ROV6D: 6D Pose Estimation Benchmark Dataset for Underwater Remotely Operated Vehicles

Jingyi Tang[1,2], Zeyu Chen[1], Bowen Fu[3], Wenjie Lu[4], Shengquan Li[2], Xiu Li[1,2] and Xiangyang Ji[3]

***Abstract*—Accurately localization between multi-robots is crucial for many underwater applications, such as tracking, convoying and subsea intervention tasks. 6D pose estimation is a fundamental task that enables precise object localization in 3D space with full six degrees of freedom. However, one critical challenge is the lack of available large-scale datasets due to the unbearable cost of labelled data collection. To overcome this difficulty, we propose a benchmark dataset, ROV6D, for 6D pose estimation of remotely operated vehicles (ROVs). The training subset consists of a large number of synthetic images with 6D pose ground truth for ROVs. These synthetic images are generated using BlenderProc and further rendered with the underwater neural rendering (UWNR) strategy to enhance their realism. The testing subsets cover different real-world scenarios, including the Pool subset and Maoming subset, focusing on challenging cases that involve partial occlusion and low visibility. Diverse recent methods are evaluated on the constructed dataset. The results show that methods based on dense coordinates currently perform best, outperforming both the keypoint-based method and the refinement-based method. Our dataset will be made publicly available soon.**

***Index Terms*—6D Pose Estimation Dataset, Underwater Remotely Operated Vehicles, Underwater Neural Rendering, Motion Capture System, Visual Perception.**

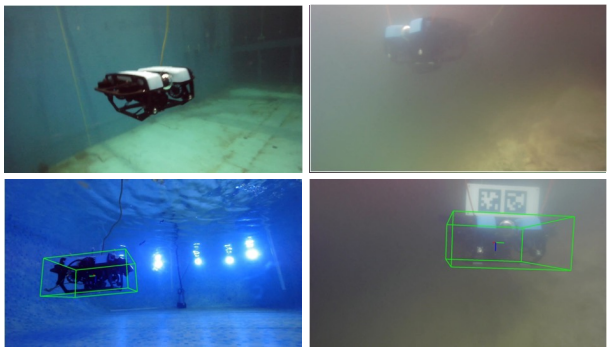

Fig. 1. The BlueROV Heavy is commonly utilized underwater, such as conducting experiments in engineering pools (left) and exploring the open lakes (right). We can provide the 6D pose annotations obtained through a motion capture system or a marker-based system (bottom). The green wireframe represents the projection of the 3D bounding box of the ground-truth pose.

I. Introduction

ESTIMATING the 6D pose, i.e., the 3D rotation and 3D translation, is essential for underwater localization, navigation and cooperative operations between unmanned underwater vehicles (UUVs), such as remotely operated vehicles (ROVs) and autonomous underwater vehicles (AUVs). The vast majority of previous work about multi-robot relative localization focus on terrestrial autonomous driving, on-orbit servicing and unmanned aerial vehicle [1]–[5]. Although researches on 6D pose estimation are proliferating in recent years, some important environments have not received significant attention thus far, such as underwater.

Traditionally, the performance of underwater acoustic localization is limited by its long latency and poor accuracy [6], [7]. In contrast, vision-based methods offer advantages such as fine discrimination and real-time localization at a near distance. However, camera-based underwater localization encounters great practical difficulties due to lighting variations, hazing, and color loss [8]–[10]. Typically, the quality of underwater images is adversely affected by the absorption and scattering [11]–[13]. Particulates, such as suspended sand and silt, further impair visibility underwater. Additionally, the presence of dynamic obstacles introduces sudden and unpredictable disturbances. All these factors mentioned above significantly impact the performance of pose estimation.

Traditional work has utilized fiducial markers on the target to simplify the visual detection task, while these markers may become invisible due to variations in the target's pose [14]. Other methods have established the correspondence between an object image and its 3D model [15], [16]. Nevertheless, these methods rely on rich textures to detect features for matching, which could fail in texture-less cases. Benefitting from the rise of deep learning methods, recent data-driven works train networks to take an image as input and predict its corresponding pose [17], [18]. Despite the prolific work, both the comprehensive study and insightful analysis of underwater

Manuscript received May 31, 2023; Revised August 27, 2023; Accepted October 22, 2023. This paper was recommended for publication by Editor C. Cadena Lerma upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by National Natural Science Foundation of China (Grant No. 62027826), in part by Shenzhen Science and Technology Project (Grant No. JCYJ20200109143041798), in part by Shenzhen Stable Supporting Program (WDZC20200820200655001), and in part by Shenzhen Key Laboratory of next generation interactive media innovative technology (Grant No. ZDSYS 20210623092001004). *(Corresponding authors: Xiu Li, Xiangyang Ji.)*

[1,2]Jingyi Tang and Xiu Li are with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: tangjy21@mails.tsinghua.edu.cn; li.xiu@sz.tsinghua.edu.cn).

[1]Zeyu Chen is with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China (e-mail: chenzy22@mails.tsinghua.edu.cn).

[3]Bowen Fu and Xiangyang Ji are with the Department of Automation and BNRist, Tsinghua University, Beijing, China (e-mail: fbw19@mails.tsinghua.edu.cn; xyji@tsinghua.edu.cn).

[4]Wenjie Lu is with the Harbin Institute of Technology Shenzhen, Shenzhen 518055, China (e-mail: luwenjie@hit.edu.cn).

[2]Shengquan Li is with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: lishq@pcl.ac.cn).

Digital Object Identifier (DOI): see top of this page.

multi-robot systems remain unsatisfactory due to the lack of a publicly available real-world underwater vehicle pose dataset. In particular, it is costly and time-consuming to photograph objects in a real underwater scene and obtain the corresponding ground truth for different water types.

In this work, we construct a large-scale underwater dataset with 6D pose annotations, focusing on estimating the 6D pose of a BlueROV Heavy, as shown in Fig. 1. For training, a large number of synthetic images with 6D pose ground truth are generated using BlenderProc [19]. To enhance the realism of the synthetic images and bridge the domain gap with real-world images, we employ the underwater neural rendering (UWNR) strategy [20] across various underwater scenarios. For testing, real-world data are collected both in a pool and the open lake at Maoming, which contain diverse and challenging scenes, such as heavy occlusions, low visibility and viewpoints ambiguity. We provide high-quality pose annotations obtained through a motion capture system and a vision-based Apriltag system for accurate quantitative evaluation. Due to the unavailability of satisfactory annotations, a part of the images in the Maoming subset is treated as unlabelled challenging data. Additionally, we evaluate the performance of three types of methods on our proposed dataset: the keypoint-based method DeepURL [17], the refinement-based method CosyPose [21] and the dense coordinates-based methods CDPN [22] and GDR-Net [23]. To the best of our knowledge, there are no relevant benchmarks specifically designed for pose estimation of underwater ROVs.

The main contributions are summarized as follows.

- **Underwater 6D pose estimation dataset** in a unified format is constructed for the remotely operated vehicles, BlueROV. The dataset contains i) **Synthetic subset:** 34K synthetic RGB images rendered from different viewpoints. ii) **Pool and Maoming subsets:** 8.7K real underwater RGB images collected in both a pool and the open lake at Maoming focusing on challenging cases that involve partial occlusion and low visibility. This dataset will soon be provided to the public for continued use in research.
- **High-quality underwater 6D pose annotations** are provided by a motion capture system and a vision-based Apriltag system for our real-world Pool subset and 1071 images of Maoming subset.
- **A comprehensive evaluation** of three types of methods on the benchmark dataset is conducted. We analyse the results and identify the open problems.

II. Related Works

The presented work relates to two major strands of research: datasets about underwater robotic relative localization and 6D pose estimation.

A. Datasets

The availability of data plays a crucial role in advancing research in underwater robotic systems, enabling method evaluation and a deeper understanding of their limitations. With the aid of motion capture systems or other vision-based marker systems, several datasets have emerged in the field of underwater 6D pose estimation. Nielsen et al. [24] collect two datasets for estimating the 6D pose between a BlueROV and a submarine connector, with pose annotations obtained by a motion capture system. Billings et al. [1] construct UWHandles, a fisheye image dataset, to explore ROI-based 6D object pose estimation methods. However, these works have mainly focused on stationary targets in clear water, limiting their applicability to scenarios with low visibility and highly dynamic tracking. In terms of relative localization between underwater robots, Koreitem et al. [25] collect the Barbados 2017 Dataset, including 188 real images of the Aqua robot. The robot's 6D poses are obtained by a custom-built annotator. Joshi et al. [17] publish a dataset of the Aqua2 robot in the ocean and swimming pool, specifically designed for AUV tracking and convoying. Nonetheless, they do not consider situations with limited visibility, such as occlusion and turbid conditions.

On the other hand, the generalization capability of data-driven deep learning methods is limited by the difficulties in collecting real-world data, as well as the large variability across different underwater regions. To tackle this issue, synthetic training data has emerged as an effective alternative [17], [26]. Previous approaches, such as placing multi-scale object images on an underwater background for data augmentation [27], often overlook the domain gap between real-world and synthetic data. To bridge this gap, researchers explore model-based and GAN-based methods to generate realistic underwater images. For example, DeepURL [17] employs CycleGAN [28] to generate synthetic training data. However, these GAN-based approaches may suffer from unstable training and non-convergence.

B. 6D Pose Estimation

Since infrared light decays rapidly, RGBD-based methods face limitations in underwater applications. Thus, our focus is on estimating 6D pose from a single RGB input. The most popular approach is the keypoint-based deep network that detects 2D keypoints from RGB images, followed by a PnP solver to predict the 6D poses [17], [29], [30]. In the underwater domain, DeepURL [17] detects eight keypoints to predict the 6D pose of Aqua2 robots. However, its robustness and accuracy are limited since the sparse 2D-3D correspondences make the network sensitive to partial occlusion. To address this challenge, dense correspondences-based methods have been investigated, demonstrating their robustness in the presence of heavy occlusions [22], [23]. However, most of these indirect methods cannot be trained end-to-end and are time-consuming during inference. Hence, some methods directly regress the 6D pose from input images. Although recent research [31], [32] performs well on standard indoor pose estimation benchmarks, e.g., LINEMOD [33], Occlusion LINEMOD [34], and TLESS [35], they are not robust to natural scenes [36]. Due to the limited datasets for 6D pose estimation in the marine context, it is hard to develop underwater 6D pose estimation methods.

Existing underwater robotic 6D pose estimation datasets [17], [25] mainly focus on AUVs and utilize a vision-based marker or a manual annotator for annotation. Compared with

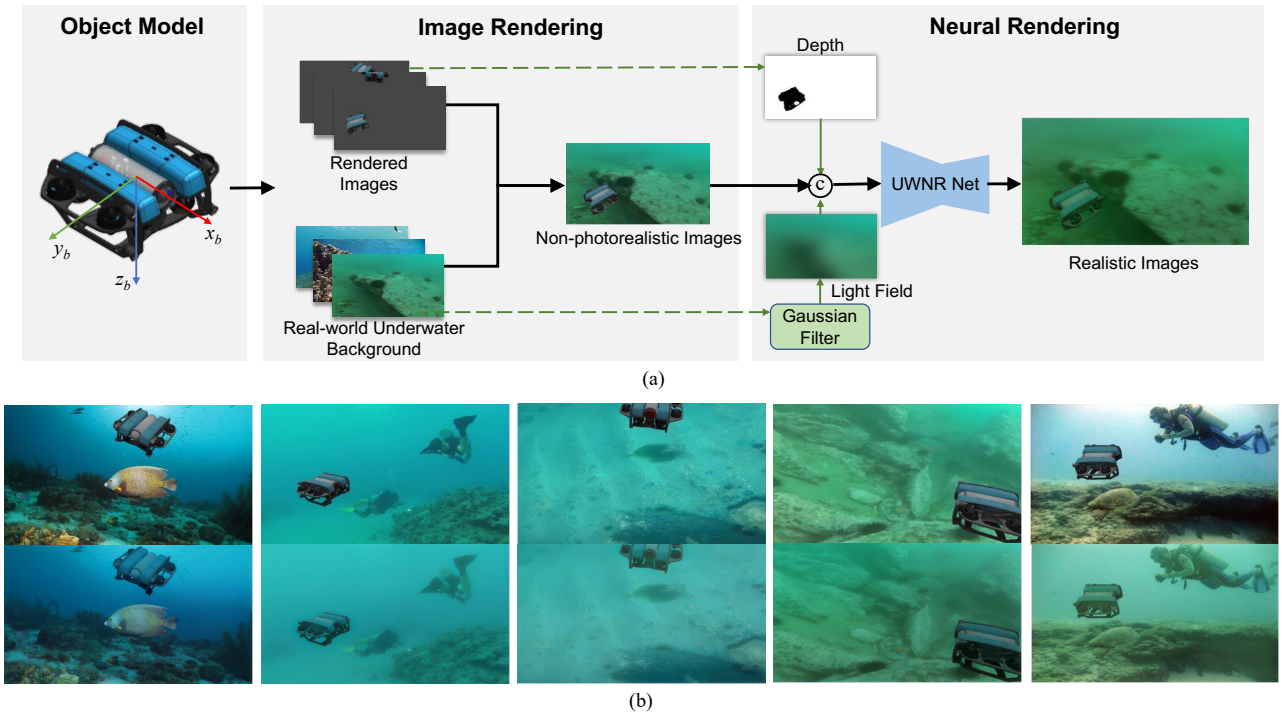

Fig. 2. (a) The training data generation process. First, the 3D BlueROV model is rendered using BlenderProc with known 6D poses and random real underwater images to generate synthetic non-photorealistic images. Then, the light field map is extracted from the real underwater background. Finally, the realistic image is generated by the UWNR network from the non-photorealistic image, its depth and light field map. (b) Top row: Samples of non-photorealistic images rendered by BlenderProc for different types of water. Bottom row: The corresponding realistic images generated by the UWNR-based network.

them, we are the first to employ a motion capture system for more accurate 6D pose annotations of underwater robots. Furthermore, we introduce a benchmark dataset specially designed for pose estimation of underwater ROVs, which covers challenging real-world scenarios. We comprehensively evaluate three types of methods on this benchmark dataset.

III. Dataset

The 6D pose estimation dataset for ROVs is introduced in detail, including the training data generation and the testing data collection in both a pool and the open lake at Maoming.

A. Synthetic Data Generation

To generate synthetic training data, we exploit BlenderProc [19], a Blender-based rendering pipeline that generates synthetic images based on physically based rendering (PBR). High-quality images can be generated by BlenderProc along with comprehensive annotations based on Pybullet physics engine. In our work, BlenderProc renders a 3D model of BlueROV with known 6D poses, projected onto real underwater images $\mathcal{B} = \{b_i | i = 1,..., M\}$, and generates the original synthetic images $\mathcal{X} = \{x_j | j = 1,..., N\}$. Nonetheless, the generated images may appear unrealistic due to the foreground-background distinction. As a pioneer, DeepURL [17] adopts CycleGAN [28] to align the feature space between real underwater and rendered images. However, a CycleGAN-based model is specific to a particular water type and can only generate images corresponding to that specific type. Consequently, in order to detect objects in various types of water, users have to collect multiple real-world data and train separate models for each water type, which is time-consuming. To mitigate the simulation-to-reality gap, we employ the UWNR-based network [20]. This network simulates the main characteristics of different underwater scenes by extracting the natural light field from real underwater images, which have been employed during the rendering stage in BlenderProc. During this process, a multi-scale Gaussian low-pass filter is performed on $\mathcal{B}$ to filter out the reflection component in the environment and retain the illumination component of ambient light. Given a non-photorealistic image x_u and its corresponding underwater background image b_u, the realistic underwater image y_u can be obtained using neural rendering. The underwater light field map is:

$$\mathcal{LF}(b_u) = \frac{1}{3}\sum_{\sigma} Gauss_{\sigma}(b_u) \tag{1}$$

where σ is set to 15, 60 and 90 following [20], respectively.

The light field consistency loss function is defined as:

$$L_{lfc} = \|\mathcal{LF}(y_u) - \mathcal{LF}(\mathcal{LF}(b_u))\|_1 \tag{2}$$

In our proposed method, the underwater real-world background images are collected from related papers, including RUIE [37], Sea-thru [38], UFO-120 [39], and USR-248 [40]. It is worth noting that the UWNR-based network utilizes a single model for generating realistic images of diverse water types. Fig. 2 shows the training data generation process.

B. Real Data Collection

As is shown in Fig. 3, the real images of the BlueROV are captured by a lightweight camera attached to the other ROV (omitted in Fig. 3 for clarity) at 30 frames per second with high resolution (1920×1080 pixels). The camera intrinsic parameters are obtained through underwater calibration [41]. We have collected two subsets of real-world data in both a pool and the open lake at Maoming for quantitative and qualitative evaluation, named Pool and Maoming subsets, respectively.

1) Frames: The coordinate systems include a body-fixed frame $\{O_b\}$, a camera frame $\{O_c\}$ and a marker-fixed tracking frame $\{O_m\}$ (see Fig. 3). The ground-truth pose labels are acquired by a motion capture system for the Pool subset and a vision-based Apriltag system for the Maoming subset. Specifically, the reflective marker frame in the motion capture system and the Apriltag marker frame in the vision-based Apriltag system represent the marker-fixed tracking frame, as shown in Fig. 3. T_b^c represents the transformation from $\{O_b\}$ to $\{O_c\}$, which can be calculated with the transformation T_b^m and T_m^c.

2) Annotations: The experimental pool with a dimension of about 6.0×3.5×1.5m is equipped with a NOKOV motion capture system, which supports the capture area of about 4.0×2.5×1.0m and tracks the reflective markers on the ROV with millimetre precision, enabling accurate annotations collection. Four reflective markers are mounted on each of the ROVs to construct the marker frames $\{O_m\}$ and $\{O_{m'}\}$. The object pose can be calculated with the output marker pose $T_m^{m'}$, as well as the transformations T_b^m from the observed ROV frame $\{O_b\}$ to its corresponding marker frame $\{O_m\}$ and $T_{m'}^c$ from the frame $\{O_{m'}\}$ to the camera frame $\{O_c\}$. In the Maoming subset, a 2D fiducial marker (Apriltag [42]) is mounted above the ROV to estimate the ground truth in the open lake. The vision-based Apriltag system provides pose measurement with accuracy at the degree and centimeter levels. To eliminate biases in the estimated targets within camera reference frames, calibration processes, as proposed in [43], are conducted. Note that the underwater real-world backgrounds in the testing subsets are not utilized in the training data generation process.

C. Dataset Description

Table I summarizes the comparison among datasets in [17], [25] and our proposed ROV6D, highlighting the diversity, challenges and accuracy of our dataset in the following three aspects. (1) **Diversity:** Compared with two realistic underwater styles of the synthetic subset in [17], our subset encompasses diverse realistic styles. (2) **Challenges:** Compared with [17], [25], our real-world subsets introduce challenging cases with partial occlusion and low visibility. (3) **Accuracy:** The annotations provided by the motion capture system achieve millimeter-level precision, outperforming the marker-based method mentioned in [17]. All images in ROV6D are annotated with 6D poses in BOP formats [44].

Training - Synthetic Subset: We first render the object's (BlueROV) 3D model in a random pose within a cube area. The rendered object is then placed onto a randomly selected

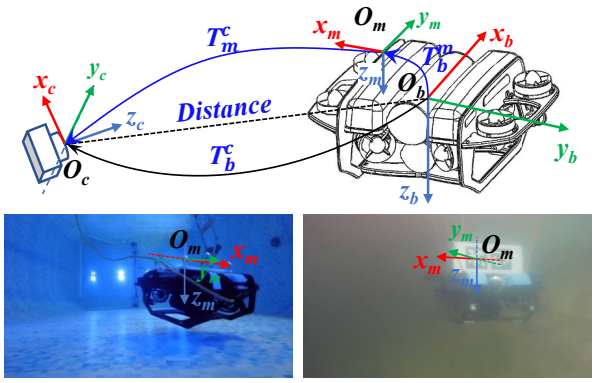

Fig. 3. Top row: Coordinates: a body frame $\{O_b\}$, a camera frame $\{O_c\}$, a marker-fixed tracking frame $\{O_m\}$. Bottom row: The ground-truth pose labels are acquired from the motion capture system (left) and the vision-based Apriltag system (right). The origin of marker-fixed tracking frame $\{O_m\}$ corresponds to the center of the marker or Apirltag. x_*: red, y_*: green, z_*: blue, $_* \in \{b, c, m\}$.

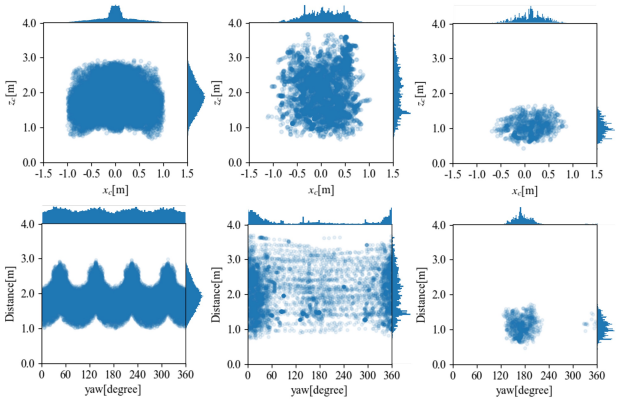

Fig. 4. Position and yaw angle distribution of the pose labels across the training and testing subsets in the camera frame $\{O_c\}$, for Synthetic (left), Pool (middle) and Maoming (right) subsets.

underwater background image. To ensure the generation of realistic images, we employ the natural light field retention scheme proposed by UWNR, which is described in III-A. The synthetic training subset consists of 34K images with a distance ranging from 0.5m to 3.0m, yaw angle ranging from -180° to 180°, pitch and roll angles ranging from -45° to 45°.

Testing - Real-world Subsets: contains **Pool subset** and **Maoming subset**. Fig. 4 visualizes the orientation and position distribution for all subsets with 6D pose annotations in the camera frame $\{O_c\}$. Notably, the orientations of the synthetic subset are well distributed across the 3D space than the real-world subsets. The samples collected from the real-world subsets are shown in Fig. 5. Note that our annotations provide the high-quality pose ground truth for accurate evaluation.

- **Pool Subset:** The Pool subset is a standard test dataset for ROV's 6D pose estimation. This subset is further divided into the **Pool-Basic group** and the **Pool-Occluded group** under different settings of scenes. The Pool-Basic group exhibits some challenges for pose estimation: object motion blur and disturbed water surface. In the Pool-Occluded group, the object is occluded by different

TABLE I
DATASETS FOR UNDERWATER 6D POSE ESTIMATION OF UUVS

Type	Subset	Object	Scale	Water Type	Cases	Pose Annotation	Accuracy	Quantitative
Rendered	Synthetic [17]	Aqua2	37K	Two&Clear	No occlusion	Simulator	/	✓
	Synthetic (Ours)	BlueROV	34K	**Multi**&Clear	No occlusion	Simulator	/	✓
Real-world (Indoor)	Pool [17]	Aqua2	11K	Single&Clear	No occlusion	Vision-based marker	Medium	✓
	Pool (Ours)	BlueROV	6.6K	Single&Clear	**Partial occlusion**	Motion capture system	High	✓
Real-world (Outdoor)	Barbado 2017 [25]	Aqua2	188	Single&Clear	No occlusion	Custom-built annotator	/	✓
	Barbado GoPro [17]	Aqua2	/	Single&Clear	No occlusion	/	/	✗
	Maoming (Ours)	BlueROV	2140	Single&**Turbid**	No occlusion	Vision-based marker	Medium	Partial

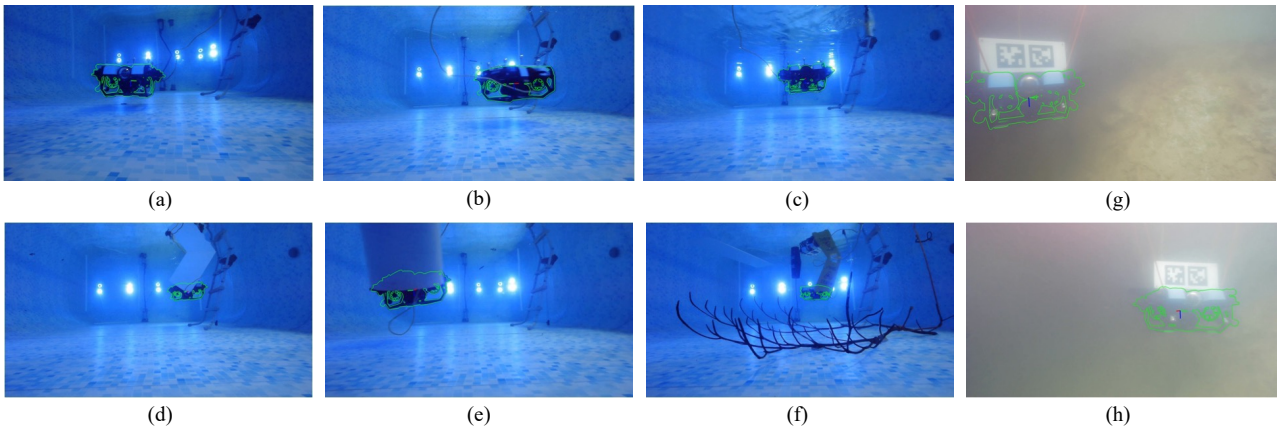

Fig. 5. Samples images from the test subsets. Images (a)-(c) represent three scenes from the Pool-Basic group: no disturbance, object motion blur and disturbed water surface; Images (d)-(f) exhibit cases from the Pool-Occluded group, where the object is occluded by a board, a pipe, and multiple obstacles, respectively. While (g) and (h) come from the Maoming subset. The green lines represent the edges of the 3D model projected from the ground-truth poses.

obstacles (e.g., board, pipe and branches). Currently, the Pool subset contains 6.6K images with a distance ranging from 0.5m to 3.5m and yaw angle ranging from -180° to 180°, as shown in Fig. 4.

- **Maoming Subset:** The Maoming subset consists of 2140 real-world underwater images, 1071 of which have 6D object pose annotations acquired by a vision-based Apriltag system. The remaining 1069 images, for which satisfactory 6D pose annotations cannot be obtained due to the Apriltag being too distant for clear visibility, are considered a challenging unlabelled group. Unlike the Pool subset, the underwater visibility in the open lake at Maoming is significantly impacted by turbidity, resulting in a limited range of underwater vision. Therefore, the real images with 6D pose annotations in the Maoming subset are collected within a distance range of 0.5m to 2.0m. In comparison, the unannotated images are captured within a distance range of 2.0m to 4.0m, and beyond that distance, the target becomes completely invisible.

IV. EVALUATION AND DISCUSSION

Given an RGB image I and the 3D CAD model $\mathcal{M}$ of the object, we aim at estimating the 6D object pose $\boldsymbol{P} = [\boldsymbol{R}|\boldsymbol{t}]$ w.r.t the camera for the object present in I with the 3D rotation $\boldsymbol{R}$ and the 3D translation $\boldsymbol{t}$ of the detected object. Using the constructed Pool and Maoming subsets, we evaluate the keypoint-based method DeepURL [17], the refinement-based method CosyPose [21], as well as two dense coordinates-based methods CDPN [22] and GDR-Net [23].

A. Metrics

We conduct the evaluation using three commonly-used metrics: 2D projection [45], ADD(-S) [33], [46] and $n°$ n cm [47]. The metric 2D projection represents the mean distance between 2D projections of 3D model points given the predicted and ground-truth poses. Generally, a pose is considered correct if the 2D projection is below 5 pixels. Following [17], we also show the results with the threshold of 10 pixels. ADD is utilized to compute the 3D Euclidean averaged distance of all model points between the estimated and the ground-truth pose. For ADD, the estimated pose is considered to be correct if the averaged distance is below 10% of the model diameter ($0.1d$). The ADD-S metric is utilized for symmetric objects to compute the mean distance to the closest model point. Similar to [23], we also provide pose estimation accuracy using the threshold of $0.02d$. Additionally, we calculate the rotation error (Re) and the translation error (Te). The $n°$ n cm metric considers an estimated pose to be correct if its rotation error is within $n°$ and the translation error is below n cm.

B. Experimental Setup

Hardware. Experiments were conducted on a desktop with an Intel 3.40GHz CPU and an NVIDIA 2080Ti GPU.

Fixed parameters. The object detector and pose estimator in the evaluation methodology are trained on our synthetic data. The parameters of each method are fixed for all the training and testing subsets.

Pool subset split for evaluation. We treat each group in the Pool subset (Pool-Basic and Pool-Occluded) as a separate

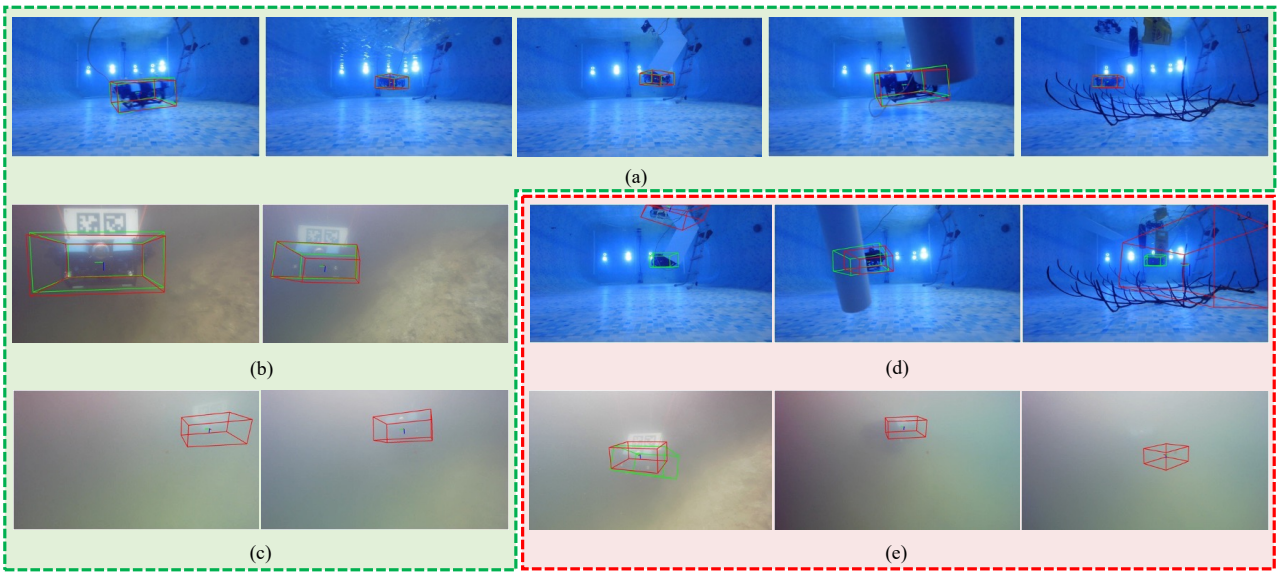

Fig. 6. Sample results from our dataset. The green and red wireframes in each image denote the ground truth and prediction, respectively. Within the green dashed box, images (a)-(c) are the relatively successful results from the Pool subset (a), the labelled group (b) and the unlabelled group (c) of the Maoming subset. Within the red dashed box, images (d) and (e) are the failure cases. In cases (d) from the Pool subset, the visible parts are too ambiguous for accurate pose estimation. In cases (e) from the Maoming subset, the object's appearance is blurred due to the water's high turbidity and the object being too far away.

challenge to avoid the evaluation results being dominated by the whole subset.

C. Experimental Results

Performance on the Pool subset. In TABLE II and TABLE III, we compare four methods [17], [21]–[23] trained on the Synthetic subset and tested on the Pool subset in terms of the ADD(-S) metric, the 2D projection metric and the 5° 5cm metric. Due to the high similarity between the front and back of the BlueROV, we analyze the influence of symmetry. The results in TABLE II demonstrate that symmetry significantly affects the performance of all methods, especially concerning rotation estimation. It is worth noting that the BlueROV is not perfectly symmetrical, but only because of the distinctions between the front and rear covers of its electronics enclosure. Similar to other works [23], [31], [48], we ignore these minor distinctions on the BlueROV and consider it a symmetric object for evaluation in the subsequent sections. For the **Pool-Basic group**, GDR-Net [23] is the top-performing method with the ADD-S 0.1d metric of 99.72% and the 5° 5 cm metric of 61.41%. Under the 2D projection metric (5px), GDR-Net is only a little inferior to CosyPose [21], with 68.97% to 69.41%. However, CosyPose is a refinement-driven method and runs significantly slower than GDR-Net to achieve comparable results. The improved performance of GDR-Net benefits from its dense correspondence network, where the multiple geometrical features can be extracted and trained in an end-to-end manner. On the other hand, the other dense correspondences-based method CDPN [22] is not end-to-end trainable since the deterministic pose calculated through PnP/RANSAC stage is inherently non-differentiable. DeepURL [17] is a sparse keypoint-based method and its potential for improving estimation accuracy is limited, especially in water. The comparison results on the **Pool-Occluded group** are shown in TABLE III. Compared to CDPN [22], GDR-Net [23] achieves a significant improvement of 32.94% under the ADD-S 0.02d metric, increasing from 45.14% to 60.01%. Moreover, GDR-Net outperforms other methods [21], [22] by a margin of 52.09% and 26.13% in terms of 5° 5 cm. These results demonstrate the robustness of dense correspondence-based geometric representations proposed in [23] to handle occlusions. Fig. 6 (a) shows some qualitative results.

TABLE II
RESULTS ON THE POOL-BASIC GROUP IN TERMS OF THE ADD(-S) METRIC, THE 2D PROJECTION METRIC AND THE 5° 5CM METRIC

Method	Sym*	ADD(-S)		2D Proj.		5°5cm	Re	Te
		0.02d	0.1d	5px	10px			
DeepURL [17]	✗	-	43.06	-	32.81	-	63.61	0.14m
CosyPose [21]	✗	-	56.17	47.67	67.21	42.98	45.01	**0.03m**
CDPN [22]	✗	2.38	57.78	43.95	66.51	36.35	48.76	0.05m
GDR-Net [23]	✗	3.39	64.80	50.08	68.44	43.63	54.02	**0.03m**
DeepURL [17]	✓	-	59.01	-	37.38	-	12.06	0.14m
CosyPose [21]	✓	-	63.54	**69.41**	92.57	60.28	5.20	**0.03m**
CDPN [22]	✓	58.28	98.47	59.08	88.66	49.15	5.57	0.05m
GDR-Net [23]	✓	**73.18**	**99.72**	68.97	**96.34**	**61.41**	**4.69**	**0.03m**

* denotes evaluating it as a symmetric object.

TABLE III
RESULTS ON THE POOL-OCCLUDED GROUP IN TERMS OF THE ADD-S METRIC, THE 2D PROJECTION METRIC AND THE 5° 5CM METRIC

Method	ADD-S		2D Proj.		5°5cm	Re	Te
	0.02d	0.1d	5px	10px			
DeepURL [17]	-	54.67	-	37.93	-	15.30	0.24m
CosyPose [21]	-	34.05	59.70	74.46	31.58	9.68	**0.16m**
CDPN [22]	45.14	88.43	67.97	86.07	38.08	9.35	0.22m
GDR-Net [23]	**60.01**	**89.74**	**76.72**	**90.09**	**48.03**	**9.24**	0.17m

Performance on the Maoming subset. TABLE IV shows the results on the labelled group of the Maoming subset.

TABLE IV
RESULTS ON THE MAOMING SUBSET (LABELLED GROUP) IN TERMS OF THE ADD-S METRIC, THE 2D PROJECTION METRIC AND THE 5° 5CM METRIC

Method	ADD-S		2D Proj.		5°5cm	Re	Te
	0.02d	0.1d	5px	10px			
CosyPose [21]	-	27.54	2.61	31.65	14.28	10.28	**0.05m**
CDPN [22]	40.24	**85.61**	5.70	45.94	22.46	**9.99**	0.06m
GDR-Net [23]	**44.82**	85.24	**5.79**	**46.13**	**23.06**	12.48	0.06m

TABLE V
EVALUATE UWNR AND CYCLEGAN FOR DATA GENERATION ON ROTATION ERROR, TRANSLATION ERROR, METRIC ADD-S AND 5° 5CM

Subset	Strategy	ADD-S	5° 5cm	Re	Te
Pool-Basic	w/ CycleGAN	99.47	54.89	5.10	0.04m
	w/o UWNR	99.67	51.34	5.01	**0.03m**
	w/ UWNR	**99.72**	**61.41**	**4.69**	**0.03m**
Pool-Occluded	w/ CycleGAN	66.54	36.16	21.51	0.59m
	w/o UWNR	**91.25**	43.15	9.55	0.18m
	w/ UWNR	89.74	**48.03**	**9.24**	**0.17m**
Maoming	w/ CycleGAN	30.72	5.14	43.96	0.40m
	w/o UWNR	69.87	14.04	20.56	0.19m
	w/ UWNR	**85.24**	**23.06**	**12.48**	**0.06m**

Coordinate-based methods [22], [23] outperform CosyPose [21]. It should be pointed out that the keypoint-based method DeepURL [17] suffers a failure since there are not enough feature points for PnP/RANSAC stage. Due to the significant domain gap between the training subset and the Maoming subset, these pose estimators perform poorly in handling low-visibility underwater images from the Maoming subset. As shown in Fig. 6 (b), when the object's appearance is relatively clear, the estimation performance is satisfactory. The visibility decreases as the distance between the two ROVs increases. In Fig. 6 (c) where the Apriltag is undetectable, learning-based methods can still estimate the object's pose.

Effectiveness of Realistic Image Generation. To verify the effectiveness of the UWNR-based image generation strategy, we train GDR-Net [23] with images generated by UWNR, images generated by CycleGAN and non-photorealistic images, respectively. TABLE V shows that when training GDR-Net with images generated by UWNR, the overall performance on all subsets exceeds that achieved with images generated by CycleGAN. Besides, the network performs well on the Pool subset even trained on the non-photorealistic data. However, its generalization ability in wild scenarios is poor. Specifically, the results obtained by training with UWNR-generated data outperform those achieved using images solely rendered by the simulator on the Maoming subset. These results fully demonstrate the advantages of our realistic image generation strategy for complex underwater environments.

Failure cases and Open Problems. Fig. 6 (d) and (e) show some typical failure cases. Some failure cases occur when the object undergoes heavy occlusion. In these situations, the available object information is often severely limited due to occlusion. Thus, occlusion is a major challenge for current methods, which can be demonstrated by the differences in scores between Pool-Basic and Pool-Occluded groups, as shown in TABLE II and TABLE III. Under the ADD-S metric, 2D projection metric (5px) and the 5° 5cm metric, all methods perform at least 10% better on the Pool-Basic group compared to the Pool-Occluded group. On the other hand, suspended sediments would lead to significant visibility degradation, resulting in poor visual quality and a huge domain gap between synthetic and real images. This domain discrepancy or domain shift can cause failure results in model generalization, as shown in Fig. 6 (e). In addition, scores on the Maoming subset show that poor visibility and severe domain gaps present significant problems for the methods that rely on synthetic training images.

V. CONCLUSIONS

This paper proposes an underwater benchmark dataset, ROV6D, for the pose estimation of ROVs, which offers large-scale synthetic images and real-world images, as well as their corresponding pose annotations. To bridge the reality gap, a natural light field extraction method is employed to generate realistic synthetic images under different water conditions. The real-world data, collected in both a pool and the open lake, contains multiple challenging scenarios, e.g., partial occlusion and low visibility. We evaluate several approaches on our dataset for underwater relative localization. Experimental results show the superior performance of a dense coordinates-based approach compared to other comparative methods on commonly used metrics. In the future, we will concentrate on the challenges of 6D object pose estimation in underwater environments, particularly heavy occlusions and high turbidity, and train a usable model with self-supervised real data.

REFERENCES

[1] G. Billings and M. Johnson-Roberson, "Silhonet-fisheye: Adaptation of a roi based object pose estimation network to monocular fisheye images," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4241–4248, 2020.

[2] M. Kisantal, S. Sharma, T. H. Park, D. Izzo, M. Märtens, and S. D'Amico, "Satellite pose estimation challenge: Dataset, competition design, and results," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, no. 5, pp. 4083–4098, 2020.

[3] S. Li, C. De Wagter, and G. C. De Croon, "Self-supervised monocular multi-robot relative localization with efficient deep neural networks," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 9689–9695.

[4] Z. Fan, Z. Chen, J. Wu, and C. Pei, "Pose recognition for dense vehicles under complex street scenario," in *2019 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2019, pp. 1–4.

[5] I. G. Zubov, "Estimation of vehicle pose with monocular camera," in *2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*. IEEE, 2019, pp. 395–397.

[6] M. T. Isik and O. B. Akan, "A three dimensional localization algorithm for underwater acoustic sensor networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 9, pp. 4457–4463, 2009.

[7] J. Yan, D. Guo, X. Luo, and X. Guan, "Auv-aided localization for underwater acoustic sensor networks with current field estimation," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 8855–8870, 2020.

[8] B. Joshi, S. Rahman, M. Kalaitzakis, B. Cain, J. Johnson, M. Xanthidis, N. Karapetyan, A. Hernandez, A. Q. Li, N. Vitzilaios *et al.*, "Experimental comparison of open source visual-inertial-based state estimation algorithms in the underwater domain," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 7227–7233.

[9] A. Quattrini Li, A. Coskun, S. M. Doherty, S. Ghasemlou, A. S. Jagtap, M. Modasshir, S. Rahman, A. Singh, M. Xanthidis, J. M. O'Kane *et al.*, "Experimental comparison of open source vision-based state estimation algorithms," in *2016 International Symposium on Experimental Robotics*. Springer, 2017, pp. 775–786.

[10] S. Skaff, J. J. Clark, and I. M. Rekleitis, "Estimating surface reflectance spectra for underwater color vision." in *BMVC*, 2008, pp. 1–10.

[11] B. McGlamery, "A computer model for underwater camera systems," in *Ocean Optics VI*, vol. 208. SPIE, 1980, pp. 221–231.

[12] J. S. Jaffe, "Computer modeling and the design of optimal underwater imaging systems," *IEEE Journal of Oceanic Engineering*, vol. 15, no. 2, pp. 101–111, 1990.

[13] D. Akkaynak and T. Treibitz, "A revised underwater image formation model," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6723–6732.

[14] D. B. dos Santos Cesar, C. Gaudig, M. Fritsche, M. A. dos Reis, and F. Kirchner, "An evaluation of artificial fiducial markers in underwater environments," in *OCEANS 2015 - Genova*, 2015, pp. 1–6.

[15] E. Simetti, F. Wanderlingh, S. Torelli, M. Bibuli, A. Odetti, G. Bruzzone, D. L. Rizzini, J. Aleotti, G. Palli, L. Moriello, and U. Scarcia, "Autonomous underwater intervention: Experimental results of the maris project," *IEEE Journal of Oceanic Engineering*, vol. 43, no. 3, pp. 620–639, 2018.

[16] J. J. Fernandez, M. Prats, P. J. Sanz, J. C. Garcia, R. Marin, M. Robinson, D. Ribas, and P. Ridao, "Grasping for the seabed: Developing a new underwater robot arm for shallow-water intervention," *IEEE Robotics Automation Magazine*, vol. 20, no. 4, pp. 121–130, 2013.

[17] B. Joshi, M. Modasshir, T. Manderson, H. Damron, M. Xanthidis, A. Q. Li, I. Rekleitis, and G. Dudek, "Deepurl: Deep pose estimation framework for underwater relative localization," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 1777–1784.

[18] K. Koreitem, J. Li, I. Karp, T. Manderson, F. Shkurti, and G. Dudek, "Synthetically trained 3d visual tracker of underwater vehicles," in *OCEANS 2018 MTS/IEEE Charleston*. IEEE, 2018, pp. 1–7.

[19] M. Denninger, M. Sundermeyer, D. Winkelbauer, D. Olefir, T. Hodan, Y. Zidan, M. Elbadrawy, M. Knauer, H. Katam, and A. Lodhi, "Blenderproc: Reducing the reality gap with photorealistic rendering," in *International Conference on Robotics: Sciene and Systems, RSS 2020*, 2020.

[20] T. Ye, S. Chen, Y. Liu, Y. Ye, E. Chen, and Y. Li, "Underwater light field retention: Neural rendering for underwater imaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 488–497.

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

[22] Z. Li, G. Wang, and X. Ji, "Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7678–7687.

[23] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 611–16 621.

[24] M. C. Nielsen, M. H. Leonhardsen, and I. Schjølberg, "Evaluation of posenet for 6-dof underwater pose estimation," in *OCEANS 2019 MTS/IEEE SEATTLE*. IEEE, 2019, pp. 1–6.

[25] K. Koreitem, J. Li, I. Karp, T. Manderson, F. Shkurti, and G. Dudek, "Synthetically trained 3d visual tracker of underwater vehicles," in *OCEANS 2018 MTS/IEEE Charleston*, 2018, pp. 1–7.

[26] J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson, "Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images," *IEEE Robotics and Automation letters*, vol. 3, no. 1, pp. 387–394, 2017.

[27] F. Shkurti, W.-D. Chang, P. Henderson, M. J. Islam, J. C. G. Higuera, J. Li, T. Manderson, A. Xu, G. Dudek, and J. Sattar, "Underwater multi-robot convoying using visual tracking by detection," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 4189–4196.

[28] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[29] B. Tekin, S. N. Sinha, and P. Fua, "Real-time seamless single shot 6d object pose prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 292–301.

[30] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "Pvnet: Pixel-wise voting network for 6dof pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4561–4570.

[31] Y. Bukschat and M. Vetter, "Efficientpose: An efficient, accurate and scalable end-to-end 6d multi object pose estimation approach," *arXiv preprint arXiv:2011.04307*, 2020.

[32] J. Lin, Z. Wei, Z. Li, S. Xu, K. Jia, and Y. Li, "Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3560–3569.

[33] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "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*. Springer, 2013, pp. 548–562.

[34] A. Krull, E. Brachmann, F. Michel, M. Y. Yang, S. Gumhold, and C. Rother, "Learning analysis-by-synthesis for 6d pose estimation in rgb-d images," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 954–962.

[35] T. Hodan, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis, "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)*. IEEE, 2017, pp. 880–888.

[36] Z. Fan, Y. Zhu, Y. He, Q. Sun, H. Liu, and J. He, "Deep learning on monocular object pose detection and tracking: A comprehensive overview," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1–40, 2022.

[37] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo, "Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light," *IEEE transactions on circuits and systems for video technology*, vol. 30, no. 12, pp. 4861–4875, 2020.

[38] D. Akkaynak and T. Treibitz, "Sea-thru: A method for removing water from underwater images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1682–1691.

[39] M. J. Islam, P. Luo, and J. Sattar, "Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception," *arXiv preprint arXiv:2002.01155*, 2020.

[40] M. J. Islam, S. S. Enan, P. Luo, and J. Sattar, "Underwater image super-resolution using deep residual multipliers," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 900–906.

[41] M. Pedersen, S. Hein Bengtson, R. Gade, N. Madsen, and T. B. Moeslund, "Camera calibration for underwater 3d reconstruction based on ray tracing using snell's law," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1410–1417.

[42] J. Wang and E. Olson, "Apriltag 2: Efficient and robust fiducial detection," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 4193–4198.

[43] A. Gouda, A. Ghanem, and C. Reining, "Dopose-6d dataset for object segmentation and 6d pose estimation," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2022, pp. 477–483.

[44] T. Hodan, F. Michel, E. Brachmann, W. Kehl, A. GlentBuch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis *et al.*, "Bop: Benchmark for 6d object pose estimation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 19–34.

[45] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, and C. Rother, "Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3364–3372.

[46] T. Hodaň, J. Matas, and Š. Obdržálek, "On evaluation of 6d object pose estimation," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 606–619.

[47] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in rgb-d images," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2930–2937.

[48] M. Rad and V. Lepetit, "Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3828–3836.