metal parts. Our dataset generated 6600 synthetic images for each type of part. In addition, it is possible to replace the background of the synthetic images with a randomly selected background provided by SUN [34]. The scripts of these processes can be found on the website.

For each object, three types of CAD models are supplied. "stl" and "ply" format files are provided because they are widely used. Then, we also provide the "sldrt" format 3d modeling file, which is designed by the designer prior to machine the part. Users can generate other special formats according to research demands or reprocess the original file to conduct other studies with this type of CAD model.

3.4 Ground truth and bounding-box

To obtain accurate ground truth poses, the external camera parameters in Sect. 3.2 are not sufficient. In the entire process, only the relative poses between the parts and the calibration board are unchanged. If this relative pose can be accurately obtained and combined with the relative pose between the camera and the calibration board, it can be detected. Then, the ground truth poses of the parts can be obtained in batches. Before generating the final position and ground truth poses of each part, we first print the positioning marks of the part on the calibration board and place the part according to the positioning marks to get a more accurate relative position between the part and the board. But there may be errors because it is manually placed, so we render the poses of the parts, generate a visual image, and obtain more accurate relative poses through the manual adjustment to ensure the accuracy of the final poses. For a set of images, we only need to do the above operation once, because the relative position between the part and the calibration plate is constant during the process of taking a set of images.

After obtaining the ground truth poses, the mask of the objects can be obtained, and the masks are processed to generate the minimum circumscribed rectangle as the bounding-box annotation. Users can read annotations and conveniently load them into memory by a python script. Users can also easily use our script to convert RT-Less annotations and images to BOP format. We also provide many other scripts for RT-Less image processing and one can see more details online at: https://github.com/transcend-lzy/RT-Less-toolbox.

4 How to use

There are many types of pose estimation methods, and the data used by different methods are also different. The comparison of different methods with large differences sometimes causes unfair results. Therefore, for the use of this database, we provide some suggestions on training modes and testing modes.

4.1 Training modes

When using RT-Less for training, there are two training modes according to the number of training images.

Training modes:

1. CAD models only: real images cannot be used for training, only CAD models are provided.
2. CAD models and real images: real images can be used for training, while CAD models are also provided.

Mode 1 is the most challenging but also the most practical since no need to capture real images for training. However, Mode 1 is too difficult for the existing methods. Currently, it is reasonable to use Mode 2 to test them. Although Mode 2 is also difficult, it is expected that the performance would be greatly improved in the next few years.

4.2 Testing modes

Most of the pose estimation methods detect the target first and then compute the pose, but due to the limited technology, the detectors are all inaccurate. Therefore, to test the pose estimation methods, an effective pose estimation algorithm in the second stage should be able to deal with these inaccurate detection results. So, we provide a uniform random bounding-box that is used to simulate the random offset and scaling of inaccurate target detectors to test the robustness of the pose estimation algorithm to the target detection results. Specifically, we randomly scale the ground truth bounding box of the target object by a factor of 1.0 to 1.2, and randomly translate it up, down, left, and right by 10% of the bounding box length. We call this ‘the object detection simulation module’. For whether to use the module we provide, there are two testing modes for pose estimation.

Testing modes:

1. Use object detection simulation module to evaluate the performance of pose estimation methods.
2. Use real object detection module to comprehensively evaluate the performance of pose estimation methods.

When using RT-Less for testing, we divide the dataset into three groups according to the degree of occlusion and clutter in different scenes: (1) Few occlusions and clutters; (2) Slight occlusion and few clutters; (3) Severe occlusion and clutter. One can obtain more detailed grouping information in the dataset website. While using RT-Less, one can choose any mode and group to train and test, but different modes or groups should be noted and compared.

5 Baselines

We use two state-of-the-art pose estimation methods of Wu et al. [35] and Haugaard et al. [36] to show how to use the dataset and the results are also can be used as a baselines for further study of new methods. These two methods perform well in the mainstream datasets like Linemod and T-Less. In the examples, we tested all three groups of testing images in Testing mode 1 using Training mode 2. We attempted to use training mode 1 for testing, but the results were not meaningful because the model trained without real images performed poorly. As an alternative, we modified the number of real images used in training mode 2 to evaluate the impact of different quantities of real images on the experimental results. The results are shown in Tables 1, 2. Table 1 evaluates the performance of the pose estimation methods on each object, while Table 2 evaluates their performance in different scenes. For Table 2, we sort the table according to the difficulty of the scene we evaluated during we designed the scenes.

5.1 Evaluation metric

We evaluate baselines from two perspectives: robustness and accuracy. Robustness refers to the probability of the successful pose estimation. Specifically, a rough accuracy threshold is set initially, when the error of the pose estimation is less than this threshold, the estimation is considered to be successful. After many times of experiments are completed, probability of the successful estimation can be calculated, which is the robustness of the pose estimation.

On the other hand, accuracy is also a quantitative metric. In a successful estimation, its 6D pose estimation error can be precisely calculated, giving a quantitative evaluation of its accuracy. We take the mean of the errors over multiple successes as the accuracy metric. It should be noted that the evaluation of the accuracy is carried out under the condition of successful estimation, otherwise the error is too large, and it is meaningless to calculate the accuracy.

In the existing researches on pose estimation, most of them only evaluated the robustness, because the existing researches are mainly aimed at non-industrial scenarios, and the requirements for accuracy are relatively low. For industrial parts, the pose estimation requires further accuracy evaluation. Consequently, our evaluations encompass both these dimensions.

5.1.1 Robustness evaluation

We evaluate the robustness of the methods in terms of the ADDS [12, 14], which is defined as the average distance of

Table 1 Pose estimation results of [35] and [36] according to different objects

Obj_id	Diameter (mm)	Symmetry	Train mode 2 (340 real images)						Train mode 2 (200 real images)					
			SurfEmb [36]			PSGMN [35]			SurfEmb [36]			PSGMN [35]		
			ADDS	R (°)	T (mm)	ADDS	R (°)	T (mm)	ADDS	R (°)	T (mm)	ADDS	R (°)	T (mm)
1	122.98	Yes	0.36	1.27	5.76	0.71	1.79	10.72	0.28	0.72	6.4	0.69	4.97	11.25
2	115.87	Yes	0.6	1.91	5.62	0.61	3.85	11.97	0.38	1.76	6.05	0.65	4.58	11.84
3	134.16	No	0.88	1.43	4.98	0.41	3.61	9.27	0.86	1.48	5.53	0.33	3.78	9.95
4	116.22	No	0.41	0.74	6.47	0.24	1.82	6.09	0.45	0.76	4.98	0.18	1.54	6.61
5	130.86	Yes	0.55	0.84	6.17	0.67	2.91	10.13	0.33	0.76	6.36	0.59	3.12	11.34
6	90.86	Yes	0.47	2.28	4.35	0.26	3.3	6.9	0.58	1.82	4.35	0.25	3.63	6.46
7	143.15	No	0.3	1.52	7.86	0.48	2.48	7.83	0.3	0.87	8.27	0.27	2.44	10.36
8	113.06	No	0.19	1.69	6.41	0.24	2.24	6.85	0.12	3.99	8.85	0.2	2.38	6.3
9	99.62	No	0.29	3.47	6.5	0.13	4.34	7.51	0.22	3.82	6.77	0.05	6.28	8.75
10	88.87	No	0.13	5.61	7.18	0.1	3.9	6.95	0.14	5.69	7.33	0.04	7.35	10.66
11	98.31	No	0.23	2.6	5.86	0.39	2.27	5.66	0.15	2.54	6.17	0.25	2.56	5.91
12	98.71	No	0.3	1.49	5.1	0.11	2.53	5.96	0.26	1.29	6.02	0.14	1.77	6.06
13	117.75	No	0.36	4.43	8.81	0.21	3.23	7.29	0.23	4.65	8.56	0.26	3.01	7.13
14	117.05	Yes	0.44	1.53	8.76	0.67	2.98	11.56	0.31	4.03	9.13	0.42	3.16	16.61
15	113.58	No	0.34	6.24	11.94	0.17	2.62	7.37	0.34	7.51	10.3	0.09	2.78	8.08
16	187.03	No	0.5	2.05	10.49	0.46	2.33	8.23	0.36	3.1	11.49	0.37	2.14	8.19
17	116.69	Yes	0.23	3.78	7.33	0.18	3.36	11.53	0.16	5.57	10.38	0.1	6.24	11.83
18	112.25	Yes	0.31	7.15	10.57	0.36	3.43	11.98	0.23	8.47	10.22	0.23	3.39	12.79
19	111.11	No	0.27	8.3	8.77	0.16	2.44	8.23	0.26	7.52	9.41	0.1	5.69	11.49
20	94.20	No	0.25	5.5	9.1	0.12	2.41	6.91	0.21	5.62	9.67	0.11	3.81	9.19
21	123.05	No	0.24	4.28	7.31	0.21	3.65	7.29	0.12	4.34	7.79	0.12	3.98	10.52
22	114.05	No	0.45	0.69	6.42	0.17	1.77	7.05	0.41	0.65	7.26	0.08	5.19	8.08
23	108.12	No	0.26	2.64	7.74	0.2	2.38	6.19	0.27	4.47	11.92	0.15	2.98	7.15
24	112.25	Yes	0.2	6.4	7.07	0.24	6.79	10.55	0.19	6.61	7.21	0.12	6.95	11.47
25	79.59	No	0.22	2.2	4.5	0.13	2.35	5.01	0.2	1.29	3.98	0.1	5.24	8.41
26	117.19	Yes	0.36	1.83	6.01	0.61	2.25	10.04	0.36	1.42	5.03	0.51	2.78	11.98
27	102.84	No	0.27	3.64	5.19	0.29	2.39	6.41	0.23	4.98	8.45	0.18	2.51	6.17
28	115.62	No	0.26	1.45	5.84	0.03	3.09	7.09	0.28	1.28	7.02	0.03	4.94	8.35
29	100.49	Yes	0.24	1.54	5.86	0.65	2.67	8.45	0.35	1.2	5.48	0.7	3.25	8.44
30	81.96	No	0.19	4.34	6.8	0.13	5.52	7.3	0.24	3.93	5.99	0.13	7.31	9.55
31	99.56	No	0.3	3.6	7.43	0.16	3.9	9.22	0.24	3.24	7.75	0.09	4.61	9.9
32	113.98	No	0.05	2.81	7.06	0.28	2.48	5.54	0.02	3.23	5.66	0.13	3.1	6.97
33	110.63	Yes	0.35	4.34	7.26	0.52	2.17	9.12	0.41	4.25	7.83	0.33	2.95	11.1
34	103.89	No	0.29	6.93	14.06	0.25	2.67	6.25	0.39	4.29	11.48	0.23	2.27	5.79
35	170.91	No	0.37	0.55	8.1	0.16	2.8	8.97	0.29	0.55	8.93	0.05	2.64	9.09
36	140.19	No	0.34	2.11	7.98	0.25	3.75	13.85	0.35	1.59	7.94	0.23	4.56	12.45
37	126.20	No	0.32	1.23	7.29	0.14	5.2	9.54	0.43	1.58	6.06	0.11	5.78	10.44
38	86.39	No	0.17	1.82	4.8	0.17	2.63	6.86	0.26	0.89	5.24	0.13	3.31	7.04
Average			0.32	3.06	7.23	0.3	3.06	8.25	0.3	3.2	7.56	0.23	3.92	9.31

Table 2 Results of [35] and [36] according to different scenes

Scene	Difficulty level	Train mode 2 (340 real images)						Train mode 2 (200 real images)					
		SurfEmb [36]			PSGMN [35]			SurfEmb [36]			PSGMN [35]		
		ADDS	R/°	T/mm	ADDS	R/°	T/mm	ADDS	R/°	T/mm	ADDS	R/°	T/mm
7	1	0.65	0.87	5.21	0.49	2.48	7.9	0.57	1	4.06	0.49	2.48	7.9
13	1	0.64	1.33	4.37	0.66	2.94	10.26	0.55	1.36	4.6	0.65	3.39	9.45
31	1	0.34	1.69	5.87	0.36	2.65	6.95	0.37	1.19	6.63	0.3	2.75	7.81
32	1	0.33	1.98	6.62	0.39	2.72	7.87	0.36	1.85	6.55	0.33	2.92	6.98
20	1	0.46	0.99	4.61	0.28	2.02	6.35	0.57	0.8	4.45	0.17	1.98	5.75
3	1	0.3	2.73	5.88	0.39	2.48	7.8	0.25	1.81	6.67	0.24	2.75	7.16
24	1	0.46	1.28	6.25	0.35	3.08	7.98	0.39	1	7.02	0.37	6.76	12.69
1	1	0.22	1.53	5.58	0.25	2.62	6.66	0.26	1.27	5.47	0.18	3.23	8.31
11	1	0.28	3.66	8.38	0.3	2.34	10.33	0.28	5.83	9.43	0.18	2.97	13.33
16	1	0.55	0.95	5.44	0.37	2.48	7.38	0.6	0.82	4.6	0.26	3	8.07
5	1	0.31	1.87	6.28	0.27	3.28	9.13	0.15	2.22	6.66	0.18	6.4	14.67
28	1	0.29	2.89	4.98	0.25	2.65	7.72	0.28	3.45	7.48	0.23	2.51	7.32
29	2	0.23	5.15	10.39	0.22	3.6	7.45	0.26	2.9	6.03	0.16	3.97	8.39
15	2	0.38	4.17	8.19	0.35	2.48	10.51	0.36	3.81	9.29	0.1	2.91	11.57
17	2	0.4	5.87	8.08	0.38	2.58	8.56	0.36	7.71	9.81	0.3	2.8	9.57
12	2	0.21	3.59	7.96	0.24	3.29	9.88	0.15	5.39	10.17	0.16	5.24	13.82
6	2	0.21	1.29	5.89	0.16	3.63	7.73	0.15	1.11	6.12	0.13	2.37	6.71
22	2	0.28	1.59	5.63	0.24	4.81	9.52	0.3	1.19	4.87	0.22	4.96	10.24
26	2	0.28	4.58	10.87	0.26	2.28	7.31	0.27	8.41	15.74	0.16	2.84	8.46
27	2	0.2	2.09	5.64	0.15	4.64	8.45	0.26	1.68	5.69	0.11	10.6	7.76
4	2	0.39	2.92	7.82	0.18	4.89	7.81	0.29	2.79	8.21	0.16	5.09	7.65
2	2	0.29	4.55	7.46	0.23	2.81	9.74	0.23	5.64	9.6	0.1	5.54	13.22
10	2	0.44	0.93	5.68	0.33	2.85	8.59	0.34	1.01	6.35	0.3	3.56	9.08
9	2	0.27	1.24	6.01	0.29	2.72	8.75	0.29	1.19	5.94	0.19	5.34	11.14
21	3	0.41	2.05	6.55	0.52	3.62	11.02	0.24	4.58	9.76	0.49	3.38	11.86
19	3	0.26	1.45	5.55	0.29	2.3	5.98	0.18	3.39	5.74	0.2	2.32	6.44
18	3	0.34	1.63	4.86	0.24	2.67	9.33	0.23	1.75	5.91	0.23	5.91	10.52
8	3	0.25	7.24	8.24	0.19	3.42	11.09	0.25	7.77	7.99	0.14	7.38	12.55
23	3	0.35	1.54	4.94	0.15	2.63	8.12	0.3	1.39	5.01	0.13	3.16	8.86
14	3	0.14	5.23	7.55	0.14	5.66	9.01	0.14	5.23	7.55	0.1	6.55	10.71
30	3	0.17	1.71	5.69	0.25	2.9	7.31	0.11	1.65	6.53	0.21	5.8	10.06
25	3	0.39	6.12	10.53	0.36	3.48	9.77	0.33	7.94	14.52	0.31	3.35	9.37
Average	–	0.34	2.71	6.66	0.3	3.09	8.51	0.3	3.1	7.33	0.23	4.19	9.61

the pairwise distances between the 3D model points transformed with the ground truth and estimated poses:

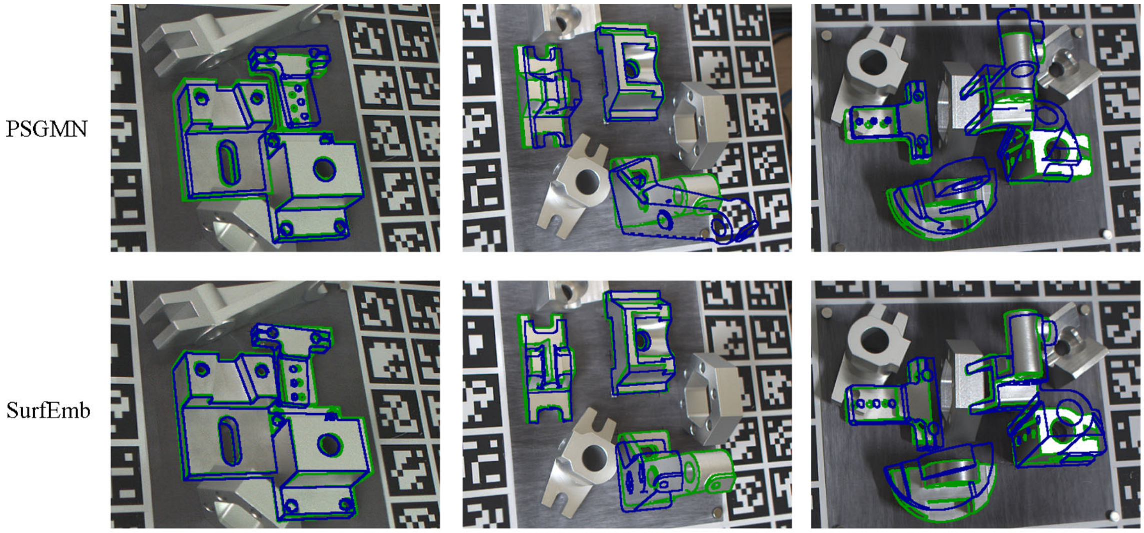
$$ADD = \frac{1}{m}\sum_{x\in M}\left\|(Rx+T)-(\tilde{R}x+\tilde{T})\right\|, \qquad (1)$$

where M is the set of 3D model points, m is the number of points, and (R, T) $(\tilde{R}, \tilde{T})$. are the rotation and translation of the ground truth pose and predicted pose.

For symmetric objects, the average distance is computed using the closest point distance because of the ambiguous situation for some views:

$$ADD-S = \frac{1}{\mathrm{m}}\sum_{x_1\in M}\min_{x_2\in M}\left\|(Rx_1+T)-(\tilde{R}x_2+\tilde{T})\right\|, \qquad (2)$$

Pose estimation is considered to fail if the error is larger than the threshold, which is 10% of the largest distance in



(a) level 1

(b) level 2

(c) level 3

Fig. 8 Performance of SurfEmb and PSGMN in scenes with different difficulty levels. The green box is the ground truth and the blue box is the predicted result. **a** Difficulty level 1. **b** Difficulty level 2. **c** Difficulty level 3

Table 3 Performance of [35] and [36] evaluated with ADDS metric on other datasets

Dataset	SurfEmb [36]	PSGMN [35]
Linemod	–	0.93
Linemod-Occluded	0.66	0.65
T-LESS	0.74	–
YCB-Video	0.79	–

the set of CAD model points. Then, we see the percentage of success as the ADDS score.

5.1.2 Precision evaluation

In addition, considering that the ADDS are relatively rough and cannot meet the requirements in the field of precision manufacturing, we therefore use the mean error of rotation R_{mean} and translation T_{mean} between estimation and ground truth in the case of successful pose estimation to evaluate the precision of the method:

$$R_{\text{mean}} = \frac{1}{n}\sum_{i \in S} \text{avg}\big(\text{abs}(\alpha_i - \alpha_i') + \text{abs}(\beta_i - \beta_i') + \text{abs}(\gamma_i - \gamma_i')\big), \tag{3}$$

$$T_{\text{mean}} = \frac{1}{n}\sum_{i \in S} \text{Euclidean}((x_i, y_i, z_i), (x_i\prime, y_i\prime, z_i\prime)), \tag{4}$$

where S is the set of images judged to be correct by ADDS, and n is the number of images in S. (α, β, γ) and $(\alpha\prime, \beta\prime, \gamma\prime)$ represent the angle values of ground truth and predicted pose, respectively.(x_i, y_i, z_i) and $(x_i\prime, y_i\prime, z_i\prime)$ represent the translation values of ground truth and predicted pose, respectively, Euclidean$(\cdot)$ indicate the calculation of Euclidean distance.

5.2 Analysis

With regard to the result, although these methods performed well on other dataset, it does not achieve comparable marvelous results on the proposed dataset.

5.2.1 Results of estimation robustness

Robustness is represented by the ADDS, Table 1 shows the ADDS of each object. Smaller and more complex parts may have lower scores. For example, the obj30 with a low ADD score is only 81 mm and has a more complex structure. Table 2 shows the ADDS of each scene. If more complex and smaller parts are included in the scene, the effect worsens. For example, although the scene28 is less occluded and not cluttered, the parts contained in the scene are relatively small and complex, so the detection results are poor. Simultaneously, the result of different scenes is related to the level of illumination and clutter. For example, scene4 and scene8 has strong occlusion (especially obj24 and obj17 in it) get bad results. And scene14 with both strong occlusions, containing small and complex parts and similar parts gave the worst result (Average result of two methods in two modes). Figure 8 presents some qualitative results, where we provide visual representations of the SurfEmb and PSGMN methods at different difficulty levels. As expected, scenes with greater difficulty levels pose more challenges for both methods.

5.2.2 Results of estimation precision

Tables 1 and 2 also show the translation and rotation errors. For these results, we can see that although the ADD-S metric of some parts are high, the average rotation and translation errors are still large (such as obj7). Such precision is unacceptable in industrial scenarios. Therefore, if the precision requirements are high in some scenarios, the ADD-S metric is not suitable for judging the performance of the method.

Few scholars have been involved in the challenging field of pose estimation of reflective texture-less parts in the past. According to the experimental results, the average ADDS of each scene is less than 35%, and the average precision can only reach the level of 7 mm and 3°. As shown in Table 3, these methods can achieve good results on other datasets, but perform poorly on RTLess, which demonstrates that the database is very challenging.

6 Conclusion and future work

This paper introduces RT-Less, which is a multi-scene RGB dataset for reflective texture-less metal objects. The dataset is proposed to facilitate research on 6D pose estimation of reflective texture-less objects. Different annotations are provided so that it can be used to research pose estimation, object detection and instance segmentation. The dataset containing 289 K true and synthetic images and test scenarios are carefully designed and graded in difficulty. And, actual industrial situation is simulated in this dataset with the setting of the parts and the test scenarios. We also provide initial baselines for RT-Less and will continuously expand the baseline methods and results online at: http://www.zju-rtl.cn/RT-Less/.

Only two methods were tested on the dataset due to the time limitation, demonstrated that it is challenging for these respective methods. Next we will conduct more methods on the dataset, especially these proposed recently and showing great potential for reflective texture-less objects [37–39], and report the results on the RT-Less dataset website. At the same time, the database will be further improved and an upgraded version of it will be launched.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grants 52275514 and 52275547, and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY21E050021.

Funding National Natural Science Foundation of China, 52275514, Zaixing He, 52275547, Xinyue Zhao, Natural Science Foundation of Zhejiang Province, LY21E050021, Xinyue Zhao.

Data availability The data and necessary code is available at http://www.zju-rtl.cn/RT-Less/.

Declarations

Conflicts of interest Xinyue Zhao, Quanzhi Li, Yue Chao, Quanyou Wang, Zaixing He, and Dong Liang declare that they have no conflict of interest or financial conflicts to disclose. The authors declares that they have no conflict of interest.

References

1. Zabulis, X., Lourakis, M.I., Koutlemanis, P.: Correspondence-free pose estimation for 3D objects from noisy depth data. Vis. Comput. **34**, 193–211 (2018)
2. Liu, D., Chen, L.: SECPNet—secondary encoding network for estimating camera parameters. Vis. Comput. **5**, 1–14 (2022)
3. Li, S., Xian, Y., Wu, W., Zhang, T., Li, B.: Parameter-adaptive multi-frame joint pose optimization method. Vis. Comput. **5**, 1–13 (2022)
4. Liang, D., et al.: Anchor retouching via model interaction for robust object detection in aerial images. IEEE Trans. Geosci. Remote Sens. **60**, 1–13 (2021)
5. Muoz, E., Konishi, Y., Murino, V., Del Bue, A.: Fast 6D pose estimation for texture-less objects from a single RGB image. In: 2016 IEEE International Conference on Robotics and Automation (icra), pp. 5623–5630 (2016)
6. Crivellaro, A., Rad, M., Verdie, Y., Yi, K.M., Fua, P., Lepetit, V.: A novel representation of parts for accurate 3D object detection and tracking in monocular images. In: 2015 IEEE International Conference on Computer Vision (ICCV), New York: IEEE, pp. 4391–4399. https://doi.org/10.1109/ICCV.2015.499
7. Wei, Z., et al.: learning calibrated-guidance for object detection in aerial images. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. **15**, 2721–2733 (2022). https://doi.org/10.1109/JSTARS.2022.3158903
8. Liang, D., Liu, X.: Coarse-to-fine Foreground Segmentation based on Co-occurrence Pixel-Block and Spatio-Temporal Attention Model. In: 2020 25th International Conference on Pattern Recognition (ICPR), Los Alamitos: IEEE Computer Soc, pp. 3807–3813. https://doi.org/10.1109/ICPR48806.2021.9412814
9. Sun, H., Chen, X., Wang, L., Liang, D., Liu, N., Zhou, H.: C(2)DAN: an improved deep adaptation network with domain confusion and classifier adaptation. Sensors (2020). https://doi.org/10.3390/s20123606
10. Rennie, C., Shome, R., Bekris, K.E., De Souza, A.F.: A dataset for improved RGBD-based object detection and pose estimation for warehouse pick-and-place. IEEE Robot. Autom. Lett. (2016). https://doi.org/10.1109/LRA.2016.2532924
11. Eppner, C., et al.: Lessons from the amazon picking challenge: four aspects of building robotic systems. Robot. Sci. Syst. **5**, 96 (2016). https://doi.org/10.15607/RSS.2016.XII.036
12. Hinterstoisser, S. et al.: Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In: Asian conference on computer vision, Springer, pp. 548–562 (2012)
13. Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J., Rother, C.: Learning 6D object pose estimation using 3D object coordinates. Comput. Vis. **8690**, 536–551 (2014). https://doi.org/10.1007/978-3-319-10605-2_35
14. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: a convolutional neural network for 6D object pose estimation in cluttered scenes
15. Calli, B., Srinivasa, S., Singh, A., Abbeel, P., Walsman, A., Dollar, A.M.: The YCB object and model set: towards common benchmarks for manipulation research. In: Proceedings of

the 17th International Conference on Advanced Robotics (icar), pp. 510–517 (2015), https://doi.org/10.1109/ICAR.2015.7251504

16. Richter-Kluge, J., Wellhausen, C., Frese, U.: ESKO6d-a binocular and RGB-D dataset of stored kitchen objects with 6D poses. IEEE/RSJ Int. Conf. Intell. Robot. Syst. **2019**, 893–899 (2019)
17. Kaskman, R., Zakharov, S., Shugurov, I., Ilic, S.: HomebrewedDB: RGB-D dataset for 6D pose estimation of 3D objects. IEEE/Cvf Int. Conf. Comput. Vis. Workshops **2019**, 2767–2776 (2019). https://doi.org/10.1109/ICCVW.2019.00338
18. Tejani, A., Kouskouridas, R., Doumanoglou, A., Tang, D., Kim, T.-K.: Latent-class hough forests for 6 DoF object pose estimation. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 1 (2018), https://doi.org/10.1109/TPAMI.2017.2665623
19. Tremblay, J., To, T., Birchfield, S.: Falling things: a synthetic dataset for 3D object detection and pose estimation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2119–21193 (2018). https://doi.org/10.1109/CVPRW.2018.00275
20. Hodan, T. et al.: Bop: Benchmark for 6D object pose estimation. In: Proceedings of the European conference on computer vision (ECCV), pp. 19–34 (2018)
21. Yuan, H., Hoogenkamp, T., Veltkamp, R.C.: RobotP: a benchmark dataset for 6D object pose estimation. Sensors **21**(4), 5 (2021). https://doi.org/10.3390/s21041299
22. Yang, J., Gao, Y., Li, D., Waslander, S.L.: ROBI: a multi-view dataset for reflective objects in robotic bin-picking. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic: IEEE, pp. 9788–9795 (2021). https://doi.org/10.1109/IROS51168.2021.9635871
23. Yang, J., Li, D., Waslander, S.L.: Probabilistic multi-view fusion of active stereo depth maps for robotic bin-picking. In: IEEE Robotics and Automation Letters, vol. 6, no. 3 (2021). https://doi.org/10.1109/LRA.2021.3068706
24. Drost, B., Ulrich, M., Bergmann, P., Haertinger, P., Steger, C.: Introducing MVTec ITODD-a dataset for 3D object recognition in industry. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW 2017), pp. 2200–2208 (2017). https://doi.org/10.1109/ICCVW.2017.257
25. Hodan, T., Haluza, P., Obdrzalek, S., Matas, J., Lourakis, M., Zabulis, X.: T-LESS: an RGB-D dataset for 6D pose estimation of texture-less objects. In: 2017 Ieee Winter Conference on Applications of Computer Vision (wacv 2017), pp. 880–888 (2017). https://doi.org/10.1109/WACV.2017.103
26. Bregier, R., Devernay, F., Leyrit, L., Crowley, J.L.: Symmetry aware evaluation of 3D object detection and pose estimation in scenes of many parts in bulk. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice: IEEE, pp. 2209–2218 (2017). https://doi.org/10.1109/ICCVW.2017.258
27. Kleeberger, K., Landgraf, C., Huber, M.F.: Large-scale 6D object pose estimation dataset for industrial bin-picking. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2573–2578 (2019). https://doi.org/10.1109/IROS40897.2019.8967594
28. Doumanoglou, A., Kouskouridas, R., Malassiotis, S., Kim, T.-K.: Recovering 6D object pose and predicting next-best-view in the crowd. IEEE Conf. Comput. Vis. Pattern Recogn. **2016**, 3583–3592 (2016). https://doi.org/10.1109/CVPR.2016.390
29. Li, X., et al.: A sim-to-real object recognition and localization framework for industrial robotic bin picking. IEEE Robot. Autom. Lett. **7**(2), 3961–3968 (2022). https://doi.org/10.1109/LRA.2022.3149026
30. Dwibedi, D., Misra, I., Hebert, M.: Cut, paste and learn: Surprisingly easy synthesis for instance detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1301–1310 (2017)
31. Feng, Y. et al.: Towards robust part-aware instance segmentation for industrial bin picking. In: 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA: IEEE, pp. 405–411 (2022). https://doi.org/10.1109/ICRA46639.2022.9811728
32. Zhang, Z.Y.: A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Mach. Intell. (2000). https://doi.org/10.1109/34.888718
33. Garrido-Jurado, S., Munoz-Salinas, R., Madrid-Cuevas, F.J., Marin-Jimenez, M.J.: Automatic generation and detection of highly reliable fiducial markers under occlusion. Pattern Recognit. (2014). https://doi.org/10.1016/j.patcog.2014.01.005
34. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In:2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, pp. 3485–3492 (2010)
35. Wu, C., Chen, L., He, Z., Jiang, J.: Pseudo-siamese graph matching network for textureless objects'6-D pose estimation. IEEE Trans. Industr. Electron. **69**(3), 2718–2727 (2021)
36. Haugaard, R.L., Buch, A.G.: SurfEmb: dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6749–6758 (2022)
37. He, Z., Li, Q., Zhao, X., Wang, J., Shen, H., Zhang, S., Tan, J.: ContourPose: monocular 6D pose estimation method for reflective texture-less metal parts. IEEE Trans. Robot. **5**, 69 (2023)
38. He, Z., Chao, Y., Wu, M., Hu, Y., Zhao, X.: G-GOP: generative pose estimation of reflective texture-less metal parts with global-observation-point priors. IEEE-ASME Trans. Mechatron. **5**, 96 (2023)
39. He, Z., Wu, M., Zhao, X., Zhang, S., Tan, J.: A generative feature-to-image robotic vision framework for 6-D pose estimation of metal parts. IEEE-ASME Trans. Mechatron. **27**(5), 3198–3209 (2022)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Xinyue Zhao received her M.S. degree in Mechanical Engineering from Zhejiang University, China in 2008, and her Ph.D. degree in Graduate School of Information Science and Technology from Hokkaido University, Japan in 2012. She is currently an associate professor in the School of Mechanical Engineering, Zhejiang University, China. Her research interests include machine vision and image processing. She has published nearly 50 peer reviewed journal papers.

Quanzhi Li received the B.S. degree in mechanical and electronic engineering from Chongqing University, Chongqing, China, in 2021. He is currently pursuing the M.S. degree in mechanical engineering with Zhejiang University, Hangzhou, China. His research interests include Robotic Vision, Pose Estimation and Artificial Intelligence.

Yue Chao received the B.S. degree in mechanical engineering from Shandong University, Shandong, China, in 2020. He is currently pursuing the M.S. degree in mechanical engineering with Zhejiang University, Hangzhou, China. His research interests include structured robotic vision, deep learning.

Quanyou Wang received his B.S. degree in Mechanical Engineering at Zhejiang University, Zhejiang, China, in 2022. He is currently pursuing his Ph.D. degree in Mechanical Engineering at the University of California, Los Angeles. His research interests include Mechanical Design and Locomotion on Humanoid Robots.

Zaixing He received his B.Sc. and M.Sc. degrees in Mechanical Engineering from Zhejiang University, China in 2006 and 2008, respectively. He received his Ph.D. degree in 2012 from the Graduate School of Information Science and Technology, Hokkaido University, Japan. He is currently an associate professor in the School of Mechanical Engineering, Zhejiang University. His research interests include robotic vision, Visual intelligence of manufacturing equipment, and optical-based measurement. He has published over 40 peer reviewed papers in prestigious journals such as IEEE TRO, TIE, TII, TIM, IEEE/ASME TMech, Pattern Recognition, Neurocomputing. He served as Lead Guest Editor or Guest Editor of several journals including IEEE TCE and Mathematics, Program Chair or TPC of more than 10 international conferences. He is a senior member of IEEE.

Dong Liang received Ph.D. at Hokkaido University, Japan in 2015. He received the B.S. degree from Lanzhou University, China, in 2008 and 2011, respectively. He is currently an Associate Professor in the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. His research interests include machine learning and pattern recognition, intelligent interaction systems, and computational imaging systems. He has published several research papers including ICCV, IJCAI, AAAI, Pattern Recognition and IEEE TIP/TMM/TNNLS/TGRS/TCSVT.