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

Active Pose Refinement for Textureless Shiny Objects using the Structured Light Camera

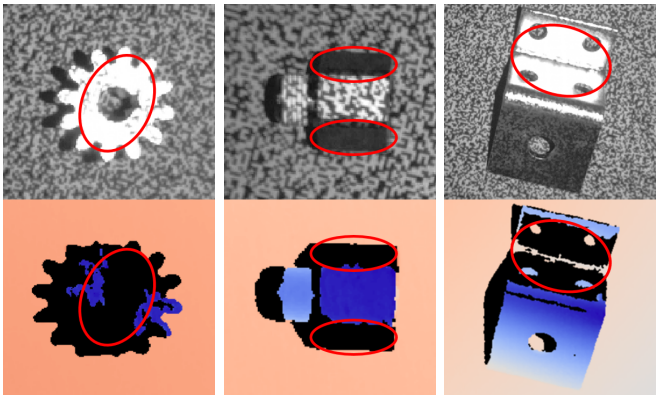
Jun Yang*, Jian Yao†, and Steven L. Waslander*

***Abstract*— 6D pose estimation of textureless shiny objects has become an essential problem in many robotic applications. Many pose estimators require high-quality depth data, often measured by structured light cameras. However, when objects have shiny surfaces (e.g., metal parts), these cameras fail to sense complete depths from a single viewpoint due to the specular reflection, resulting in a significant drop in the final pose accuracy. To mitigate this issue, we present a complete active vision framework for 6D object pose refinement and next-best-view prediction. Specifically, we first develop an optimization-based pose refinement module for the structured light camera. Our system then selects the next best camera viewpoint to collect depth measurements by minimizing the predicted uncertainty of the object pose. Compared to previous approaches, we additionally predict measurement uncertainties of future viewpoints by online rendering, which significantly improves the next-best-view prediction performance. We test our method on the real-world ROBI dataset. The results show that our pose refinement module outperforms the traditional ICP-based approach when given the same input depth data, and our next-best-view strategy can achieve high object pose accuracy with significantly fewer viewpoints than the heuristic-based policies.**

I. INTRODUCTION

Textureless shiny objects, such as metal parts, are essential components of many products. Detecting and estimating the 6-DoF (Degree of Freedom) poses of these objects is crucial in many applications, such as robot manipulation [1], [2]. With the explosive growth of deep learning, many RGB-based solutions have been developed to estimating 6D object poses [3], [4], [5]. While these approaches show high detection rates in 2D space, the actual 6D pose accuracy is often low due to scale and perspective ambiguities in RGB images. This limitation restricts the use of object poses in tasks like robotic bin-picking, which require highly accurate 6D object poses (e.g., 5-mm and 5-degree accuracy) for grasping and assembly [6]. Hence, depth data is often necessary to refine the object pose accuracy in many object pose estimation systems.

To acquire reliable depth maps, the structured light illumination (SLI) camera is usually used because of its high accuracy and resolution. It projects light patterns onto objects to simplify the stereo-matching problem and excels on diffuse surfaces. However, when imaging shiny objects, the SLI camera produces depth maps with low accuracy and missing data. Due to specular reflection, a large amount of incident light is reflected, either directly back to the camera (image saturation), completely missing the camera (low SNR), or reflected within the object surfaces before reaching the camera (inter-reflection). As illustrated in Figure 1, each effect can result in inaccurate or missing depth measurements. To address this issue, our recent work fuses multi-view depth maps to improve scene completion [7]. Remaining challenges include selecting camera viewpoints to maximize information gain and using multi-view depth for object pose estimation, crucial for fast scene understanding.



(a) Image Saturation (b) Low SNR (c) Inter-reflection

Fig. 1: Missing depth measurements on shiny objects' surfaces using a structured light camera.

Some approaches have been proposed to predict the next-best-view (NBV) to complete the depth data on the target objects and estimate object poses [8], [9]. These studies assume that complete depth data is necessary for the optimal object pose estimation, and aim to find camera viewpoints that can recover as much depth data as possible on all objects. However, this strategy is usually inefficient since, for the object pose estimation, depth from some areas of the scene is far more important than others (e.g., areas with lower measurement uncertainties and better constraints for object pose refinement).

The above observations motivated us to introduce a tightly coupled framework of 6D pose refinement and next-best-view prediction for textureless shiny objects. Inspired by [5], [10], we first develop an optimization approach based on signed distance functions (SDF) to refine object poses. In addition, to mitigate the impact of depth errors, we estimate depth uncertainties from the SLI camera and integrate them into our pose refinement module. Given the initial object pose, we iteratively refine it and determine the next best camera viewpoint to enhance

This work was supported by Epson Canada Ltd.

*Jun Yang and Steven L. Waslander are with University of Toronto Institute for Aerospace Studies and Robotics Institute. {jun.yang, steven.waslander}@robotics.utias.utoronto.ca

†Jian Yao is with Epson Canada jian.yao@ea.epson.com

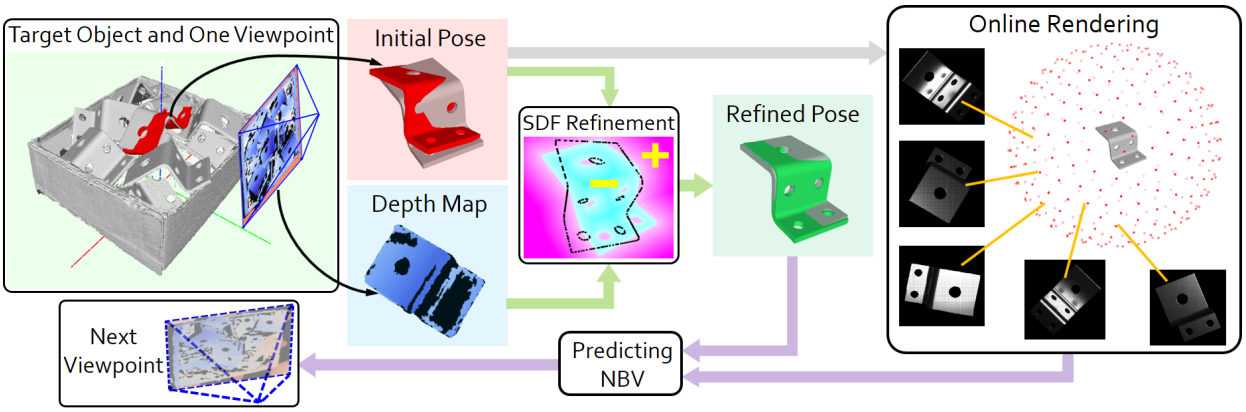


Fig. 2: An overview of the proposed multi-view pose refinement and the next-best-view prediction for the shiny objects.

pose accuracy. For NBV prediction, our proposed method includes two main parts: a) a surface reflection model to predict depth uncertainties, b) the NBV prediction for object pose refinement by incorporating the reflection model. In the first part, we estimate the object's surface reflection parameters by differentiable rendering techniques. The estimated parameters are then used by online rendering to predict object's depth uncertainties from a future viewpoint. In the second part, we integrate our reflection model into an information-theoretic NBV policy. For each candidate viewpoint, we predict the expected uncertainty of the object pose and determine the NBV by minimizing the predicted uncertainties. Figure 2 shows an overview of the framework.

We perform the evaluation on the real-world ROBI dataset [11]. We first evaluate our pose refinement approach, demonstrating its superiority over the widely used iterative closest point (ICP) method when given same input depth measurements. To demonstrate the advantages of our NBV approach, we compare it with two heuristic-based strategies. The results indicate our method can achieve high pose refinement accuracy using significantly fewer viewpoints. In summary, our key contributions are:

- A 6D pose refinement approach for textureless shiny objects designed for SLI cameras. Our approach comprises (a) the estimation of pixel depth uncertainties, (b) the integration of uncertainty estimates within our SDF-based object pose refinement module.
- A surface reflection model to predict the object's depth uncertainties for unseen camera viewpoints. Our reflection model recovers the object's reflection parameters with the differentiable renderer.
- An active vision system that integrates our pose refinement approach and reflection model via the online rendering to predict the NBV for pose estimation.

II. RELATED WORK

A. Object Pose Refinement

To acquire highly accurate object poses, pose refinement is a critical step and has been mostly addressed using depth data. Iterative Closest Point (ICP) is the most classical approach and has been used in many object pose estimation pipelines [3], [4]. Given the initial object pose, ICP refines it iteratively by establishing the point-to-point correspondences from the point cloud to the object model and minimizing the distances. To improve the runtime performance, several approaches [5], [10], [12] have been proposed to reduce the computation cost. Among them, Deng et al. [5] avoid the costly point correspondence building and refine the object pose by aligning the 3D point cloud with the SDF of the target object. Learning based approaches, such as DenseFusion [2], improve pose refinement by replacing ICP with a neural network. DenseFusion fuses the RGB and depth features and trains a refinement network to iteratively regress a pose offset.

B. Depth Acquisition with Structured Light Camera

Structured Light Illumination (SLI) cameras are one of the most used indoor 3D sensors, but they produce inaccurate and missing depth measurements when target objects have shiny surfaces. To overcome the image saturation problem (Figure 1a), high dynamic range (HDR)-based methods are widely used in many SLI systems [13]. HDR methods fuse a set of images under multiple exposures into a single image for stereo matching. To reduce HDR's time cost, Liu et al. [14] employed a neural network to directly enhance single exposure-captured images. Despite good performance, these methods cannot solve the low SNR and inter-reflection problems (illustrated in Figure 1b and 1c) from a single viewpoint. In comparison, when the setup permits, multi-view acquisition [7], [9] can provide a high level of depth completion.

C. Active Vision

Active vision refers to actively manipulating the camera viewpoint to obtain the maximum information for different tasks. Active vision has received a lot of attention from the robotics community and has been employed in many applications, such as robot manipulation [8], reconstruction [9], [15] and SLAM [16], [17]. Recent studies show that active vision can be achieved by maximizing the Fisher information of the robot state [15], [16], [17]. For example, the authors in [16], [17] use the Fisher information to find highly-informative trajectories and achieve high localization accuracy.

III. MULTI-VIEW OBJECT POSE REFINEMENT

This section presents our multi-view 6D pose refinement formulation for shiny objects with the SLI camera. Given the 3D object model and multi-view acquired depth maps, we aim to refine the rigid pose $\boldsymbol{T}_{ow} \in SE(3)$ from a global (world) coordinate W to the object coordinate O. We assume that we know the camera poses $\boldsymbol{T}_{wc} \in SE(3)$ relative to the world coordinate. Our pose refinement module consists of two parts: (a) depth uncertainty estimation from the SLI camera, (b) an optimization-based object pose refinement module that takes the uncertainty estimates into account. In the following subsections, we describe these two parts in detail.

A. Estimating Measurement Uncertainty

For an SLI camera, the depth measurement uncertainty is a function of the depth, camera parameters (e.g., intrinsics), and the photometric appearance of the projected light patterns. In this section, we describe how to compute the depth uncertainty, starting from estimating the uncertainty for the disparity, $\boldsymbol{\sigma}_d^2$, and propagating it through a non-linear model to obtain the depth uncertainty, $\boldsymbol{\sigma}_z^2$.

Depending on the hardware design of an SLI camera, the stereo matching is performed from camera-to-camera or camera-to-projector. We use the camera-to-camera in our derivation, which can be easily adapted to the camera-to-projector design. Given the stereo pair of left $\boldsymbol{I}_L$ and right $\boldsymbol{I}_R$ pattern projected images, the disparity uncertainty, $\boldsymbol{\sigma}_d^2$, accounts for the appearance ambiguities between image patches. Intuitively, matching is reliable for image patches with strong image intensity gradients. When the dominant gradient direction is parallel to the epipolar line (x axis for the left-right setup), the obtained disparity becomes more reliable, which results in lower uncertainty and vice versa. Inspired by [15], we use the sum of squared differences (SSD) to quantify the disparity and define the photometric error as:

$$e = \boldsymbol{I}_L(u_L, v) - \boldsymbol{I}_R(u_R, v) = \boldsymbol{I}_L(u_L, v) - \boldsymbol{I}_R(u_L - d, v) \tag{1}$$

where d is the acquired disparity for the pixel $\mathbf{u} = [u, v]$ in the left image $\boldsymbol{I}_L$. We assume the disparity of a pixel is normally distributed and compute its variance, $\boldsymbol{\sigma}_d^2$, through the Fisher information:

$$\boldsymbol{\sigma}_d^2 = \boldsymbol{\sigma}_I^2 \left(\mathbf{J}_d^T \mathbf{J}_d\right)^{-1} \tag{2}$$

where $\boldsymbol{\sigma}_I^2$ denotes the variance of the image noise and Jacobian $\mathbf{J}_d$ is derived by:

$$\mathbf{J}_d = \frac{\partial e}{\partial d} = -\frac{\partial \boldsymbol{I}_R}{\partial u_R}\frac{\partial u_R}{\partial d} = \frac{\partial \boldsymbol{I}_R}{\partial u_R} \tag{3}$$

which is the image gradient along the x-axis over a patch from image $\boldsymbol{I}_R$, centered at the pixel $\boldsymbol{u}_R = [u_R, v]$. To obtain the measurement variance of the depth, $\boldsymbol{\sigma}_z^2$, we propagate the disparity variance, $\boldsymbol{\sigma}_d^2$, through:

$$\boldsymbol{\sigma}_z^2 = \boldsymbol{F}\,\boldsymbol{\sigma}_d^2\,\boldsymbol{F}^T \tag{4}$$

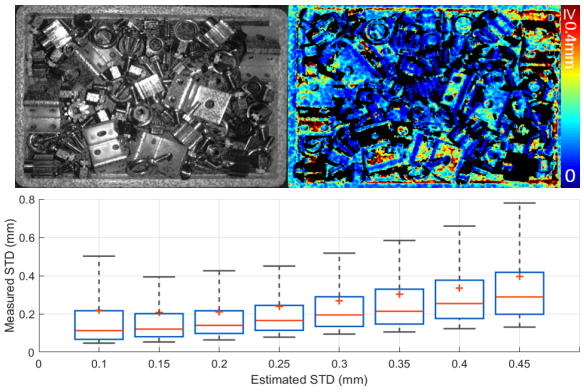


Fig. 3: Upper: The pattern projected image and the estimated uncertainties on the depth map. Lower: The correlation between estimated and measured uncertainty.

where $\boldsymbol{F}$ is the Jacobian of depth, $\boldsymbol{z}$, with respect to disparity, $\boldsymbol{d}$. For the camera-to-camera setup, the acquired depth will have high uncertainty when the image gradient is weak in the left or right image. Hence, we need to compute $\boldsymbol{\sigma}_{z,L}^2$ and $\boldsymbol{\sigma}_{z,R}^2$ for $\boldsymbol{I}_L$ and $\boldsymbol{I}_R$, respectively, and final depth uncertainty is obtained by:

$$\boldsymbol{\sigma}_z^2 = \max\left(\boldsymbol{\sigma}_{z,L}^2, \boldsymbol{\sigma}_{z,R}^2\right) \tag{5}$$

We demonstrate our estimated uncertainty measure in Figure 3. We can see that the estimated uncertainty accurately correlates with the actual measured depth variance.

B. Pose Refinement With SDFs

We formulate the object pose refinement as an optimization problem and solve it iteratively. Our refinement process is illustrated in Figure 4. Based on a signed distance function (SDF) approach [5], [10], we refine the object pose $\boldsymbol{T}_{ow}$ by matching the depth measurement, z, which is defined for each pixel $\mathbf{u} = [u, v]$, against the **SDF** of the target object model.

Given the depth map $\boldsymbol{Z}_k(\boldsymbol{u})$ from viewpoint k, we first extract the object mask, $\boldsymbol{M}_k$, and obtain the object's depth measurements. We utilize an instance segmentation network from [18] which provides pixel-level instance predictions. By back-projecting the pixels in $\boldsymbol{M}_k$, we obtain the point cloud of the object, $\boldsymbol{P}_{c,k} \in \mathbb{R}^3$, defined in the k^{th} camera coordinate as:

$$\boldsymbol{P}_{c,k} = \left\{\boldsymbol{Z}_k(\boldsymbol{u})\,\boldsymbol{K}^{-1}\,[\boldsymbol{u}, 1]^T\,,\, \boldsymbol{u} \in \boldsymbol{M}_k\right\} \tag{6}$$

where $\boldsymbol{K}$ represents the camera intrinsic matrix. We transform the point cloud, $\boldsymbol{P}_{c,k}$, to the world coordinate frame, W, with the known camera pose, $\boldsymbol{T}_{wc,k}$:

$$\boldsymbol{P}_w = \left\{\boldsymbol{T}_{wc,k}\,\boldsymbol{P}_{c,k}\,,\, k = 1:K\right\} \tag{7}$$

where $\boldsymbol{P}_w$ is the point cloud defined in the world frame. We optimize the object pose $\boldsymbol{T}_{ow}$ by matching the 3D points against the **SDF** of the target object model:

$$\boldsymbol{T}_{ow}^* = \operatorname{argmin} \sum_{\boldsymbol{p}_{w,i} \in \boldsymbol{P}_w} \left\|\mathbf{SDF}\left(\boldsymbol{T}_{ow}\,\boldsymbol{p}_{w,i}\right)\right\|^2 \tag{8}$$

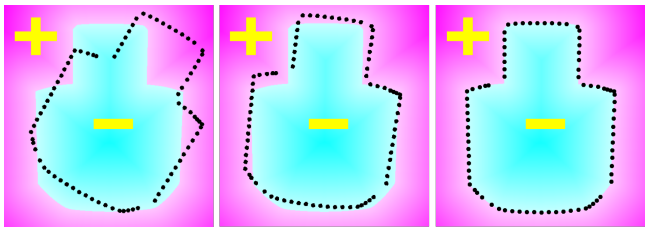


Fig. 4: Object pose refinement module. Pink and blue represents the space with positive and negative distance, respectively. Black dots are the transformed point cloud. From left to right: the object pose, $\boldsymbol{T}_{ow}$, is refined iteratively by minimizing the SDF loss (Equation 8).

where $\boldsymbol{p}_{w,i}$ is a 3D point in the point cloud $\boldsymbol{P}_w$. The function $\mathbf{SDF}\left(\boldsymbol{T}_{ow}\,\boldsymbol{p}_{w,i}\right)$ denotes the signed distance value by transforming the 3D point $\boldsymbol{p}_{w,i}$ from the world frame to the object model frame with a pose estimate $\boldsymbol{T}_{ow}$.

To estimate the pose, $\boldsymbol{T}_{ow}$, from measurements $\boldsymbol{P}_w$, we formulate the problem as a nonlinear least squares (NLLS) problem. We solve this problem with the Gauss-Newton algorithm and integrate the **SDF** measurement uncertainties, $\boldsymbol{\Sigma}_{\mathbf{sdf}}$, in each iterative step:

$$\left(\mathbf{J}_{\xi_{ow}}^{T}\,\boldsymbol{\Sigma}_{\mathbf{sdf}}^{-1}\,\mathbf{J}_{\xi_{ow}}\right)\delta_{\xi_{ow}} = \mathbf{J}_{\xi_{ow}}^{T}\,\boldsymbol{\Sigma}_{\mathbf{sdf}}^{-1}\,\mathbf{SDF} \tag{9}$$

where $\mathbf{J}_{\xi_{ow}}$ is the stacked Jacobian matrix of **SDF**:

$$\mathbf{J}_{\xi_{ow}} = \frac{\partial \mathbf{SDF}}{\partial \boldsymbol{\xi}_{ow}} = \frac{\partial \mathbf{SDF}}{\partial \boldsymbol{P}_o}\,\frac{\partial \boldsymbol{P}_o}{\partial \boldsymbol{\xi}_{ow}} \tag{10}$$

where $\boldsymbol{\xi}_{ow} \in \mathfrak{se}(3)$ is the Lie algebra representation of the transformation $\mathbf{T}_{ow}$, and $\boldsymbol{P}_o$ is the point cloud, transformed to the object frame. We acquire the uncertainty, $\boldsymbol{\Sigma}_{\mathbf{sdf}}$, by propagating the depth uncertainty (obtained from Section III-A), $\boldsymbol{\sigma}_z^2$, through a nonlinear model:

$$\boldsymbol{\Sigma}_{\mathbf{sdf}} = \boldsymbol{G}\,\boldsymbol{\sigma}_z^2\,\boldsymbol{G}^T \tag{11}$$

where $\boldsymbol{G}$ is the Jacobian of **SDF** value with respect to the depth measurement $\boldsymbol{z}$.

IV. Predicting Depth Uncertainty

The object pose refinement performance relies heavily on the input depth measurements from different viewpoints. For an SLI camera, to find the optimal viewpoint, it is important to quantify the depth uncertainty for future viewpoints. In this section, we detail how to predict the depth uncertainty by the rendering technique. The predicted uncertainties will be used to find the next-best-view for the object pose refinement (Section V).

A. Image Acquisition Process

The depth acquisition of an SLI camera is influenced by the light sources, camera viewpoint, and object characteristics (e.g., surface materials). Figure 5 illustrates the image acquisition process of the SLI camera. While the details of this process are given in [9], we include a brief description in this paper for completeness.

Typically, two light sources need to be considered: the ambient light, $\boldsymbol{L_a}$, and the projector light, $\boldsymbol{L_p}$. Since the projector light is the primary light source and the ambient light is negligible in comparison, we define the total light source $\boldsymbol{L_{total}}$ as:

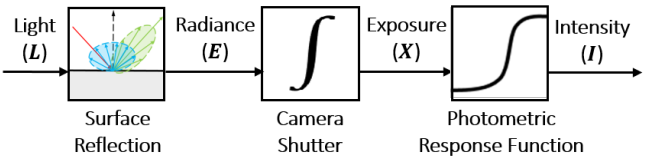


Fig. 5: Image acquisition process of the SLI camera.

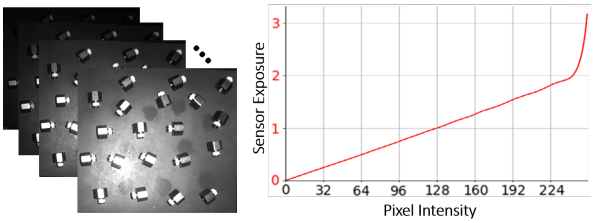


Fig. 6: Left: Input images for calibrating the photometric response function. Right: The recovered function.

$$\boldsymbol{L_{total}} = \boldsymbol{L_p} + \boldsymbol{L_a} \approx \boldsymbol{L_p} \tag{12}$$

Given the light source and other scene parameters (e.g., object poses, materials), the reflection function, $f(\cdot)$, returns the radiance, $\boldsymbol{E}$, which is the amount of light that reflects into the camera lens per time unit. We recover the reflection function using a differentiable rendering algorithm, as detailed in Section IV-B. The sensor exposure, $\boldsymbol{X}$, then integrates the received radiance, $\boldsymbol{E}$, within the camera exposure time, $\boldsymbol{\Delta t}$, via the camera shutter. The photometric response function, $g(\boldsymbol{X})$, finally maps the exposure $\boldsymbol{X}$ to the pixel intensity $\boldsymbol{I}$ in the pattern projected image:

$$\boldsymbol{I} = g(\boldsymbol{X}) = g(\boldsymbol{E}\boldsymbol{\Delta t}) \tag{13}$$

We obtain the function $g(\boldsymbol{X})$ and its inverse, $g^{-1}(\boldsymbol{I})$, using a photometric calibration approach, presented in [19]. The input to the calibration process is a number of images taken from a static scene with different known exposures, $\boldsymbol{\Delta t}$. A white pattern is projected onto the scene during the capture. An example of input images and the recovered photometric response function, $g(\boldsymbol{X})$, is shown in Figure 6.

B. Recovering Reflection Function

The reflection function, $f(\cdot)$, describes how light interacts with surfaces in the scene. As illustrated in Figure 5, it takes the physical attributes of a scene (e.g., lighting source, objects' poses, and materials) and outputs the radiance, $\boldsymbol{E}$. We implement the reflection function, $f(\cdot)$, as a rendering process, and solve this inverse problem using the differentiable rendering technique. The reflection function is differentiable. Its derivative $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ provides a first-order approximation of how a desired output $\mathbf{y}$ (rendered radiance) can be achieved by optimizing the inputs $\mathbf{x}$ (scene parameters). The differentiable loss function, $l(\mathbf{y})$, is used to quantify the rendering output $\mathbf{y}$. As illustrated in Figure 7a, scene parameters (e.g., object materials) can be estimated by minimizing the loss function.

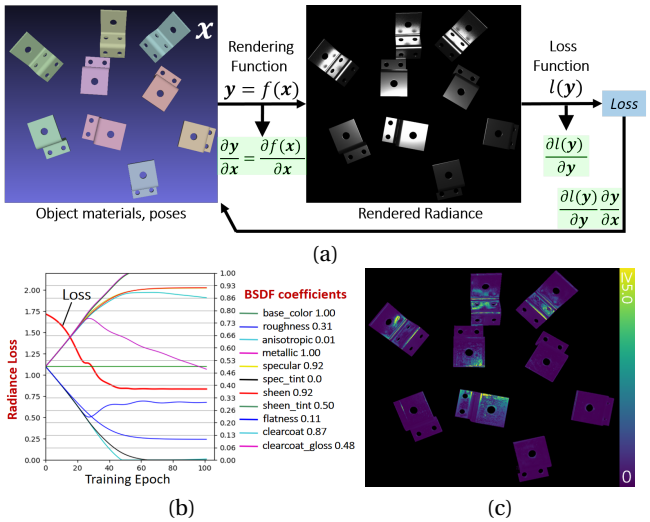


Fig. 7: (a) Estimating the scene parameters (object BSDF coefficients) with differentiable rendering [20]. (b) Loss curve and estimated object BSDF coefficients over the differentiable rendering epochs. (c) Error map between target and estimated radiance when the training converged.

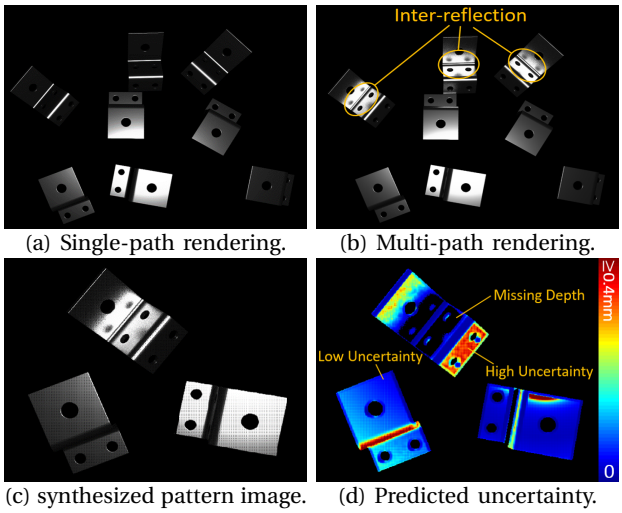


Fig. 8: (a)&(b) Rendering with single- and multi-path ray-tracing, demonstrating the presence of inter-reflection. (c)&(d) The synthesized pattern projected image of the objects from a future viewpoint and the corresponding predicted depth uncertainty map.

For each object, we estimate its materials with the known projector light, $\boldsymbol{L_p}$, radiance map, $\mathbf{E}_{obj}$ and ground truth 6D object poses, $\mathbf{T}_{c2o}$. The radiance map and object poses are obtained by capturing a static scene of the target objects. We calculate the radiance for each pixel on the object's surface using the photometric response function (Section IV-A). To acquire object poses, we capture a depth map of the scene and manually label the 6D pose for each object in the camera coordinate.

We assume the projector light source, $\boldsymbol{L_p}$, is a point light emitter, which radiates the uniform illumination to all directions. For the surface reflection, we use the principled BSDF (bidirectional scattering distribution function) [21] as the surface reflection model. We estimate the BSDF coefficients with the differentiable rendering technique. The optimization problem can be solved with gradient-based methods iteratively. In our approach, we implement the differentiable rendering using the Mitsuba 3 library [20]. All parameters are initialized to the medium value and optimized with the L2 loss and the Adam optimizer [22]. Figure 7b illustrates the loss curve and estimated BSDF coefficients of a textureless shiny object. The corresponding error map between the target and estimated radiance map is shown in Figure 7c.

C. Predicting Measurement Uncertainty

For an object, we predict its depth uncertainty, $\check{\boldsymbol{\sigma}}_z^2$, from a future camera viewpoint, $\mathbf{T}_{c2w}$, using the forward rendering process. With the recovered reflection function, $f(\cdot)$, and photometric response function, $g(\cdot)$, we can generate a white pattern projected image, $\boldsymbol{I}_w$, of the target object. An object is defined with its CAD model and a 6D pose hypothesis, $\check{\mathbf{T}}_{w2o}$, defined in the world coordinate:

$$\boldsymbol{I}_w = g(\boldsymbol{E}\boldsymbol{\Delta t}) \tag{14}$$

$$\boldsymbol{E} = f\left(\boldsymbol{L_p}, \check{\mathbf{T}}_{c2o}\right) = f\left(\boldsymbol{L_p}, \mathbf{T}_{c2w}\check{\mathbf{T}}_{w2o}\right) \tag{15}$$

where $\check{\mathbf{T}}_{c2o}$ is the object pose hypothesis in the camera frame. To predict the missing depth measurement caused by the inter-reflection problem (shown in Figure 1), we render the object with both multi- and single-path ray tracing. As illustrated in Figure 8a, the rendered image with single-path, $\mathbf{I}_{single}$, only contains direct reflections, which serves as the signal portion to acquire the depth for an SLI camera. The multi-path rendered image (Figure 8b), $\mathbf{I}_{multi}$, contains both direct and inter-reflections. We treat a pixel depth as missing if the intensity ratio between $\mathbf{I}_{single}$ and $\mathbf{I}_{multi}$ is smaller than a threshold τ_I:

$$\left\{\boldsymbol{z} = \emptyset \;\middle|\; \forall \frac{\mathbf{I}_{single}}{\mathbf{I}_{multi}} < \tau_I\right\} \tag{16}$$

To predict the depth uncertainty from the SLI camera, we synthesize a random pattern projected image and compute the uncertainty, $\check{\boldsymbol{\sigma}}_z^2$, using Equations (1)-(5). We synthesize the random pattern image by combining two multi-path rendered white pattern images with two different lighting intensities (one strong and one weak). Figure 8c-8d show an example of our synthesized pattern image and the predicted depth uncertainties. A pixel depth is considered missing when the predicted uncertainty is larger than a pre-defined threshold τ_σ:

$$\left\{\boldsymbol{z} = \emptyset \;\middle|\; \forall \check{\boldsymbol{\sigma}}_z > \tau_\sigma\right\} \tag{17}$$

Note that, for a candidate viewpoint, a pixel depth measurement is considered to be missing if any condition from Equations (16)-(17) is fulfilled.

V. Active Pose Refinement with Next-Best-View

In Section III, we formulate the multi-view object pose refinement problem and solve it using an iterative approach. However, collecting many viewpoints is usually

not practical. Hence, in this section, we present an active vision approach for object pose refinement. We developed our NBV policy based on the Fisher information. Compared to most previous NBV approaches [16], [17], which neglect the measurement uncertainty, we exploit our predicted depth uncertainties (Section IV-C) when computing the Fisher information. For each iteration, we estimate the uncertainty of the object pose and find the NBV, which minimizes the predicted uncertainty.

We assume the initial object pose is obtained (e.g., from an external pose estimator) and refine the pose by optimizing the Equation (8) with the Jacobian, $\mathbf{J}_{\boldsymbol{\xi}}$, and measurement uncertainty, $\boldsymbol{\Sigma}_{\mathbf{sdf}}$. We compute the covariance of the refined object pose, $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$, through a first-order approximation of the Fisher information matrix (FIM):

$$\boldsymbol{\Sigma}_{\boldsymbol{\xi},\mathbf{Z}_{1:K}} = \left(\mathbf{J}^T_{\boldsymbol{\xi},\mathbf{Z}_{1:K}} \boldsymbol{\Sigma}^{-1}_{\mathbf{sdf},\mathbf{Z}_{1:K}} \mathbf{J}_{\boldsymbol{\xi},\mathbf{Z}_{1:K}}\right)^{-1} \tag{18}$$

where $\mathbf{Z}_{1:K}$ denote the collected depth measurement sets from K viewpoints. The stacked Jacobian, $\mathbf{J}_{\boldsymbol{\xi},\mathbf{Z}_{1:K}}$, and measurement uncertainties, $\boldsymbol{\Sigma}_{\mathbf{sdf},\mathbf{Z}_{1:K}}$, are represented by:

$$\mathbf{J}_{\boldsymbol{\xi},\mathbf{Z}_{1:K}} = \begin{bmatrix} \mathbf{J}_{\boldsymbol{\xi},\boldsymbol{Z}_1} \\ \vdots \\ \mathbf{J}_{\boldsymbol{\xi},\boldsymbol{Z}_K} \end{bmatrix}, \boldsymbol{\Sigma}_{\mathbf{sdf},\mathbf{Z}_{1:K}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{sdf},\mathbf{Z}_1} & & \\ & \ddots & \\ & & \boldsymbol{\Sigma}_{\mathbf{sdf},\mathbf{Z}_K} \end{bmatrix} \tag{19}$$

The row-blocks, $\mathbf{J}_{\boldsymbol{\xi},\boldsymbol{Z}_k}$ and $\boldsymbol{\Sigma}_{\mathbf{sdf},\mathbf{Z}_k}$, correspond to the Jacobian matrix and **SDF** uncertainty with the k^{th} viewpoint, and can be calculated using Equation (10) and (11), respectively. To compute the uncertainty for the object pose covariance, we use the differential entropy, $h_e\left(\boldsymbol{\Sigma}_{\boldsymbol{\xi},\mathbf{Z}_{1:K}}\right)$:

$$h_e\left(\boldsymbol{\Sigma}_{\boldsymbol{\xi},\mathbf{Z}_{1:K}}\right) = \frac{1}{2}\ln\left((2\pi e)^n \left|\boldsymbol{\Sigma}_{\boldsymbol{\xi},\boldsymbol{Z}_{1:K}}\right|\right) \tag{20}$$

where $h_e\left(\boldsymbol{\Sigma}_{\boldsymbol{\xi},\mathbf{Z}_{1:K}}\right)$ is expressed in nats.

To increase the object pose accuracy, we aim to find the next best camera viewpoint $\boldsymbol{v}^*$ from a set of candidate viewpoints $\{\boldsymbol{V}\}$ which will minimize the entropy of the object pose, $h_e\left(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\right)$. Suppose we have collected the depth measurement sets, $\mathbf{Z}_{1:K}$, from K viewpoints. For a future camera viewpoint, $\widehat{\boldsymbol{v}}$, the stacked Jacobian and measurement uncertainties have the following form:

$$\mathbf{J}_{\boldsymbol{\xi},\overline{\boldsymbol{Z}}} = \begin{bmatrix} \mathbf{J}_{\boldsymbol{\xi},\boldsymbol{Z}_{1:K}} \\ \mathbf{J}_{\boldsymbol{\xi},\widehat{\boldsymbol{Z}}} \end{bmatrix}, \boldsymbol{\Sigma}_{\mathbf{sdf},\overline{\boldsymbol{Z}}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{sdf},\mathbf{Z}_{1:K}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{sdf},\widehat{\boldsymbol{Z}}} \end{bmatrix} \tag{21}$$

where $\overline{\boldsymbol{Z}} = \{\boldsymbol{Z}_{1:K}, \widehat{\boldsymbol{Z}}\}$ includes acquired measurement sets $\boldsymbol{Z}_{1:K}$ from viewpoints $\boldsymbol{v}_{1:K}$ and predicted measurement set $\widehat{\boldsymbol{Z}}$ for the future viewpoint, $\widehat{\boldsymbol{v}}$. With the FIM evaluation, we can predict the object pose covariance by:

$$\boldsymbol{\Sigma}_{\boldsymbol{\xi},\overline{\boldsymbol{Z}}} = \left(\mathbf{J}^T_{\boldsymbol{\xi},\overline{\boldsymbol{Z}}} \boldsymbol{\Sigma}^{-1}_{\mathbf{sdf},\overline{\boldsymbol{Z}}} \mathbf{J}_{\boldsymbol{\xi},\overline{\boldsymbol{Z}}}\right)^{-1} \tag{22}$$

Note that, in Equation (21), we compute the Jacobian $\mathbf{J}_{\boldsymbol{\xi},\widehat{\boldsymbol{Z}}}$ and uncertainty $\boldsymbol{\Sigma}_{\mathbf{sdf},\widehat{\boldsymbol{Z}}}$ before actually moving to the camera viewpoint $\widehat{\boldsymbol{v}}$. The computation of Jacobian $\mathbf{J}_{\boldsymbol{\xi},\widehat{\boldsymbol{Z}}}$ is based on the initial object pose guess. We compute the **SDF** uncertainty, $\boldsymbol{\Sigma}_{\mathbf{sdf},\widehat{\boldsymbol{Z}}}$, using the online rendering process (described in Section IV-C).

Our NBV is determined over candidate viewpoints $\{\boldsymbol{V}\}$ by minimizing the predicted entropy of the object pose:

$$\boldsymbol{v}^* = \underset{\widehat{\boldsymbol{v}}}{\operatorname{argmin}}\ h_e\left(\boldsymbol{\Sigma}_{\boldsymbol{\xi},\overline{\boldsymbol{Z}}}\right) \tag{23}$$

Once the next-best-view $\boldsymbol{v}^*$ is determined, the camera is moved, and a measurement set $\boldsymbol{Z}^*$ is collected from the corresponding viewpoint. We append the measurement set by $\mathbf{Z}_{1:K} \cup \mathbf{Z}^* \rightarrow \mathbf{Z}_{1:K+1}$ to recompute the object pose and perform the NBV selection again using Equations (21)-(23). This process is repeated until the predicted entropy falls below a user-defined threshold or a maximum number of views is selected.

VI. EXPERIMENTS

To demonstrate the effectiveness of our active pose refinement system, we aim to address two questions: (1) Can our pose refinement module accurately recover object poses given depth measurements? (2) Can our active vision policy achieve optimal performance with minimal viewpoints? To answer the question (1), we compare our pose refinement module with the classical ICP algorithm, given the same input depth data. For question (2), we use our pose refinement module and test our active vision approach against two heuristic-based policies. We obtain initial object poses using LINE-MOD [23], starting with object segmentation using an instance segmentation network from [18], followed by feeding the segmented objects into the pose estimator. An initial object pose will be used for evaluation only if its translation error is smaller than 30 mm and the rotation error is smaller than 30 degrees. In our evaluation, a ground truth pose will be considered if its occlusion rate is less than 20%.

A. Datasets and Evaluation Metrics

In our experiments, we use an industrial-grade SLI camera (IDS ENSNESO N35), which equips with two cameras and a visible-light projector. We evaluate our method on the ROBI dataset [11], which was captured using this camera. The ROBI dataset provides multi-view depth maps and pattern-projected images for shiny objects. The dataset also includes precisely labeled ground truth 6D object poses. In our experiments, We select five of the shiniest objects and evaluate each object individually.

Objects\Metrics	ADD	(5 mm, 5°)	(2 mm, 2°)
Eye Bolt	75.8	9.3	0.19
Tube Fitting	63.4	18.4	0.86
Chrome Screw	76.2	34.5	5.35
Gear	69.7	24.7	1.31
Zigzag	96.0	19.9	1.14
ALL	**76.2**	**21.4**	**1.77**

TABLE I: **Correct detection rate (%) on initial object poses (before the refinement)** with different evaluation metrics. Compared to the widely used ADD metric, the (5 mm, 5°) and (2 mm, 2°) metrics have much more strict criteria to consider an object pose as the correct detection.

Methods	Eye Bolt		Tube Fitting		Chrome Screw		Gear		Zigzag		ALL	
	(5,5)	(2,2)	(5,5)	(2,2)	(5,5)	(2,2)	(5,5)	(2,2)	(5,5)	(2,2)	(5,5)	(2,2)
Initial Pose	9.3	0.19	18.4	0.86	34.5	5.35	24.7	1.31	19.9	1.14	21.4	1.77
SDF+Random	91.0	73.5	84.1	60.3	69.2	19.3	93.8	89.5	96.6	87.5	86.9	66.0
ICP+Random	88.8	72.0	77.3	45.4	72.3	19.5	93.1	85.6	95.2	81.2	85.3	60.7
SDF+Max	87.4	68.0	77.6	43.0	60.3	14.7	96.6	91.6	96.4	**89.0**	83.7	61.3
ICP+Max	82.7	64.2	71.6	34.5	67.5	15.4	94.0	85.4	95.1	82.5	82.2	56.4
SDF+NBV	**94.5**	**81.5**	**94.1**	**76.8**	**77.5**	**24.3**	**97.8**	**97.7**	**98.8**	88.6	**92.5**	**73.8**

TABLE II: Object pose refinement results with different refinement approaches and view selection strategies on the ENSENSO test set from the ROBI dataset [11], expressed as the correct detection rate. An object pose is considered correct if it lies within (5 mm, 5°), or (2 mm, 2°) of ground truth. The maximum number of viewpoints is set to 2.

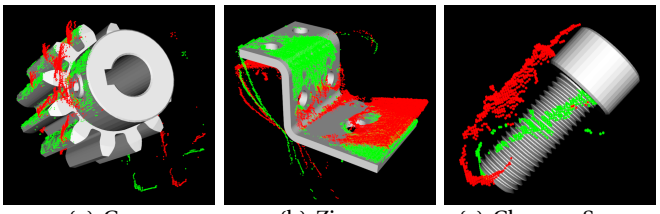
To evaluate the object pose performance, most works use the average model distance (ADD) metric [23], [3], [4], [5], [2]. However, this metric is not strict enough and cannot guarantee a high pose accuracy (e.g., it tolerates up to 30-degree rotation error on some parts). Instead, we evaluate the 6D object pose accuracy using the 5-mm/5-degree (5,5) and 2-mm/2-degree (2,2) metrics. The 5-mm/5-degree metric considers a pose correct if the translation error is under 5 mm and the rotation error is under 5 degrees. In Table I, we demonstrate the correct detection rate of initial object poses using ADD, (5,5) and (2,2) metrics. Although initial object poses can achieve a high detection rate with the ADD metric, but performance drops significantly with the (5,5) and (2,2) metrics, showing the importance of pose refinement and NBV.

B. Object Pose Refinement Evaluation

We first visualize how the poses of objects are refined in Figure 9. Our pose refinement module reliably and accurately refines object poses even in the presence of many outliers and the initial pose has a large error.

For quantitative evaluation, we compare our SDF-based refinement approach with the widely used ICP algorithm. For the viewpoint selection, we compare our NBV approach against two heuristic-based strategies as the baselines. The first baseline, "Random", selects viewpoints randomly from candidate viewpoints. The second baseline, "Max-Distance" moves the camera to the furthest distance location from previous viewpoints. For all approaches, we use the same initial object poses and perform the pose evaluation on each individual object. To obtain the results, the maximum number of viewpoints is set to 2.

The results of the object refinement are summarized in Table II. It is evident that the initial object poses have low detection rates for both 5-mm/5-degree and 2-mm/2-degree metrics. The results can be significantly improved when performing the pose refinement with the depth data. With the **same viewpoint selection strategy** ("Random" and "Max-Distance"), our SDF refinement module outperforms the ICP algorithm in almost all tests. Using the 5-mm/5-degree metric, our pose refinement module outperforms ICP by 1.6% for the "Random" strategy, and 1.5% for the "Max-Distance" strategy. With the stricter 2-mm/2-degree metric, our refinement method surpasses ICP by a more significant margin of 5.3% and 4.9% for "Random" and "Max-Distance" strategy, respectively. However, it is noteworthy that our SDF refinement module performs worse than ICP on object "Chrome Screw", mainly due to its extreme shininess, resulting in many missing depth data on the surface. Moreover, as shown in Figure 9c, this object's cylindrical shape lacks geometric constraints, making it difficult for the optimization to find the global minima. Hence, selecting informative viewpoints to acquire sufficient depth measurements is crucial.



(a) Gear (b) Zigzag (c) Chrome Screw

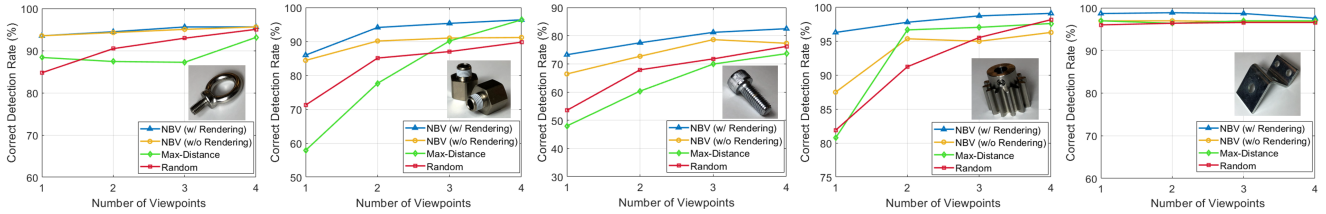
Fig. 9: Example refinement results on the ROBI dataset. The red and green point clouds are transformed by initial and refined object pose, respectively.

C. Next-Best-View Evaluation

To demonstrate the advantage of our next-best-view approach, we use the same pose refinement module (SDF-based approach) for a fair comparison. As shown in Table II, comparing to the "Random" and "Max-Distance" baselines, our method outperforms them by a large margin of at least 5.6% with the 5-mm/5-degree metric and 7.8% with the 2-mm/2-degree metric, respectively.

Figure 10 further presents the NBV results when using different number of viewpoints with the 5-mm/5-degree metric. To achieve the same level of correct detection rate, our NBV policy (blue curve) requires much fewer viewpoints than the baselines (red and green curves). This phenomenon is more obvious when using fewer and fewer viewpoints. When compared to the "Random" strategy, our NBV approach outperforms it by 12.1% for 1-view and 5.1% for 3-view test set. Compared to the "Max-Distance", the NBV policy exceeds it by 15.1% and 5.6% for 1-view and 3-view, respectively. For the shiniest object "Chrome Screw", our NBV policy achieves a high detection rate, 73.2%, with only one viewpoint. This result is comparable to the baseline policies when using four views (76.2% for the "Random", 73.6% for the "Max-Distance").

As presented in Section IV and V, a key component of our NBV policy is the depth uncertainty prediction of



(a) Eye Bolt (b) Tube Fitting (c) Chrome Screw (d) Gear (e) Zigzag

Fig. 10: Evaluation of our next-best-view policy when comparing against two heuristic-based baselines. We use our pose refinement module for all the viewpoint selection strategies. The results are evaluated using the correct detection rate with the 5-mm/5-degree metric. Our approach can achieve a high correct detection rate with much fewer viewpoints.

future viewpoints by online rendering. To demonstrate its effectiveness, we implement an alternative version of our NBV approach, one which assumes the depth uncertainty is constant for different future camera viewpoints. This version does not require online rendering and predicts the object pose covariance (Equation 22) with the Jacobian approximation only. As shown in Figure 10, the NBV can achieve high performance without predicting the depth uncertainty (yellow curve) for the object "Eye Bolt" and "Zigzag". This is because these two objects have low specular reflection, and the depth uncertainty is consistent for a wide range of different viewpoints. However, when objects have strong specular reflection (e.g., "Chrome Screw", "Tube Fitting", "Gear"), the NBV performance can be significantly improved by including the depth uncertainty prediction module.

VII. CONCLUSIONS AND FUTURE WORK

In this paper, we present an active vision framework for refining 6D pose and predicting next-best-view for shiny objects. Based on the SLI camera, we first estimate depth measurement uncertainties and integrate them into our object pose refinement module. Our framework refines the object pose and selects the NBV by minimizing predicted uncertainties. We evaluate our approach on a challenging real-world dataset for shiny objects. Results demonstrate that our pose refinement module outperforms the classical ICP algorithm when using the same input depth data. Additionally, our NBV policy achieves high pose refinement accuracy with significantly fewer viewpoints compared to heuristic baselines. In future work, we will investigate how to include the initial object pose estimation into our active vision framework, and explore how RGB images can be leveraged in a similar way, eliminating the specialization of our approach to the SLI camera setting.

REFERENCES

[1] X. Deng, Y. Xiang, A. Mousavian, C. Eppner, T. Bretl, and D. Fox, "Self-supervised 6d object pose estimation for robot manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[2] 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 *IEEE/CVF Conference on Computer Vision and Pattern recognition (CVPR)*, 2019.

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

[4] M. Sundermeyer, Z.-C. Marton, M. Durner, M. Brucker, and R. Triebel, "Implicit 3d orientation learning for 6d object detection from rgb images," in *European Conference on Computer Vision (ECCV)*, 2018.

[5] X. Deng, A. Mousavian, Y. Xiang, F. Xia, T. Bretl, and D. Fox, "Poserbpf: A rao–blackwellized particle filter for 6-d object pose tracking," *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1328–1342, 2021.

[6] K.-T. Song, C.-H. Wu, and S.-Y. Jiang, "Cad-based pose estimation design for random bin picking using a rgb-d camera," *Journal of Intelligent & Robotic Systems*, vol. 87, pp. 455–470, 2017.

[7] J. Yang, D. Li, and S. L. Waslander, "Probabilistic multi-view fusion of active stereo depth maps for robotic bin-picking," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4472–4479, 2021.

[8] Y. Wu, Y. Zhang, D. Zhu, X. Chen, S. Coleman, W. Sun, X. Hu, and Z. Deng, "Object slam-based active mapping and robotic grasping," in *International Conference on 3D Vision (3DV)*, 2021.

[9] J. Yang and S. L. Waslander, "Next-best-view prediction for active stereo cameras and highly reflective objects," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[10] T. Schmidt, R. A. Newcombe, and D. Fox, "Dart: Dense articulated real-time tracking.," in *Robotics: Science and Systems (RSS)*, 2014.

[11] J. Yang, Y. Gao, D. Li, and S. L. Waslander, "Robi: A multi-view dataset for reflective objects in robotic bin-picking," in *International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[12] H. Zhang and Q. Cao, "Fast 6d object pose refinement in depth images," *Applied Intelligence*, vol. 49, pp. 2287–2300, 2019.

[13] S. Zhang and S.-T. Yau, "High dynamic range scanning technique," *Optical Engineering*, vol. 48, no. 3, p. 033604, 2009.

[14] X. Liu, W. Chen, H. Madhusudanan, J. Ge, C. Ru, and Y. Sun, "Optical measurement of highly reflective surfaces from a single exposure," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, 2020.

[15] C. Forster, M. Pizzoli, and D. Scaramuzza, "Appearance-based active, monocular, dense reconstruction for micro aerial vehicles," in *Robotics: Science and Systems (RSS)*, 2014.

[16] Z. Zhang and D. Scaramuzza, "Perception-aware receding horizon navigation for mavs," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2534–2541, IEEE, 2018.

[17] Z. Zhang and D. Scaramuzza, "Beyond point clouds: Fisher information field for active visual localization," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[18] J. Yang, W. Xue, S. Ghavidel, and S. L. Waslander, "6d pose estimation for textureless objects on rgb frames using multi-view optimization," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[19] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *ACM SIGGRAPH*, pp. 1–10, 2008.

[20] W. Jakob, S. Speierer, N. Roussel, M. Nimier-David, D. Vicini, T. Zeltner, B. Nicolet, M. Crespo, V. Leroy, and Z. Zhang, "Mitsuba 3 renderer," 2022. https://mitsuba-renderer.org.

[21] B. Burley, "Extending the disney brdf to a bsdf with integrated subsurface scattering," *Physically Based Shading in Theory and Practice'SIGGRAPH Course*, 2015.

[22] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, (San Diega, CA, USA), 2015.

[23] 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 *Asian Conference on Computer Vision (ACCV)*, 2012.