2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

MV-ROPE: Multi-view Constraints for Robust Category-level Object Pose and Size Estimation

Jiaqi Yang[1]*, Yucong Chen[1]*, Xiangting Meng[1]*, Chenxin Yan[1], Min Li[1], Ran Cheng[2], Lige Liu[2], Tao Sun[2], and Laurent Kneip[1]†

***Abstract*— Recently there has been a growing interest in category-level object pose and size estimation, and prevailing methods commonly rely on single view RGB-D images. However, one disadvantage of such methods is that they require accurate depth maps which cannot be produced by consumer-grade sensors. Furthermore, many practical real-world situations involve a moving camera that continuously observes its surroundings, and the temporal information of the input video streams is simply overlooked by single-view methods. We propose a novel solution that makes use of RGB video streams. Our framework consists of three modules: a scale-aware monocular dense SLAM solution, a lightweight object pose predictor, and an object-level pose graph optimizer. The SLAM module utilizes a video stream and additional scale-sensitive readings to estimate camera poses and metric depth. The object pose predictor then generates canonical object representations from RGB images. The object pose is estimated through geometric registration of these canonical object representations with estimated object depth points. All per-view estimates finally undergo optimization within a pose graph, culminating in the output of robust and accurate canonical object poses. Our experimental results demonstrate that when utilizing public dataset sequences with high-quality depth information, the proposed method exhibits comparable performance to state-of-the-art RGB-D methods. We also collect and evaluate on new datasets containing depth maps of varying quality to further quantitatively benchmark the proposed method alongside previous RGB-D based methods. We demonstrate a significant advantage in scenarios where depth input is absent or the quality of depth sensing is limited.**

I. Introduction

The detection of objects and the estimation of their 6D pose and size is an important problem in applications such as robotics and augmented reality. Generally, the problem can be divided into instance-level and category-level pose estimation. In the former, we assume knowledge about a small number of exact shape priors (e.g. meshes, CAD models), thus reducing the problem to discrete model selection, correspondence estimation, and pose estimation. The present work looks at the more general case of category-level pose estimation, in which the exact shape and appearance of the observed objects are assumed to be unknown. It defines the center, orientation, and size of objects at the category level, ultimately providing absolute pose and size estimations.

With the introduction of Normalized Object Coordinate Spaces (NOCS) [1], there has been a surge in research efforts in recent years, continuously improving the accuracy of category-level pose and size estimation. These methods significantly broaden the applicability of object pose and size estimation in real-world scenarios. However, while NOCS methods enable the extraction of canonical object representations from RGB images, achieving robust object pose and size estimation still requires integration with additional depth information for accurate alignment. The depth information is commonly given in the form of a direct depth channel reading or image-based depth predictions. However, it is well-understood that depth camera readings may easily suffer from measurement partiality or artifacts [2], and that single-view depth prediction may fail to accurately reflect depth details or absolute scale [3]. Current RGB-D based methods are intrinsically limited in their real-world applicability due to their reliance on high-quality depth sensing while methods that purely rely on images are not yet able to perform on par with the state-of-the-art.

The motivation of our work is fueled by three important insights: (1) Obtaining an accurate depth map poses practical challenges for consumer-grade devices such as the iPhone or Azure Kinect, especially when it comes to small objects with smooth surfaces and intricate structures. (2) We recognize the fact that in many practical applications, we do not only have a single image being taken of the environment, but the sensor is often mobile and continuously gathers novel views of the scene. As a result, we may indeed continuously generate predictions from nearby images and incrementally and robustly generate improved object pose predictions over time. (3) We may rely on motion stereo to perform dense depth reconstruction and thereby bypass the need for a depth camera or inaccurate single-view depth predictions.

We exploit these insights in a novel framework, denoted *MV-ROPE*. It combines a scale-aware dense monocular SLAM module, a lightweight object pose estimator, as well as an object-level pose graph optimization module. Our SLAM module takes a monocular RGB image sequence along with additional scale information as input. It outputs accurate camera poses as well as dense metric depth estimations through a dense bundle adjustment layer. The lightweight object pose estimator predicts pixel-wise object canonical representations from RGB images. By performing geometric registration between depth and canonical object points, we can obtain the object pose within each keyframe. All keyframe object poses are subsequently optimized by an object-level pose graph optimization module. This allows us to ultimately obtain accurate and robust 6D object pose and size estimations. In summary, we make the following

[1] ShanghaiTech University
[2] Midea RoboZone
* Authors contributed equally to this work.
† Corresponding author.

contributions:

- We present the first multi-view RGB framework for robust and accurate category-level object pose estimation. Our primary modules comprise a scale-aware dense monocular SLAM module, a lightweight object pose estimator, and an object-level pose graph optimization module, which is utilized for averaging multi-view object pose estimations. The parallel operation of these modules yields robust estimations of canonical object pose and size in interactive level real-time.
- In an aim to circumvent the inherent scale invariance property of monocular SLAM, we propose a novel scale-aware monocular dense SLAM module making use of either inertial readings, stereo images, or direct depth readings as obtained from an RGB-D camera.
- Even when relying solely on images, our results demonstrate performance on par with or even better than previous state-of-the-art RGB-D methods on publicly available datasets. Additionally, we introduce a new dataset MEREAL that captures real-world scenarios, including RGB-D sequences and ground truth annotations obtained from different types and qualities of depth sensors. This dataset serves as a benchmark for evaluating the performance of our proposed method and priorly proposed RGB-D based algorithms.

We demonstrate through extensive experimental results that our algorithm achieves superior performance in situations where direct depth input is unavailable or when using low-quality depth readings.

II. Related Works

Typical object pose and size estimation is different from object pose tracking, which is also called relative object pose estimation. Pose tracking methods [4], [5], [6], [7] do not reason about an object's intrinsic shape or absolute pose, but only aim at capturing relative pose variations, primarily by tracking 3D keypoints across different frames. On the other hand, absolute pose and size estimation of objects generally falls into one of two categories: instance-level, and category-level. Instance-level pose estimation [8], [9], [10]—while precise—imposes strict requirements in terms of data. It necessitates knowledge of exact CAD models or meshes corresponding to the observed objects to define each object's pose. However, securing such information is often a significant challenge in practical applications. It is not further discussed here, as the present paper focuses on category-level pose estimation. The latter expands the practical usability of object pose and size estimation in real-world scenarios. Although a few works directly learn object poses [11], [12], [13], the fundamental idea behind category-level pose estimation is to replace CAD models with canonical object representations. By using canonical representations and real-world object points, we can directly obtain a 7DoF similarity transformation by employing geometric solvers [14], [15] or neural networks for implicit prediction. The process of obtaining absolute object poses can therefore be classified by the employed canonical representation:

Deformed shape priors. Even in the absence of an exact CAD model, a categorical shape prior can still be utilized by deforming it to match the canonical object model [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26]. ShapePrior [16] learns a deformation field and applies it to a categorical shape prior, allowing for the reconstruction of the object representations in canonical space. DPDN [18] uses a self-supervised approach to generate better-deformed shape priors. Although it avoids the need for exact CAD models, it still needs representative CAD models at the category level. Recent work uses learnable queries as shape prior alternatives [27] or explores the potential of prior-free methods [28].

Generated object models. In GCASP [29], a latent code is utilized to reconstruct canonical object points. Additionally, Sim(3)-invariant 3D features are employed to estimate the object pose. On the other hand, works such as CenterSnap [30] and ShAPO [31] utilize neural implicit representations for the canonical object model. These approaches optimize the latent code for better poses and reconstructed objects. Methods like GPV-Pose [32] directly regress transformation to canonical space and then reconstruct object points in canonical views for refinement. These methods, however, also need geometric information from accurate depth channels for their features or latent codes to generate models.

Regression of canonical representations. Our method belongs to the category that utilizes semantic or geometric features to regress a canonical representation [1], [33], [34], [35], [36]. NOCS [1] utilizes a single RGB image to predict normalized object coordinates for each pixel. It then employs correspondences between NOCS points and object points from the depth image to get object pose and size. MetricScale [37] independently predicts object metric scale and object center by networks and utilizes predicted NOCS to recover object orientations. [35] enhances NOCS by semantically-aware keypoints guidance, which requires additional depth input. Regression of canonical representations is intuitive and easy to operate. More importantly, such methods have the potential to use the semantic information contained in RGB images to obtain geometric information, which avoids the need for direct depth input channels.

To the best of our knowledge, our method is the first to employ multi-view constraints of input RGB video streams to obtain reliable absolute object pose estimation results.

III. Methodology

We will first provide an overview of the system before going into further details on camera pose and dense depth estimation, object instance segmentation and association, object pose estimation, and pose graph optimization.

A. Overview

From a high-level perspective, our framework accepts an RGB image sequence $\{\mathscr{I}_i\}$ and one of a few possible scale-sensitive readings. Additional information to help resolve the global scale factor can be in the form of stereo images

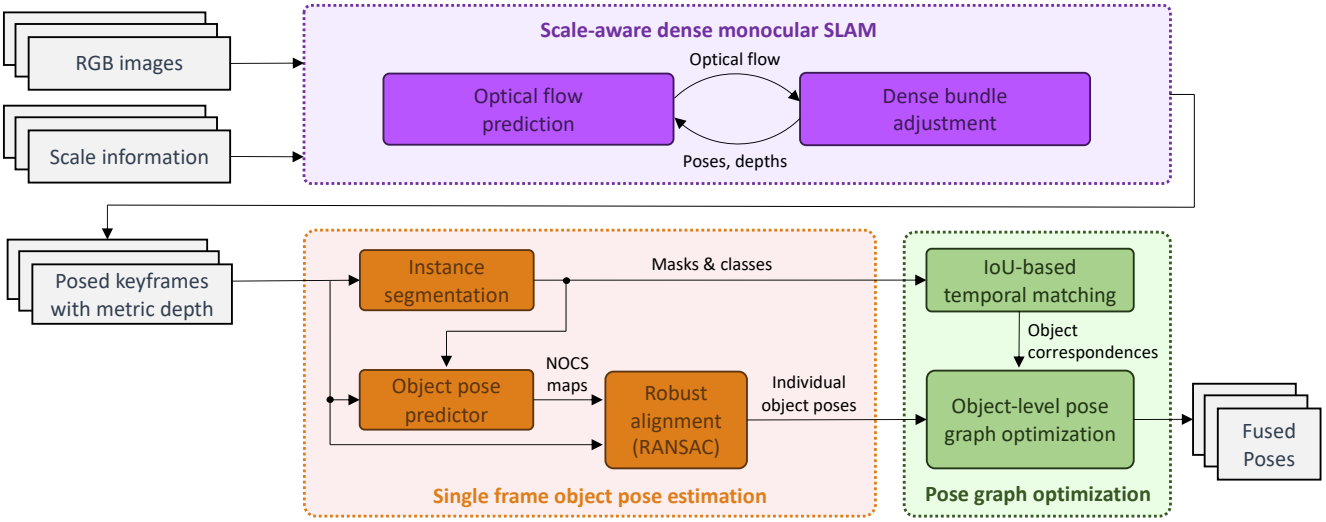

Fig. 1: Overview of the complete, proposed system. The input for the first block (in purple), is a continuous image stream accompanied by scale information. The output of this block includes keyframe camera poses and dense metric depth maps. The second block (in orange) is the single-view object pose estimation module. It utilizes an object pose estimator to obtain NOCS maps for each segmented object. These NOCS maps are then aligned with partial dense depth maps using a RANSAC framework to estimate the pose of each individual object. A third block (in green) finally establishes object correspondences and optimizes object poses over time.

or merely IMU measurements. For comparison purposes, a third alternative using direct depth readings is also supported. The output of our framework consists of camera poses and dense metric depth maps for each keyframe i, represented as $\xi_{\text{ref}}^{c_i} \in \text{SE}(3)$ and $\mathscr{D}_i \in \mathbb{R}^{H \times W}$, respectively. Additionally, our framework provides the pose of each object instance k within a global reference frame, denoted as $\xi_{\text{ref}}^{\mathbf{o}_k} \in \text{Sim}(3)$. Note that $\text{Sim}(3)$ typically refers to the group of similarity transformations in 3D space. To facilitate comprehension, we designate the first keyframe as the reference frame, and $\mathscr{I}_{\text{ref}} = \mathscr{I}_0$.

For ease of representation, going forward, we will consistently use superscript $\mathbf{o}$ to represent objects, $\mathbf{b}$ to represent the IMU body, and superscript $\mathbf{c}$ to represent the camera frame. We use R and t to represent the rotation and translation part of pose ξ. Additionally, we will use indices i and k to refer to frames and objects, respectively.

Our framework is illustrated in Figure 1. The first block runs scale-aware dense monocular SLAM with the purpose of obtaining accurate poses and metric depth maps for each keyframe (details in Sec. III-B). Towards object-centric perception, our system includes a second block that consists of an instance segmentation network and a lightweight object predictor. Inspired by [1], the object pose predictor predicts NOCS maps for each individual object instance. Having NOCS maps and the back-projected 3D points derived from the depth maps, the second, single-view block concludes by employing the Umeyama algorithm [15] within RANSAC to compute individual object poses. Concurrently, a third block takes all single view object pose estimates and incorporates them along with keyframe poses into a multi-view object-level pose graph optimization module. The entire process results in incrementally refined, globally consistent canonical object poses $\xi_{\text{ref}}^{\mathbf{o}_k}$ for each instance $\mathbf{o}_k$.

B. Camera Pose and Metric Depth Estimation

Our camera pose and metric depth estimation module is inspired by other dense geometric bundle adjustment frameworks such as DROID-SLAM [38], and VOLDOR++ [39], and consists of two elements: recurrently refined optical flow and a scale-aware dense bundle adjustment layer. The optical flow module is designed based on [40], and the accompanying scale-aware dense bundle adjustment layer optimizes both pose and depth while incorporating scale constraints from different sources of scale information. The objective includes a reprojection error term and a scale constraint term.

1) Reprojection Error Minimization: The first objective aims to minimize the covariance reweighted sum of squared reprojection errors. The reprojection errors are calculated as

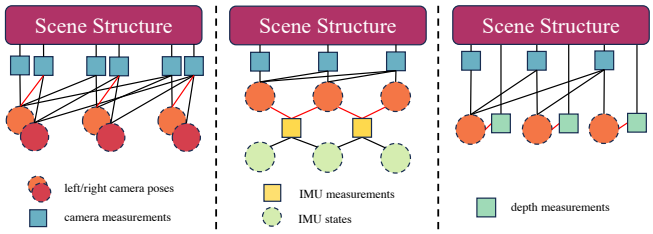

Fig. 2: Graphical models of our scale-aware dense SLAM system. From left to right: model using stereo images, IMU measurements, and depth readings. The red lines represent scale-aware factors in our bundle adjustment layer.

the difference between the target location of 2D image points as predicted by optical flow and as obtained by applying the estimated relative pose and the dense depth to perform image warping. The reprojection objective is given by:

$$E_r(\xi^{\mathbf{c}}_{\text{ref}}, \mathscr{D}) = \sum_{\{i,j\}\in\mathscr{E}} ||\mathscr{P}_{ij} - \Pi_{\mathbf{c}}\left(\tau\left(\xi^{\mathbf{c}_j}_{\mathbf{c}_i}, \Pi^{-1}_{\mathbf{c}}(\mathscr{D}_i)\right)\right)||^2_{\sum_{ij}}, \tag{1}$$

where $\xi^{\mathbf{c}}_{\text{ref}} = \{\xi^{\mathbf{c}_i}_{\text{ref}}\}$ is the set of all keyframe poses to be estimated, $\mathscr{D} = \{\mathscr{D}_i\}$ is the set of all dense depth maps to be estimated, $\mathscr{E}$ is the set of all keyframe-pairs between which residuals are evaluated, $\Pi_{\mathbf{c}}$ is a camera projection function that takes a $H\times W\times 3$ tensor of 3D world points and returns the $H\times W\times 2$ tensor of corresponding image points, and $\Pi^{-1}_{\mathbf{c}}$ is the corresponding inverse mapping that takes a dense depth field $\mathscr{D}_i$ and returns the corresponding $H\times W\times 3$ tensor of 3D points expressed in the camera frame. $\xi^{\mathbf{c}_j}_{\mathbf{c}_i} = \xi^{\mathbf{c}_j}_{\text{ref}}(\xi^{\mathbf{c}_i}_{\text{ref}})^{-1}$ is the euclidean transformation from frame i to j, and $\tau(\cdot,\cdot)$ is a function defined to take a euclidean transformation and a $H\times W\times 3$ tensor of 3D world points, and return the equal-size tensor of transformed 3D world points. Finally, $\mathscr{P}_{ij}$ is the target location of the points in frame i in frame j as hypothesized by the optical flow prediction between frames i and j [40], and $\sum_{ij}$ expresses the uncertainty of these predictions.

2) Scale Constraints: For scale-aware bundle adjustment, please refer to Figure.2. When the scale input is in the form of stereo images, we can input both the left and right camera images into the bundle adjustment layer. At the same time, the extrinsics of the left and right cameras $\xi^{\mathbf{c}'}_{\mathbf{c}}$ are known and fixed. The right camera is marked with a prime, and the objective is given by

$$E = E_r + \sum_{\{i,i'\}\in\mathscr{E}} ||\mathscr{P}_{ii'} - \Pi_{\mathbf{c}}\left(\tau\left(\xi^{\mathbf{c}'}_{\mathbf{c}}, \Pi^{-1}_{\mathbf{c}}(\mathscr{D}_i)\right)\right)||^2_{\sum_{ii'}}. \tag{2}$$

When the scale input is in the form of IMU measurements, we use visual-inertial alignment as introduced in [41] to recover metric scale of visual SLAM estimates and initialize IMU gravity g, velocity v and bias b. After a successful initialization, the vision reference frame can be aligned with the inertial frame, and rescaled based on the identified velocity. The continuous constraint of scale and inertial alignment is then realized by the addition of IMU pre-integration terms. The latter are finally used to form inertial residuals between consecutive keyframes as given by

$$e_{\mathbf{b}_i} = ||\mathrm{R}^{\mathbf{b}_i}_{\text{ref}}(\mathrm{t}^{\text{ref}}_{\mathbf{b}_{i+1}} - \mathrm{t}^{\text{ref}}_{\mathbf{b}_i} + \frac{1}{2}\mathrm{g}\Delta t_i^2 - \mathrm{v}^{\text{ref}}_{\mathbf{b}_i}\Delta t_i) - \hat{\alpha}^{\mathbf{b}_{i+1}}_{\mathbf{b}_i}||^2 \tag{3}$$

$$+ ||\mathrm{R}^{\mathbf{b}_i}_{\text{ref}}(\mathrm{v}^{\text{ref}}_{\mathbf{b}_{k+1}} - \mathrm{g}\Delta t_i) - \hat{\beta}^{\mathbf{b}_{i+1}}_{\mathbf{b}_i}||^2 \tag{4}$$

$$+ ||\log(\mathrm{R}^{\text{ref}^{-1}}_{\mathbf{b}_i}\mathrm{R}^{\text{ref}}_{\mathbf{b}_{i+1}}\hat{\gamma}^{\mathbf{b}_{i+1}}_{\mathbf{b}_i})||^2 \tag{5}$$

$$+ ||\mathrm{b}^a_i - \mathrm{b}^a_{i+1}||^2 + ||\mathrm{b}^g_i - \mathrm{b}^g_{i+1}||^2, \tag{6}$$

where $\hat{\alpha},\hat{\beta},\hat{\gamma}$ are the measured translation, velocity and rotation changes between keyframe IMU states $\mathbf{b}_i$ and $\mathbf{b}_{i+1}$ as given by IMU pre-integration. Δt_i is the time interval. Visual-inertial bundle adjustment differs in that it also optimizes the keyframe IMU states $\mathbf{b}_i$ consisting of translational velocity and gyroscope and accelerate-meter biases. The objective now reads $E = E_r + \sum_{i\in\mathscr{V}} e_{\mathbf{b}_i}$, where $\mathscr{V}$ is the set of keyframes in the bundle adjustment layer.

In the case where the scale input is in the form of a depth map, it is possible to recover the correct scale by directly incorporating a constraint term into the estimated depth map. The final scale-aware objective is given by

$$E = E_r + \sum_{i\in\mathscr{V}} ||\hat{\mathscr{D}}_i - \mathscr{D}_i||^2, \tag{7}$$

where $\hat{\mathscr{D}}_i$ is the depth measurements from the depth camera. Note that the inclusion of direct depth readings is not our main focus, but the objective is used for comparative purposes.

We use Gauss-Newton-style optimization methods to solve these nonlinear optimization objective, and—in analogy to bundle adjustment methods—use the Schur complement trick to accelerate the computation.

C. Single View Object Pose Estimation

1) Instance Segmentation and Object Association: We can use off-the-shelf toolbox like SAM-based models [42], [43] or Mask-RCNN [44] for instance segmentation. To build associations between each object, we then calculate the Intersection over Union (IoU) across all instances of each two frames. If two bounding boxes have the same category and the IoU exceeds a certain threshold, the two object instances are assigned to the same instance ID.

2) Object Pose Predictor: Our object pose predictor is built upon NOCS [1] and aims to produce canonical representations of objects. It processes RGB keyframes and incorporates object masks and class labels from our instance segmentation and tracking module. The output comprises NOCS maps for each object, serving as a crucial input for the subsequent object pose estimation module. The fundamental design of our NOCS predictor is inspired by the original NOCS implementation [1]. However, we have already acquired the necessary masks and class labels, thereby eliminating the need for the Mask R-CNN-like framework [44]. Instead, we leverage an encoder adapted from the U-Net architecture [45] to attain a feature map for each complete, full-resolution input keyframe. This step is followed by using the object masks to crop the ROI features and resize them to a standard 32×32 grid. Subsequently, three separate predictors are engaged to acquire the x, y, and z components of the NOCS output.

3) Robust Geometric Registration for Object Pose: For each masked object, its pose is estimated by registering the NOCS points $X^{nocs}\in\mathbb{R}^{3\times N}$ and the back-projected object depth points $X^{xyz}\in\mathbb{R}^{3\times N}$ using Umeyama's algorithm [15]. Note that the correspondences between X^{nocs} and X^{xyz} are simply given by the question whether or not they originate from the same pixel in the image. In order to deal with inaccurate instance segmentations and noisy NOCS predictions, we apply RANSAC for outlier rejection. The fitted model is the similarity transformation

$$X^{xyz} = s\mathrm{R}X^{nocs} + \mathrm{t}, \tag{8}$$

where $s \in \mathbb{R}$ is the scale of the object, $R \in SO(3)$ is the orientation of the object, and $t \in \mathbb{R}^3$ its position. RANSAC with post-refinement over all inliers hence aims at a robust minimization of the energy objective

$$s^*, R^*, t^* = \arg\min_{s,R,t} \sum_{i}^{N} ||X_i^{xyz} - sRX_i^{nocs} - t||^2. \quad (9)$$

D. Object-Level Pose Graph Optimization

One of our assumptions is that the object is static. Therefore, the camera pose and object pose can be optimized through the proposed object-level pose graph. Our graph contains where K nodes representing object poses and N node representing camera poses. There are two types of edges in our pose graph. The first one is camera-camera edges, which we can obtain from our scale-aware dense monocular SLAM-based relative pose estimations $\xi_{\mathbf{c}_i}^{\mathbf{c}_{i+1}}$. Its corresponding covariance matrix $\Sigma_{\mathbf{c}_i}$ is derived from the bundle adjustment layer. The second type is given by object-camera edges. The relative pose between object node k and camera node i is denoted $\xi_{\mathbf{o}_k}^{\mathbf{c}_i}$, which is the individual object pose obtained from each keyframe. The covariance matrix of the object camera edges can be approximated as

$$\Sigma_{\mathbf{o}_{ki}} = (J_{\mathbf{o}_{ki}}^T \cdot J_{\mathbf{o}_{ki}})^{-1}, \quad (10)$$

where $J_{\mathbf{o}_{ki}}$ is the jacobian matrix of the estimated object pose, defined as

$$J_{\mathbf{o}_{ki}} = \frac{\partial(\xi_{\mathbf{c}_i}^{\mathbf{o}_k} \cdot X_{\mathbf{o}_{ki}}^{nocs} - X_{\mathbf{o}_{ki}}^{xyz})}{\partial \xi_{\mathbf{c}_i}^{\mathbf{o}_k}} |_{\xi_{\mathbf{c}_i}^{\mathbf{o}_k} = \xi_{\mathbf{c}_i}^{\mathbf{o}_k *}}, \quad (11)$$

where $X_{\mathbf{o}_{ki}}^{xyz}$ and $X_{\mathbf{o}_{ki}}^{nocs}$ are back-projected depth points and predicted NOCS points of object $\mathbf{o}_k$ in frame i. Then our object-level pose graph optimization problem can be formulated as:

$$\arg\min_{\{\xi_{\text{ref}}^{\mathbf{o}_k}\},\{\xi_{\text{ref}}^{\mathbf{c}_i}\}} \sum_{k}^{K}\sum_{i}^{N} e_{k,i}^T \cdot \Sigma_{\mathbf{o}_{ki}}^{-1} \cdot e_{k,i} + \sum_{i}^{N-1} e_{i,i+1}^T \cdot \Sigma_{\mathbf{c}_i}^{-1} \cdot e_{i,i+1}, \quad (12)$$

where $e_{k,i} = \log(\xi_{\text{ref}}^{\mathbf{o}_k}(\xi_{\text{ref}}^{\mathbf{c}_i})^{-1}\xi_{\mathbf{o}_k}^{\mathbf{c}_i})$ and $e_{i,i+1} = \log(\xi_{\text{ref}}^{\mathbf{c}_i}(\xi_{\text{ref}}^{\mathbf{c}_{i+1}})^{-1}\xi_{\mathbf{c}_{i+1}}^{\mathbf{c}_i})$ are the residual terms derived from camera-object edge and camera-camera edges, respectively.

We can use robust loss functions like a Huber loss to reduce the influence of outlier object poses. Our implementation is based on [46], which can produce certifiably globally optimal results to our object-level pose graph optimization problem in the presence of outlier object-camera edges.

IV. Experiments

We now proceed to our experimental evaluation. We start by introducing implementation details, and discuss our baseline methods and the benchmarks used for evaluating the 6DoF pose and size estimations. Later, we will introduce details of a self-collected dataset named *MEREAL* (MV-ROPE Extended Real dataset), which is used to benchmark state-of-the-art methods as well as our proposed method with varying qualities of depth sensing. While the detailed quantitative performance of our proposed method is showcased on public datasets, we also provide qualitative and quantitative results on MEREAL, ultimately revealing the superiority of the proposed method.

(a) ZED (b) Azure Kinect (c) Structure Sensor

Fig. 3: Illustration of object point clouds captured by different sensors.

A. Implementation Details

In our scale-aware dense monocular SLAM module, the recurrent optical flow estimator and dense bundle adjustment layer are inspired by the design of DROID-SLAM [38]. Building upon this foundation, we extended the additional inputs to include depth images, stereo images, or IMU measurements, encompassing various forms of scale information. Furthermore, we seamlessly integrated these pieces of information into the bundle adjustment layer. Note that we employ its pre-trained weights of [38] for the optical flow estimation without performing any fine-tuning. The object pose predictor is trained on the NOCS dataset [1]. The dataset consists of two splits: CAMERA and REAL. The CAMERA split is generated by rendering synthetic objects into 300k real-world images. The Real split contains RGB-D image sequences from 31 indoor scenes. Our training strategy aligns with the policies set forth by ShAPO [31]. Initially, the model undergoes training on the CAMERA split, followed by fine-tuning on the REAL training split. To address the challenge of rotational symmetries, the loss function is designed to rotate symmetric objects (bottle, bowl, and can) around the y-axis of the predicted NOCS coordinates. It then selects the smallest loss from rotations along the circle. This ensures a more accurate and reliable model. The encoder is structured as a modified U-Net, where the last two up-sampling blocks are excluded. Instead, we utilize 4×interpolation to regain the full resolution. This modification is crucial to maintain the channel depth, thereby ensuring that the feature encapsulates sufficient information. Following the encoder, each of the three separate predictors is a shallow CNN composed of five convolutional layers. This design choice contributes to the efficiency and compactness of the model. Overall, the object pose predictor is an efficient, lightweight solution and can be trained on a single NVIDIA GeForce RTX 2080 Ti. Additionally, the core module of our framework achieves real-time performance at an interactive level, with a frame rate of about 5 keyframes per second.

B. MEREAL Dataset

This dataset comprises sequences recorded using different depth sensors, and offers depth images and IMU measurements. Our depth sensors include ZED, Azure Kinect,

and Structure Sensor. ZED is a stereo camera that has undergone precise calibration for both intrinsic and extrinsic parameters. It can also output depth maps through stereo reconstruction algorithms. Azure Kinect is a commonly used consumer-grade RGB-D camera that utilizes Time-of-Flight (ToF) technology for depth sensing. It also incorporates a hardware-synchronized IMU sensor. Structure Sensor is a high-precision 3D scanner and the same device used in the NOCS REAL dataset. It is capable of generating high-quality depth maps. Furthermore, Structure Sensor can be used in conjunction with an iPad, allowing for the joint calibration of the depth camera and iPad's color camera to obtain well-aligned RGB-D data. As presented in Figure. 3, we can observe that the Azure Kinect has inaccurate depth, while ZED has substantial noise.

Our dataset comprises three different scenes, each with different objects and object placements. During the recording process, each scene was captured 3 times using each of the different cameras. As a result, we have a total of 27 sequences, with each sequence containing 500-1000 frames. In each scene, we have an additional calibration board placed to acquire ground truth camera poses. Before recording the data, we perform 3D reconstruction for each object in the scene to obtain accurate object models. Regarding the pose annotation of the object, we will manually align the back-projected point cloud of the object to the reconstructed object model for the object pose. Furthermore, we only need to first annotate the object pose in the first frame. Then, we can propagate the object pose from the first frame to all the subsequent frames using the camera motion obtained from the calibration board.

Method	Extra inputs	IoU25	IoU50	IoU75	5°2cm	5°5cm	10°2cm	10°5cm
NOCS[1]	**d**	84.8	78.0	30.1	7.2	10.0	13.8	25.2
DPDN[18]	**d+p**	-	83.4	76.0	46.0	50.7	70.4	78.4
GPV-Pose[32]	**d**	84.2	83.0	64.4	32.0	42.9	-	73.3
VI-Net[13]	**d**	-	-	48.3	**50.0**	**57.6**	70.8	82.1
Query6DoF[27]	**d**	-	82.9	76.0	46.8	54.7	67.9	81.6
HS-Pose[12]	**d**	84.2	82.1	74.7	46.5	55.2	68.6	**82.7**
IST-Net[28]	**d**	84.3	82.5	**76.6**	47.5	53.4	**72.1**	80.5
SOCS[35]	**d**	-	82	75	49	56	72	82
Ours	**s**	**99.9**	**93.6**	60.3	31.4	46.6	46.1	73.7

TABLE I: Quantitative comparisons of mAP on REAL[1]. We marked the extra inputs of each method to distinguish their technical routes. **d** means depth images, **p** means categorical shape priors, and **s** means scale information. For the metrics, IoUx means mAP defined by 3D IoU over a threshold of x%; **m**°n cm represents mAP defined by rotation error less than n° and transformation error less than m cm.

C. Results on NOCS Dataset

Our method uses RGB-D image sequences as an input on this dataset. The CAMERA split of the NOCS dataset [1] consists of single view RGB-D images, only, hence we conducted our experiments on the REAL test split. Regarding baseline methods, we selected state-of-the-art approaches [32], [28], [13], [12], [27], [35], from the three different categories as discussed in the related work section (Sec. II).

In our experimental results, we present the mean Average Precision (mAP) based on the 3D Intersection over Union (3D IoU) as well as translation and rotation errors, as shown in Table I. Figure 4 indicates qualitative results. Figure 5 shows its AP curve.

Fig. 4: Qualitative results of our method obtained on all 6 sequences of the REAL test dataset [1].

On the REAL dataset which contains high-quality depth maps, our method achieves comparable performance to RGB-D methods across various metrics. Specifically, our method shows significant improvements in IoU25 and IoU50 metrics, highlighting the robustness of our approach. The latter stems from three main factors. Firstly, better camera pose and depth estimation can be obtained by utilizing multi-view images. Secondly, by integrating object pose information from multiple frames, we reduce errors and improve the accuracy of pose estimation. Finally, the utilization of multi-view constraints helps mitigate the negative impact of occlusions and false detections that can occur in single view images. This combination of factors contributes to the overall robustness of our method. Furthermore, it should be noted that the reason for the lower IoU75 and rotation accuracy compared to RGB-D based methods is due to the significant intra-class variation in the camera category, where the quality of NOCS maps is often insufficient.

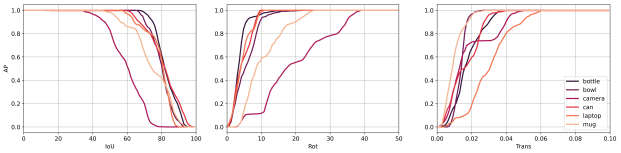

Fig. 5: The AP curve of our approach, with the vertical axis representing AP and the horizontal axis representing IoU, rotation error, and translation error of each category.

Method	mIoU(%)	mE_{rot}(°)	mE_{trans}(cm)
VI-Net[13]	42.3 / 59.0 / 79.2	48.0 / 36.0 / 12.4	36.6 / 15.7 / 8.5
IST-Net[28]	48.3 / 55.7 / **80.8**	54.2 / 49.1 / **10.7**	13.5 / 6.9 / 8.6
Ours	**72.5** / **75.2** / 78.5	**15.2** / **14.4** / 12.9	**8.3** / **3.6** / **6.6**

TABLE II: mIoU, mE_{rot}, mE_{trans} respectively denote the mean IoU, rotation and translation errors. The three numbers within the same cell each represent the results in sequences captured by the ZED, Azure Kinect, and the Structure Sensor.

Fig. 6: Qualitative results on MEREAL. The four rows, respectively, are: VI-Net [13], IST-Net [28], the proposed method, and ground truth. The three columns represent the ZED, Azure Kinect, and Structure Sensor sequence, respectively.

D. Results on MEREAL

On the MEREAL dataset, we adopt mean IoU and mean translation and rotation error as our comparison criteria. Due to the arbitrary definition of ground truth isotropic scale, we do not include the scale factor when calculating 3D IoUs. The benchmark results are given by Table II. Moreover, we tested the performance of the proposed method that uses RGB images and IMU as output on the Azure Kinect sequence. With the use of visual-inertial input, the three metrics of the proposed method are respectively 75.0%, 15.9 degrees, and 6.3 cm, on the same level as that of the RGB-D modality. As the quality of the depth map decreases, the accuracy of single view RGB-D based methods shows a significant decline. However, the proposed method is far less sensitive to this. This occurs because the training depth data of previous methods has been overly idealistic, leading to a significant domain gap with the real world, while our lightweight object pose predictor does not rely on depth and can achieve reliable depth estimation through multi-view images. These factors ultimately enable the proposed method to achieve robust object pose and size estimations.

V. Conclusions

The present work introduces a novel category-level object pose and size estimation framework that distinguishes itself from previous approaches by leveraging a continuous stream of images in conjunction with varying types of scale input. It effectively reduces reliance on accurate but expensive depth sensors, and offers multiple flexible ways for deployment in real-world scenarios. We believe that our method will be of strong interest in applications such as robotic manipulation and virtual reality. Our continued efforts consist of utilizing the geometric information from metric predicted depth to enhance the NOCS prediction, and then intensify the pose and size estimation of objects with large intra-class variations.

VI. Acknowledgment

We would like to acknowledge the funding support provided by project 62250610225 by the Natural Science Foundation of China, as well as projects 22DZ1201900, 22ZR1441300, and dfycbj-1 by the Natural Science Foundation of Shanghai. We would furthermore like to acknowledge the support provided by Midea RoboZone.

References

[1] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2642–2651.

[2] M. Camplani and L. Salgado, "Efficient spatio-temporal hole filling strategy for kinect depth maps," *Proc SPIE*, vol. 8290, pp. 13–, 02 2012.

[3] W. Yin, Y. Liu, and C. Shen, "Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 7282–7295, 2022.

[4] C. Wang, R. Martín-Martín, D. Xu, J. Lv, C. Lu, L. Fei-Fei, S. Savarese, and Y. Zhu, "6-pack: Category-level 6d pose tracker with anchor-based keypoints," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10 059–10 066.

[5] J. Sun, Z. Wang, S. Zhang, X. He, H. Zhao, G. Zhang, and X. Zhou, "Onepose: One-shot object pose estimation without cad models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6825–6834.

[6] B. Wen and K. Bekris, "Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8067–8074.

[7] B. Wen, J. Tremblay, V. Blukis, S. Tyree, T. Müller, A. Evans, D. Fox, J. Kautz, and S. Birchfield, "Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 606–617.

[8] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *Robotics: Science and Systems (RSS)*, 2018.

[9] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3343–3352.

[10] Y. Xu, K.-Y. Lin, G. Zhang, X. Wang, and H. Li, "Rnnpose: Recurrent 6-dof object pose refinement with robust correspondence field estimation and pose optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 880–14 890.

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

[12] L. Zheng, C. Wang, Y. Sun, E. Dasgupta, H. Chen, A. Leonardis, W. Zhang, and H. J. Chang, "Hs-pose: Hybrid scope feature extraction for category-level object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 17 163–17 173.

[13] J. Lin, Z. Wei, Y. Zhang, and K. Jia, "Vi-net: Boosting category-level 6d object pose estimation via learning decoupled rotations on the spherical representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 14 001–14 011.

[14] K. Arun, T. Huang, and S. Blostein, "Least-squares fitting of two 3-d point sets. ieee t pattern anal," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-9, pp. 698 – 700, 10 1987.

[15] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 13, no. 04, pp. 376–380, 1991.

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

[17] H. Lin, Z. Liu, C. Cheang, Y. Fu, G. Guo, and X. Xue, "Sar-net: Shape alignment and recovery network for category-level 6d object pose and size estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6707–6717.

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

[19] X. Deng, J. Geng, T. Bretl, Y. Xiang, and D. Fox, "icaps: Iterative category-level object pose and shape estimation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1784–1791, 2022.

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

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

[22] J. Wang, K. Chen, and Q. Dou, "Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4807–4814.

[23] Q. Meng, J. Gu, S. Zhu, J. Liao, T. Jin, F. Guo, W. Wang, and W. Song, "Kgnet: Knowledge-guided networks for category-level 6d object pose and size estimation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 6102–6108.

[24] L. Zhou, Z. Liu, R. Gan, H. Wang, and M. H. Ang Jr, "Dr-pose: A two-stage deformation-and-registration pipeline for category-level 6d object pose estimation," *arXiv preprint arXiv:2309.01925*, 2023.

[25] P. Liu, Q. Zhang, and J. Cheng, "Gsnet: Model reconstruction network for category-level 6d object pose and size estimation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2898–2904.

[26] H. Wang, W. Li, J. Kim, and Q. Wang, "Attention-guided rgb-d fusion network for category-level 6d object pose estimation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 10 651–10 658.

[27] R. Wang, X. Wang, T. Li, R. Yang, M. Wan, and W. Liu, "Query6dof: Learning sparse queries as implicit shape prior for category-level 6dof pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 14 055–14 064.

[28] J. Liu, Y. Chen, X. Ye, and X. Qi, "Ist-net: Prior-free category-level pose estimation with implicit space transformation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 13 978–13 988.

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

[30] M. Z. Irshad, T. Kollar, M. Laskey, K. Stone, and Z. Kira, "Centersnap: Single-shot multi-object 3d shape reconstruction and categorical 6d pose and size estimation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 10 632–10 640.

[31] M. Z. Irshad, S. Zakharov, R. Ambrus, T. Kollar, Z. Kira, and A. Gaidon, "Shapo: Implicit representations for multi-object shape, appearance, and pose optimization," in *European Conference on Computer Vision*. Springer, 2022, pp. 275–292.

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

[33] L. Zou, Z. Huang, N. Gu, and G. Wang, "6d-vit: Category-level 6d object pose estimation via transformer-based instance representation learning," *IEEE Transactions on Image Processing*, vol. 31, pp. 6907–6921, 2022.

[34] D. Chen, J. Li, Z. Wang, and K. Xu, "Learning canonical shape space for category-level 6d object pose and size estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[35] B. Wan, Y. Shi, and K. Xu, "Socs: Semantically-aware object coordinate space for category-level 6d object pose estimation under large shape variations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 14 065–14 074.

[36] 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.

[37] T. Lee, B.-U. Lee, M. Kim, and I. S. Kweon, "Category-level metric scale object shape and pose estimation," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8575–8582, 2021.

[38] Z. Teed and J. Deng, "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras," *Advances in neural information processing systems*, vol. 34, pp. 16 558–16 569, 2021.

[39] Z. Min and E. Dunn, "Voldor+ slam: For the times when feature-based or direct methods are not good enough," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 813–13 819.

[40] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 402–419.

[41] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[42] Y. Cheng, L. Li, Y. Xu, X. Li, Z. Yang, W. Wang, and Y. Yang, "Segment and track anything," *arXiv preprint arXiv:2305.06558*, 2023.

[43] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[44] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[46] D. M. Rosen, L. Carlone, A. S. Bandeira, and J. J. Leonard, "Se-sync: A certifiably correct algorithm for synchronization over the special euclidean group," *The International Journal of Robotics Research*, vol. 38, no. 2-3, pp. 95–125, 2019.