2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

Probabilistic Multi-View Fusion of Active Stereo Depth Maps for Robotic Bin-Picking

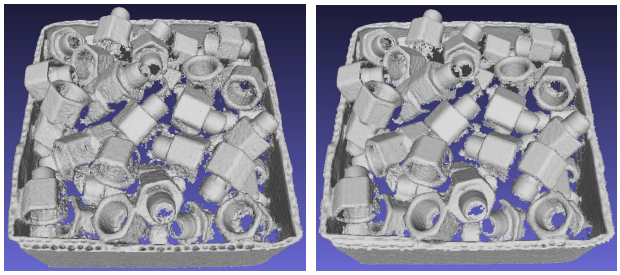
Jun Yang[1], Dong Li[2], and Steven L. Waslander[1]

***Abstract*— The reliable fusion of depth maps from multiple viewpoints has become an important problem in many 3D reconstruction pipelines. In this work, we investigate its impact on robotic bin-picking tasks such as 6D object pose estimation. The performance of object pose estimation relies heavily on the quality of depth data. However, due to the prevalence of shiny surfaces and cluttered scenes, industrial grade depth cameras often fail to sense depth or generate unreliable measurements from a single viewpoint. To this end, we propose a novel probabilistic framework for scene reconstruction in robotic bin-picking. Based on active stereo camera data, we first explicitly estimate the uncertainty of depth measurements for mitigating the adverse effects of both noise and outliers. The uncertainty estimates are then incorporated into a probabilistic model for incrementally updating the scene. To extensively evaluate the traditional fusion approach alongside our own approach, we will release a novel representative dataset with multiple views for each bin and curated parts. Over the entire dataset, we demonstrate that our framework outperforms a traditional fusion approach by a 12.8% reduction in reconstruction error, and 6.1% improvement in detection rate. The dataset will be available at `https://www.trailab.utias.utoronto.ca/robi`.**

I. INTRODUCTION

Bin-picking is a high value robotic task as it is able to take a tedious, repetitive and dangerous task out of workers' hands. The goal is to have a vision-guided robot to pick up known objects with random poses from a bin. Towards this goal, highly accurate 6D object pose estimation [1], [2], [3], [4], [5] is required prior to grasp planning. Due to the requirements of high accuracy and short cycle time in bin-picking, active stereo cameras have been used regularly for this task. Equipped with two cameras and a light projector, its imaging technology simplifies the stereo matching problem and provides reliable depth data. Historically, such cameras are mounted above the bin, and the object pose estimation has been addressed from a static viewpoint. However, from this viewpoint, active stereo cameras often fail to sense complete depths throughout the field of view due to reflective object materials, limited sensor resolutions and occlusions in industrial environments. To overcome these sensing limitations, a multi-view strategy can be employed, but the reliable fusion of these views is a critical step in realizing the benefits of multiple views.

To efficiently fuse multi-view depth maps, volumetric fusion based on truncated signed distance functions (TSDF) is a commonly used approach [6], [7]. Given sequential depth maps with associated camera poses, the TSDF method is able to perform local updates cumulatively via uniformly weighted averaging. The high memory usage for operations on 3D voxel grids can be reduced by voxel hashing [8] or octrees [9]. All these methods assume that all depth measurements have the same noise level, and are treated equally in TSDF fusion. However, this assumption is usually incorrect since depth uncertainty is typically dependent on the object surface as well as the type of camera sensor and its viewpoint. Moreover, the TSDF fusion does not explicitly handle outliers, therefore depth maps have to be pre-filtered before the fusion step.



(a) Standard TSDF Fusion (b) Ours

Fig. 1: Standard TSDF Fusion vs. Our fusion approach.

To tackle the limitations of TSDF, a combination of probabilistic modeling and volumetric representation has been proposed in [10], which fuses measurements using a probabilistic signed distance function (PSDF). This method requires pre-computed depth uncertainty as input for handling noise and outliers. However, the depth uncertainty model is generally difficult to develop since it is determined by various aspects, such as imaging technology, sensor viewpoint as well as scene characteristics. Due to the importance of active stereo camera in robot bin-picking, in this work, we explicitly model depth uncertainties from such sensor at different viewpoints, and incorporate them into a voxel-based probabilistic fusion framework. Our approach is able to perform reliable depth fusion and reconstruct the scene with more details and fewer outliers, as depicted in Figure 1. This will improve 6D object pose estimation performance by providing high quality input depth data.

To demonstrate the advantages of our framework, we present a real-world dataset in bin-picking scenarios. Compared to existing bin-picking related datasets [1], [3], [11], [12], our dataset has unique characteristics. It

[1]J. Yang and S.L. Waslander are with University of Toronto Institute for Aerospace Studies and Robotics Institute. {junyang.yang@mail, stevenw@utias}.utoronto.ca

[2]D. Li is with Epson Canada Ltd. dong.li@ea.epson.com

consists of reflective parts and over 30 individual bin instances. We use a robot arm to capture monochrome images and depth maps at multiple viewpoints. For each object, we provide the object model, annotations of 6D object poses, and ground truth depth in real-world.

In summary, we make the following contributions:

- A probabilistic framework for scene reconstruction in the bin-picking problem. The framework comprises of a) the explicit estimation of pixel depth uncertainties for active stereo cameras, b) the integration of uncertainty estimates with a probabilistic model for multi-view depth map fusion.
- A new multi-view dataset for robotic bin-picking. To the best of our knowledge, this is the first bin-picking dataset with ground truth depth in real-world.
- The extensive evaluation of the traditional fusion approach and our method of multi-view depth fusion, as well as its impact on 6D object pose estimation in the bin-picking scenario.

II. RELATED WORK

Object Detection and 6D Pose Estimation. Accurate 6D object pose estimation is a key step in most robotic bin-picking solutions. In past works, the most representative methods are based on template matching, such as LINEMOD [1], and its extensions. These methods rely on RGB or RGB-D images as inputs, and are effective in cluttered scene. Point-to-point techniques [2] represent another category, and are based on depth data only. Many learning-based approaches have been proposed recently to solve 6D object pose estimation problem in an end-to-end manner [4], [5]. The networks are normally trained on synthetic data using 3D CAD models, and tested on real data. Despite the varied pose estimation strategies, the performance of these methods usually relies heavily on the quality of measured depth data.

Multi-View Volumetric Depth Fusion. When operating limits permit, multi-view fusion is able to provide higher levels of scene completion than single-view methods. In [6], Curless and Levoy initially presented an effective method to fuse multi-view depth maps based on TSDF. This method has been later integrated into influential works, such as KinectFusion [7], and further improved with lower memory usage [8], [9]. However, TSDF methods fuse noisy depth maps via uniform weighted averaging in each voxel. Hence, they do not account for the fact that, when observing from different viewpoints, depth measurements may present different levels of noise and outliers. To tackle this limitation, [13] and [14] proposed end-to-end learning-based approaches for volumetric depth map fusion. These methods are able to handle sensor noise and outliers, but require a large amount of training data and are relatively prone to over-fitting to a particular sensor or dataset. Based on a model from [15], Wei *et al.* proposed the PSDF Fusion framework [10] for general scene reconstruction. This method requires pre-computed depth uncertainty for incremental Bayesian updating. However, the depth uncertainty is dependent on many scene and sensor characteristics, and is, in general, difficult to acquire.

Uncertainty Estimation for Stereo. In the context of stereo matching, a wide range of approaches has been proposed for uncertainty estimation. Such estimates were used for outlier removal [16] and depth map fusion [14]. A detailed review of uncertainty estimation for stereo can be found in [16]. Due to the success of deep learning, some recent works have explored it to learn the uncertainties of disparities intrinsically from data [17], [18]. The ground truth disparities are usually required for the supervised training, which can be difficult to acquire in practice.

III. METHOD

Our framework consists of two main parts: (a) uncertainty estimation of active stereo depth maps, (b) volumetric probabilistic integration. In the first part, we use raw disparity maps and corresponding active stereo pairs, and estimate the measurement uncertainties by evaluating photometric uniqueness and geometric consistency. Our uncertainty models are especially useful for determining the pixel-level reliability of depth measurements from active stereo camera sensors. The second part, volumetric probabilistic integration, integrates our uncertainty estimates into a voxel-based probabilistic framework [10]. Based on Bayesian inference, the scene can be updated in an incremental fashion. Figure 2 shows an overview of the framework. We assume the camera poses are known, and all computations are performed in a single world coordinate system.

A. Uncertainty from 2D Photometry

To determine photometric uncertainty from active stereo data, we operate on individual pixels by examining their cost curves as a function of the disparity hypothesis. The conventional active stereo camera employs a light projector to create light patterns (e.g., pseudorandom dots) on texture-less regions. This imaging technology simplifies the stereo matching problem by pushing the cost curve to an ideal form. As demonstrated in Figure 3a, the ideal cost curve has a distinct global minimum. However, due to occlusion, shadows and surface reflection, the projected pattern texture is not always visible, which leads to ambiguous cost curves. Figure 3b shows this ambiguity, where multiple local minima with similar cost make localization of the global minimum hard. To determine this uncertainty, we compute a confidence score for each pixel in the raw disparity map. The confidence value indicates the distinctiveness of assigned disparity with adjacent disparity hypotheses.

Given a stereo pair of rectified, grayscale, left I_L and right I_R pattern projected images, as well as its disparity map D, we first assign a cost values $c(x,y,d)$ to the integer disparity hypothesis $d \in \mathbb{Z}_+$ of pixel $(x,y) \in I_L$. To reduce illumination sensitivity, we compute the cost value $c : I \times \mathbb{Z}_+ \rightarrow \mathbb{C} = [0,2]$ by converting the normalized

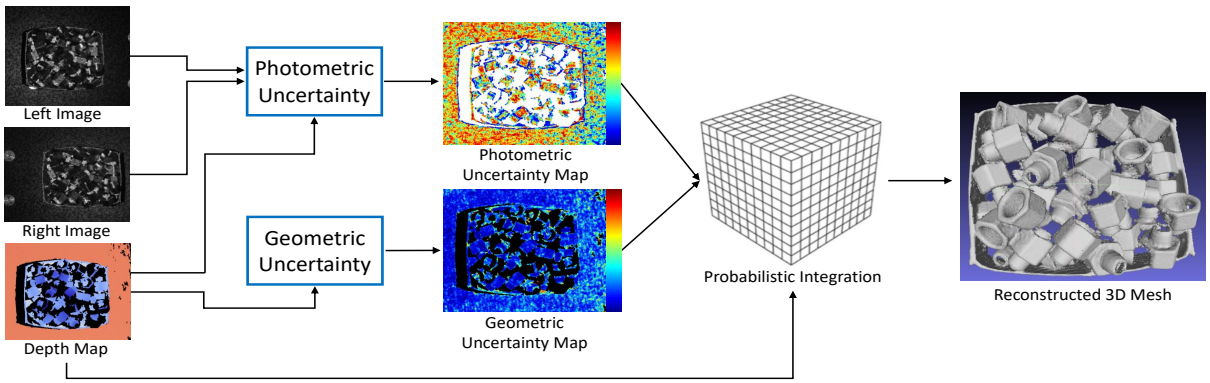


Fig. 2: **Overall system pipeline.** Given depth maps and corresponding active stereo pairs, our system estimates depth uncertainty from both photometric and geometric properties. These uncertainties are then incorporated into a voxel-based probabilistic framework for incrementally updating of the scene.

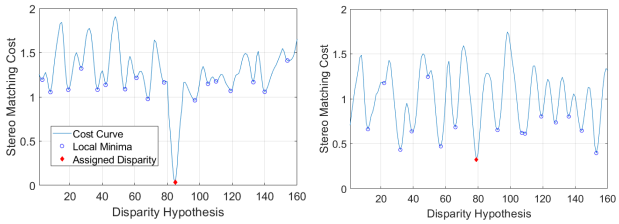


Fig. 3: (a) Ideal cost curve. (b) Ambiguous cost curve.

cross correlation (NCC) as $c(x,y,d) = 1 - NCC(x,y,d)$. The normalized cross correlation is defined as:

$$NCC(x,y,d) = \frac{\sum_{i \in W} \left(I_L(x_i,y_i) - \mu_L\right)\left(I_R(x_i - d, y_i) - \mu_R\right)}{\sigma_L \sigma_R} \tag{1}$$

where $\mu_L, \mu_R \in \mathbb{I}$ denote the means of all pixel intensities within the square window W of the left and right image, respectively, and σ_L and σ_R are corresponding standard deviations. To compute the confidence score, we choose one of the most promising metrics from [16], namely maximum likelihood measure (MLM), and its metric is modified to become:

$$C_{MLM}(x,y,d_1) = \frac{e^{-\frac{c(x,y,d_1)}{2\sigma^2}}}{\sum_{d'} e^{-\frac{c(x,y,d')}{2\sigma^2}}} \tag{2}$$

where parameter σ denotes the disparity uncertainty, $c(d_1)$ and $c(d')$ represent the cost value of assigned disparity and local minima, respectively. MLM measures the distinctiveness of assigned integer disparity d_1. When margin between $c(d_1)$ and $c(d')$ is larger, it is more likely the assigned disparity is correct. However, the MLM still lacks information about the "strength" of left-right correlation. To account for this, we utilize the original NCC and the final confidence score is computed as:

$$C(x,y,d_1) = \frac{NCC(x,y,d_1)+1}{2} \times C_{MLM}(x,y,d_1) \tag{3}$$

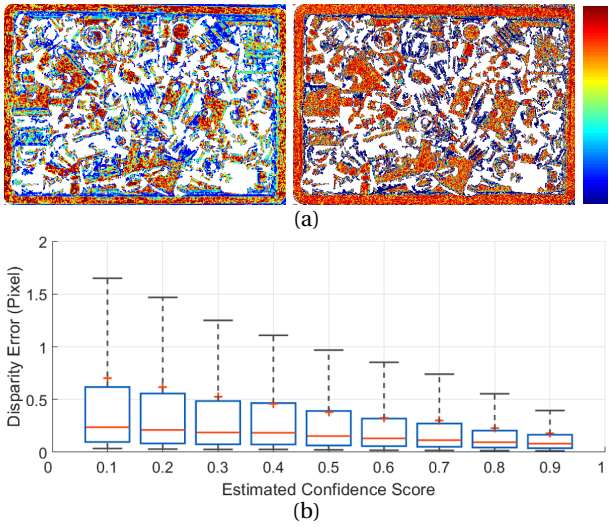


Fig. 4: **Photometric Uncertainty Estimation.** (a) Left: Our confidence measure overlaid on the disparity map. Right: The corresponding error map. Red indicates high confidence (low error) of assigned disparity, and blue low confidence (high error). (b) The correlation between confidence value and disparity error. The paired monochrome images and depth map are shown in Figure 8a.

For practical reasons, we scale the confidence metric to the interval $[0,1]$, and the results of our estimated confidence measure are illustrated in Figure 4.

B. Uncertainty from 3D Geometry

In addition to the photometric uncertainty, the uncertainty of geometry is also calculated in 3D space. We examine consistency of each 3D point $\mathbf{p} \in \mathbb{R}^3$, originating from raw disparity map, D, with its local neighbors. For this purpose, we fit a local surface and measure how far $\mathbf{p}$ is to it. We utilize several steps for this computation, which were inspired by data resampling techniques [19].

For each point, $\mathbf{p}$, from input point cloud, we first project it from camera frame xyz to a local reference

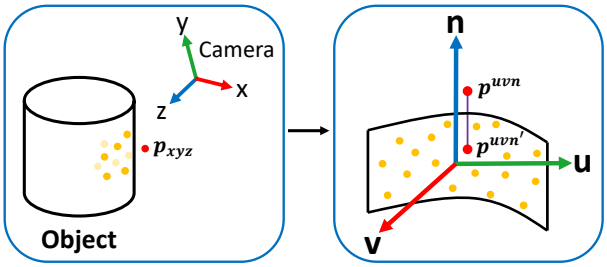


Fig. 5: The 3D point $\mathbf{p}$ (red dot) and its N nearest neighbors $N(\mathbf{p})$ (yellow dots) are transformed from camera coordinate xyz to local reference frame uvn.

frame uvn, fitted via Principal Component Analysis (PCA) on its N nearest neighbors, $N(\mathbf{p})$. This is illustrated in Figure 5. Under the coordinate system of uvn, a bivariable quadratic function is fit to the heights of the points above the plane:

$$f(u,v) = au^2 + bv^2 + cuv + du + ev + f \tag{4}$$

The coefficients $[a,b,c,d,e,f]$ can be computed via a closed-form least squares solution. More details on the computation of polynomial surface can be found in [19].

As demonstrated in Figure 5, the signed offset from $\mathbf{p}$ to the local fitted surface is:

$$\epsilon_{\mathbf{p}} = n_{\mathbf{p}} - n'_{\mathbf{p}} \tag{5}$$

where $n_{\mathbf{p}}$ is the height of point $\mathbf{p}^{uvn}$ and $n'_{\mathbf{p}}$ is its recalculated height. The offset for its neighbors $N(\mathbf{p})$ can be computed in the same way, and is represented as $\epsilon_{N(\mathbf{p})}$. With a Gaussian assumption for inlier measurements, we let $\epsilon_{\mathbf{p}}$ be the expectation of the distribution, and treat $\epsilon_{N(\mathbf{p})}$ as its observations. The geometric uncertainty for 3D point $\mathbf{p}$ is finally obtained as the variance of the distribution:

$$\sigma^2_{\mathbf{p}} = \frac{\sum \epsilon^2_{N(\mathbf{p})}}{N} - \epsilon^2_{\mathbf{p}} \tag{6}$$

The results of the estimated and measured geometric uncertainty are illustrated in Figure 6.

C. Probabilistic Volumetric Integration

Given the depth map $Z_k \in \mathbb{R}$ (originating from raw disparity map, with known baseline and focal length) and its camera pose $\mathbf{T}_k \in SE(3)$, the classic TSDF fusion [6], [7] integrates them to signed distance functions (SDF) F_k:

$$F_k = Z_k - Z^g \tag{7}$$

where $Z^g \in \mathbb{R}$ is a constant value, defined in global frame. To integrate our estimated uncertainties into the global SDF volume, we leverage a per-voxel probabilistic framework, introduced in [10].

Based on the model proposed in [15], we describe the k^{th} SDF observation for a voxel, F_k, as a $Gaussian + Uniform$ mixture model distribution. Specifically, an inlier measurement is normally distributed around the true

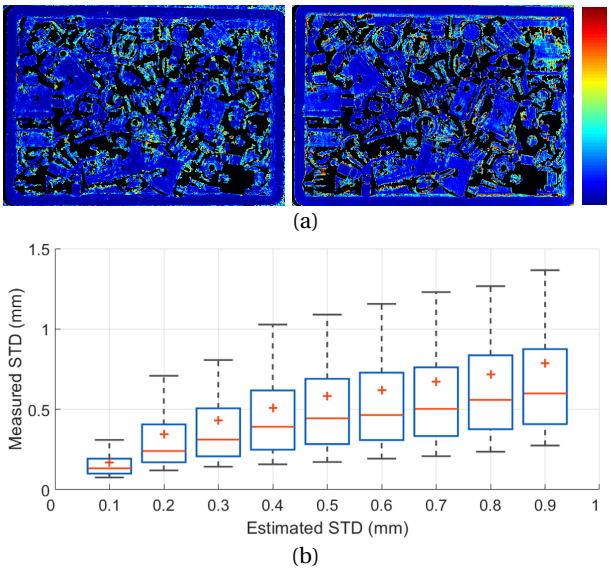


Fig. 6: **Geometric Uncertainty Estimation.** (a) Left: Our estimated STD on depth map. Right: The measured STD from 50 depth maps. Large and small STD of depth measurement are colored in red and blue, respectively. (b) The correlation between estimated and measured STD. The paired monochrome images and depth map are shown in Figure 8a.

SDF value $\hat{F}$, while an outlier is uniformly distributed in an interval $[F_{min}, F_{max}]$:

$$p\left(F_k|\hat{F},\pi\right) = \pi N\left(F_k|\hat{F},\tau_k^2\right) + (1-\pi)\, U(F_k|F_{min},F_{max}) \tag{8}$$

where τ_k^2 and π are the variance and probability of an inlier measurement, respectively. For Bayesian updates, the posterior mixture model can be approximated by the product of a Gaussian and a Beta distribution [15] representing the inlier probability:

$$q\left(\hat{F},\pi|a_k,b_k,\mu_k,\sigma_k^2\right) \triangleq Beta(\pi|a_k,b_k)\, N\left(\hat{F}|\mu_k,\sigma_k^2\right) \tag{9}$$

where a_n and b_n are parameters of the Beta distribution. The update takes the following form:

$$q\left(\hat{F},\pi|a_k,b_k,\mu_k,\sigma_k^2\right) \approx p\left(F_k|\hat{F},\pi\right) q\left(\hat{F},\pi|a_{k-1},b_{k-1},\mu_{k-1},\sigma_{k-1}^2\right) \tag{10}$$

The true posterior of (10) by equating first and second order moments for $\hat{F}$ and π. The updates for μ_k and σ_k^2 are derived as:

$$L_1 = \frac{a_{k-1}}{a_{k-1}+b_{k-1}} N\left(F_k|\mu_{k-1},\sigma_{k-1}^2+\tau_k^2\right) \tag{11}$$

$$L_2 = \frac{b_{k-1}}{a_{k-1}+b_{k-1}} U(F_k|F_{min},F_{max}) \tag{12}$$

$$\frac{1}{s^2} = \frac{1}{\sigma_{k-1}^2} + \frac{1}{\tau_k^2}, \qquad \frac{m}{s^2} = \frac{\mu_{k-1}}{\sigma_{k-1}^2} + \frac{F_k}{\tau_k^2} \tag{13}$$

$$\mu_k = \frac{L_1}{L_1+L_2} m + \frac{L_2}{L_1+L_2} \mu_{k-1} \tag{14}$$

$$\mu_k^2 + \sigma_k^2 = \frac{L_1}{L_1+L_2}\left(s^2+m^2\right) + \frac{L_2}{L_1+L_2}\left(\sigma_{k-1}^2+\mu_{k-1}^2\right) \tag{15}$$

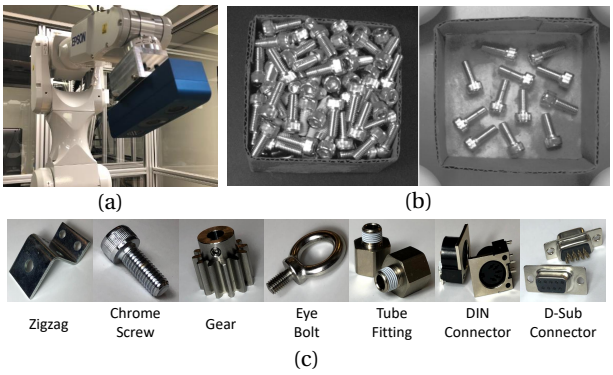


Fig. 7: **Overview of ROBI dataset.** (a) we mount an Ensenso N35 camera on an EPSON C4L robot arm. (b) Two bin-picking scenarios. Left: Full-bin scene. Right: Low-bin scene. (c) An overview of all objects of the dataset.

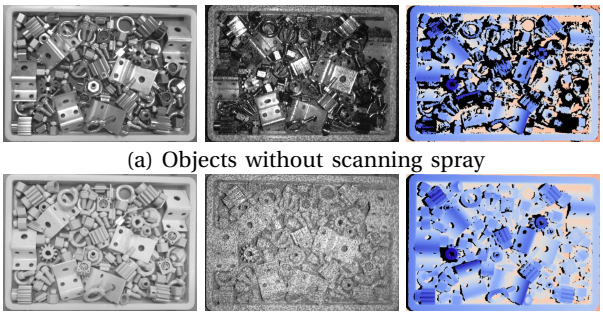


Fig. 8: **Illustration of using scanning spray on objects. Left**: Left stereo image from monochrome configuration. **Middle**: Left stereo image from depth configuration. **Right**: Depth map from depth configuration.

The computation of a_k and b_k are the same as [15] and are excluded here for compactness.

Due to the linear form in (7), we use our geometric uncertainty directly as the approximation of the variance of the inlier SDF: $\tau_k \approx \sigma_{\mathbf{p}}$. Note $\sigma_{\mathbf{p}}$ is computed per-frame for updating Equations (11) - (15). To fully take advantage of estimated uncertainties, a per-frame inlier probability π_k is computed for F_k in [10], to replace the expectation of $Beta(\pi|a_{k-1}, b_{k-1})$ in Equation (11) and (12). Inspired by this, we employ the estimated photometric uncertainty C for evaluating inlier probability π_k. For this purpose, a mapping from C to the actual inlier probability is required. We leverage a training-based approach from [20], and construct the mapping as:

$$\pi_k = p(\mathbf{i}|C) = \frac{p(C|\mathbf{i}) \cdot p(\mathbf{i})}{p(C|\mathbf{i}) \cdot p(\mathbf{i}) + p(C|\mathbf{o}) \cdot \big(1 - p(\mathbf{i})\big)} \tag{16}$$

where $\mathbf{i}$ and $\mathbf{o}$ represent inlier and outlier, respectively. By using the labeled training data, we are able to extract $p(C|\mathbf{i})$, $p(C|\mathbf{o})$ as well as $p(\mathbf{i})$. The computed π_k is finally used to replace the simple $\frac{a_{k-1}}{a_{k-1}+b_{k-1}}$, and π is still parameterized by a and b.

D. Surface Extraction

Given SDF values, we use the Marching Cubes algorithm [21] to extract connected surfaces. Having sufficient photometric and geometric information within voxels, we check their convergence and determine zero-crossing points when the following conditions are satisfied:

$$\mu^{v_1} \cdot \mu^{v_2} < 0, \tag{17}$$

$$\sigma^{v_1} < \sigma_{thr} \quad \text{and} \quad \sigma^{v_2} < \sigma_{thr}, \tag{18}$$

$$\frac{a^{v_1}}{a^{v_1} + b^{v_1}} > \pi_{thr} \quad \text{and} \quad \frac{a^{v_2}}{a^{v_2} + b^{v_2}} > \pi_{thr}, \tag{19}$$

where v_1 and v_2 are adjacent voxels. The threshold σ_{thr} and π_{thr} control the estimation convergence and can be set according to the accuracy requirements of robotic bin-picking. This operation will reject outliers while preserving distinct features of the scene, which leads to higher quality data for later object pose estimation.

IV. REFLECTIVE OBJECTS IN BINS DATASET

To demonstrate the advantages of our framework, we provide ROBI (Reflective Objects in BIns), a multi-view dataset consisting of reflective objects in bin-picking scenes. For each scene and viewpoint, monochrome images, depth maps and annotations of 6D object poses are included. In addition, we provide ground truth depth maps captured by an active stereo camera with parts coated in anti-reflective scanning spray. We will release a public version of our dataset upon publication of this work, which can be used for scene reconstruction, 6D object pose estimation and depth completion tasks.

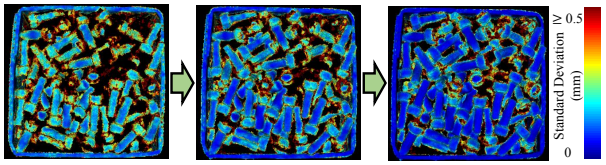
A. Sensor Setup

We captured the raw multi-view sensor data with an Ensenso N35 active stereo camera. The camera is equipped with a short working distance optical lens, and has the working distance from 240 mm to 520 mm. We captured images with following two configurations:

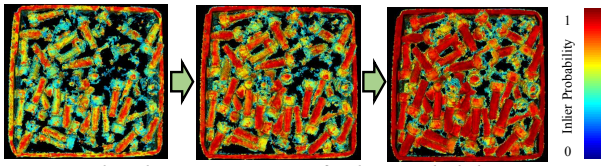
- **Depth configuration.** We turn camera projector on and use a low exposure time to eliminate the impact of ambient light. The raw disparity maps and pattern projected stereo pairs are collected.
- **Monochrome configuration.** The camera projector is turned off for this configuration. We use a high exposure time to obtain optimal contrast for objects. Only stereo pair data is saved with this configuration.

B. Data Capture Pipeline

The camera is mounted to an EPSON 6-Axis C4L robot arm, illustrated in Figure 7a. We program the robot arm to move in a trajectory that traverses a spherical dome and approaches the bin. The robot end-effector stays pointed towards the center of the workstation. The initial camera extrinsics are obtained by leveraging robot end-effector poses and an offline "eye-in-hand" calibration [22]. To further refine camera poses, we apply the iterative closest point (ICP) algorithm on calibration spheres, which are placed around the bin. The average closest-point residual error was successfully reduced from 0.33 mm to 0.26 mm.

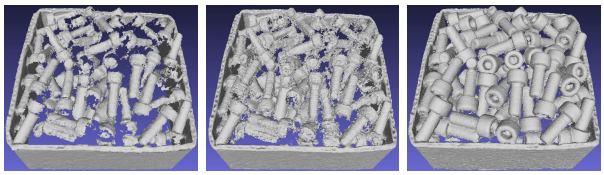


(a) The propagation of standard deviation for inlier meshes.

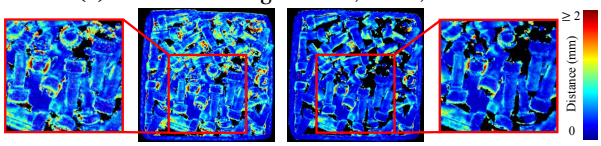


(b) The propagation of inlier probability.

Fig. 9: Incremental scene updating for object "Chrome Screw". From left to right: Reconstruction from 10 views, 30 views and 88 views. The paired monochrome image is shown on the left of Figure 7b.



(a) From left to right: Ours, TSDF, GT mesh.



(b) Left: TSDF. Right: Ours.

Fig. 10: (a) Comparison of output mesh between TSDF and our framework for object "Chrome Screw" in a full-bin scene. (b) Heatmap of the point-to-point distance from reconstructed model to the ground truth mesh.

We use seven challenging industrial objects, with different levels of reflectivity. An overview of all objects is depicted in Figure 7c. To demonstrate the bin-picking problem in realistic conditions, we separate our data capture into 2 scenarios: (a) full-bin: multiple objects are stacked on a bin with severe occlusions and clutter, (b) low-bin: a small number of objects are used, with spatial separation between parts. These two scenarios are shown in Figure 7b.

For each object, we capture 5 bin scenarios (3 for full-bin and 2 for low-bin). We sample a total of 88 views for full-bin scenarios, from approximately 45° to 90° of sphere elevation, with distances from 400 mm to 520 mm. Due to the occlusion of bin wall, in low-bin data capture, the sphere elevation is limited to the range of $[65°, 90°]$, and there are 42 views in total.

C. Ground Truth Acquisition

Depth Maps. Depth measurement by Ensenso camera suffers from large errors up to $2-3\,mm$ on reflective surfaces [23], which are common to industrial parts. To acquire ground truth depths, we apply a scanning spray [24] on objects to create Lambertian surfaces, so that the Ensenso camera can achieve its optimal accuracy (less than $0.2\,mm$). The scanning spray generates a thin and homogeneous layer with $8-15\,\mu m$ thickness, which is one to two orders of magnitude less than the expected depth accuracy. As illustrated in Figure 8, we capture the ground truth and test images with two scans. In the initial scan, we apply the scanning spray and capture ground truth depth maps. After spray evaporation, the test images are captured during the second scan.

Scene Model. For each scene, we construct the ground truth mesh by applying TSDF fusion [7] on ground truth depth maps. To demonstrate that the reconstruction of the ground truth scene is not biased to any fusion method, we apply both the TSDF and our method to produce two sets of ground truth meshes, and compute mean point-to-point distance between these two meshes. The mean point-to-point distance is $0.03\,mm$, indicating our ground truth accuracy is not sensitive to different fusion methods.

6D Object Poses. To annotate 6D object poses of a bin instance, we first manually align object CAD models to ground truth scene model. To increase accuracy, we upsample the CAD model to a higher resolution and use ICP algorithm to refine 6D object poses. The misalignment can be identified from scene models, and poses were then manually adjusted. We repeat this process several times until a satisfactory alignment was achieved.

V. EXPERIMENTS

The experiments regarding reconstruction and 6D pose estimation were performed on ROBI dataset, described in Section IV. Our framework is compared to traditional TSDF fusion as the baseline. To make the comparison fair, we also utilized a strategy from [25], [26] to reduce outliers for TSDF. Specifically, all voxels that have received the number of measurements under a given threshold W_{thr} are deleted. In our evaluation, parameter W_{thr} is empirically set to 3. For all experiments, we use a voxel size of $0.5\,mm$, the truncation distance is set to $1.5\,mm$.

A. Reconstruction Evaluation

We first visualize the propagation of surface uncertainties by rendering the inlier probability π_k and standard deviation σ_k in colormaps. As shown in Figure 9, when compared to the matte surface (bin wall), depth measurements for high gloss surface (objects in the bin) are more prone to noise and missed detections, hence high gloss reconstruction requires more views to converge. After the outlier rejection, the qualitative comparison, shown in Figure 10, illustrates that our framework produces smoother surfaces and fewer outliers than traditional TSDF.

For quantitative evaluation, we use three metrics by computing point-to-point distances between reconstructed mesh $\hat{\mathbf{M}}$ and ground-truth mesh $\mathbf{M}$. For each point in source mesh, its corresponding point is the closest point in target mesh. A point is defined as inlier if point-to-point Euclidean distance is smaller than $2\,mm$:

Bin	Object Category	Mean Point-to-Point Distance (mm)			Outlier Percentage (%)			Scene Completeness (%)		
		TSDF	TSDF w/ W_{thr}	Ours	TSDF	TSDF w/ W_{thr}	Ours	TSDF	TSDF w/ W_{thr}	Ours
Full	Large Size	0.42	0.29	**0.25**	18.3	**3.18**	4.1	**99.5**	98.8	99.3
	Complex Shape	0.58	0.46	**0.39**	13.4	2.18	**1.86**	**98.9**	95.8	96.8
	High Gloss	0.61	0.47	**0.41**	11.1	1.60	**1.51**	**95.9**	89.3	91.5
Low	Large Size	0.24	0.17	**0.13**	3.1	**0.29**	0.41	**99.9**	99.4	99.7
	Complex Shape	0.44	0.34	**0.3**	3.28	**0.35**	0.38	**90.1**	78.5	79.9
	High Gloss	0.43	0.36	**0.24**	1.88	**0.16**	**0.16**	**85.03**	74.5	77.0
Total		0.52	0.39	**0.34**	10.1	**1.56**	1.57	**95.1**	89.5	91.1

TABLE I: Quantitative reconstruction results in different object categories with varied bin scenarios.

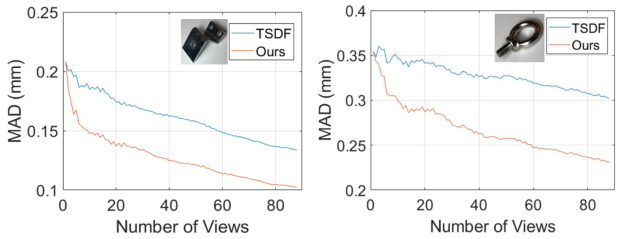
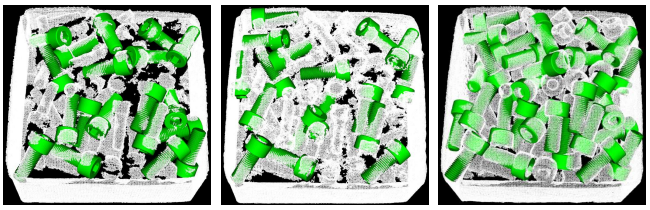


Fig. 11: Comparison of mean absolute distance (MAD) between TSDF and our methods. Left: Object "Zigzag" in full-bin. Right: Object "Eye Bolt" in full-bin.



(a) Ours (b) TSDF (c) GT mesh

Fig. 12: Correctly detected 6D poses for object "Chrome Screw" in a full-bin scene. The corresponding meshes are shown in Figure 10.

- **Mean Point-to-Point Distance.** We build correspondences from $\hat{\mathbf{M}}$ to $\mathbf{M}$, the mean point-to-point distance is computed over all inliers, and measures the reconstruction accuracy.
- **Outlier Percentage.** Given correspondences from $\hat{\mathbf{M}}$ to $\mathbf{M}$, the outlier percentage is computed as the fraction of the number of outliers over the total number of vertices in the ground truth mesh.
- **Scene Completeness.** We build correspondences from $\mathbf{M}$ to $\hat{\mathbf{M}}$, the scene completeness is computed as the fraction of the number of inliers over the total number of ground truth vertices.

These metrics are measured only on the surface of objects, which has a significant impact on later 6D object pose estimation. To illustrate reconstruction quality on different properties, we split all objects into three categories, based on their major characteristics: (a) Large Size ("Zigzag"), (b) Complex Shape ("DIN Connector" and "D-Sub Connector"), (c) High Gloss (remaining four objects).

Table I shows the quantitative reconstruction results of each object category with the accumulation of all viewpoints. Our method not only achieves higher reconstruction accuracy (12.8% improvement), but also makes a better trade-off between scene completeness and outlier percentage. Table I also reveals that the reconstruction is particularly difficult for high gloss objects using both TSDF and our methods. This indicates that our method is upper bounded by depth measurement quality and the viewpoint coverage. Moreover, despite higher accuracy and fewer outliers, our method also sacrifices some completeness of the scene. This phenomenon is more obvious for low bin data, which have occlusions of bin walls and lower viewpoint coverage for object surfaces to converge.

To demonstrate the fast convergence of our method, we compute the mean absolute distance (MAD) of SDF value over all occupied voxels. As illustrated in Figure 11, as the accumulation of viewpoints, TSDF generally has a slow convergence in terms of MAD. In comparison, our method requires much fewer viewpoints to achieve same accuracy, making it highly useful for active vision techniques [3].

B. 6D Object Pose Estimation Evaluation

To evaluate the performance of 6D object pose estimation, we choose the algorithm proposed by Drost et al. [2]. This method only relies on 3D point cloud data and solves object poses by coupling the idea of point-pair features (PPF) and a dense voting scheme. We compare our method against TSDF, its variant with thresholding strategy, ground truth mesh and the point cloud from a single (top center) viewpoint. All meshes are converted to point cloud by extracting their vertices.

As in [2], [5], we compute the correct detection rate for each object and each scene. To measure the 6D pose error, we use the standard ADD score from [1]. A pose hypothesis is counted as correct when its ADD score is less than 10% of the object diameter.

We present the correct detection rate in Table II. It can be seen that, given optimal 3D data (GT mesh), the PPF pose estimator can provide close to perfect detection rates. In comparison, due to a large amount of missing depths, the detection performance is generally poor on single view based point cloud. This performance can be improved with multi-view volumetric depth fusion. And as a result of the complete scene and few outliers, our method outperforms TSDF and its variant by a large margin of 6.1%. It is noteworthy that our improvements in detection rate are more significant for full-bin scenarios, which has more noise and outliers due to the clutter.

Bin	Object	Detection Rate (%)				
		Single View	TSDF	TSDF w/ W_{thr}	Ours	GT Mesh
Full	Zigzag	57.1	78.5	75.0	**87.5**	87.5
	Chrome Screw	0	33.3	37.2	**49.0**	83.3
	Gear	14.3	60.0	62.8	**67.6**	94.3
	Eye Bolt	42.1	**78.9**	68.4	**78.9**	97.3
	Tube Fitting	23.2	88.4	86.9	**91.3**	95.6
	DIN Connector	15.0	75.0	73.6	**76.6**	83.3
	D-Sub Connector	3.6	25.6	23.1	**32.9**	57.3
Low	Zigzag	41.6	**91.6**	**91.6**	**91.6**	100
	Chrome Screw	0	72.2	68.1	**77.2**	90.9
	Gear	38.4	53.8	46.1	**61.5**	100
	Eye Bolt	12.5	86.6	86.6	**93.7**	93.7
	Tube Fitting	9.5	71.4	66.7	**76.2**	100
	DIN Connector	9.1	**63.6**	31.8	59.1	90.9
	D-Sub Connector	0	**22.7**	9.1	**22.7**	95.4
Total		14.9	58.1	54.9	**64.2**	85.9

TABLE II: Detection rate results for different objects.

Further comparison in Figure 12 illustrates that, as a result of the smoother surface and fewer outliers (shown in Figure 10), the PPF pose estimator provides more correct object poses on the mesh that was constructed by our method versus the TSDF method.

VI. CONCLUSIONS

In this paper, we proposed a probabilistic framework for scene reconstruction in the bin-picking problem. Based on active stereo camera data, we explicitly estimated two types of depth uncertainties, and incorporated them into a probabilistic framework for incrementally updating of the scene. The high-quality mesh can be used as the input 3D data for improving the performance of 6D object pose estimation. To evaluate the performance of reconstruction and object pose estimation, we construct ROBI, a real-world dataset of reflective objects in bin-picking scenarios. Our approach outperforms the traditional TSDF for both reconstruction and 6D object pose estimation. As a future work, we want to investigate how to strategically select viewpoints, and achieve the optimal performance of reconstruction accuracy as well as 6D object pose estimation with the minimum number of views.

REFERENCES

[1] 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*, pp. 548–562, Springer, 2012.

[2] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3d object recognition," in *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 998–1005, Ieee, 2010.

[3] A. Doumanoglou, R. Kouskouridas, S. Malassiotis, and T.-K. Kim, "Recovering 6d object pose and predicting next-best-view in the crowd," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3583–3592, 2016.

[4] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1521–1529, 2017.

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

[6] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 303–312, 1996.

[7] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pp. 127–136, IEEE, 2011.

[8] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger, "Real-time 3d reconstruction at scale using voxel hashing," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 6, pp. 1–11, 2013.

[9] F. Steinbrucker, C. Kerl, and D. Cremers, "Large-scale multi-resolution surface reconstruction from rgb-d sequences," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3264–3271, 2013.

[10] W. Dong, Q. Wang, X. Wang, and H. Zha, "Psdf fusion: Probabilistic signed distance function for on-the-fly 3d data fusion and scene reconstruction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 701–717, 2018.

[11] T. Hodan, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis, "T-less: An rgb-d dataset for 6d pose estimation of texture-less objects," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 880–888, IEEE, 2017.

[12] K. Kleeberger, C. Landgraf, and M. F. Huber, "Large-scale 6d object pose estimation dataset for industrial bin-picking," *arXiv preprint arXiv:1912.12125*, 2019.

[13] D. Rozumnyi, I. Cherabier, M. Pollefeys, and M. Oswald, "Learned semantic multi-sensor depth map fusion," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[14] S. Weder, J. Schonberger, M. Pollefeys, and M. R. Oswald, "Routed-fusion: Learning real-time depth map fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4887–4897, 2020.

[15] G. Vogiatzis and C. Hernández, "Video-based, real-time multi-view stereo," *Image and Vision Computing*, vol. 29, no. 7, pp. 434–441, 2011.

[16] X. Hu and P. Mordohai, "A quantitative evaluation of confidence measures for stereo vision," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2121–2133, 2012.

[17] F. Tosi, M. Poggi, A. Benincasa, and S. Mattoccia, "Beyond local reasoning for stereo confidence estimation with deep learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 319–334, 2018.

[18] M. Poggi and S. Mattoccia, "Learning from scratch a confidence measure.," in *BMVC*, 2016.

[19] R. B. Rusu, Z. C. Marton, N. Blodow, M. Dolha, and M. Beetz, "Towards 3d point cloud based object maps for household environments," *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 927–941, 2008.

[20] D. Pfeiffer, S. Gehrig, and N. Schneider, "Exploiting the power of stereo confidences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 297–304, 2013.

[21] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," *ACM siggraph computer graphics*, vol. 21, no. 4, pp. 163–169, 1987.

[22] K. Daniilidis, "Hand-eye calibration using dual quaternions," *The International Journal of Robotics Research*, vol. 18, no. 3, pp. 286–298, 1999.

[23] J. R. Hodgson, P. Kinnell, L. Justham, N. Lohse, and M. R. Jackson, "Novel metrics and methodology for the characterisation of 3d imaging systems," *Optics and Lasers in Engineering*, vol. 91, pp. 169–177, 2017.

[24] "Aesub blue: Vanishing 3d scanning spray." `https://aesub.com`.

[25] T. Duhautbout, J. Moras, and J. Marzat, "Distributed 3d tsdf manifold mapping for multi-robot systems," in *2019 European Conference on Mobile Robots (ECMR)*, pp. 1–8, IEEE, 2019.

[26] M. Fehr, F. Furrer, I. Dryanovski, J. Sturm, I. Gilitschenski, R. Siegwart, and C. Cadena, "Tsdf-based change detection for consistent long-term dense reconstruction and dynamic object discovery," in *2017 IEEE International Conference on Robotics and automation (ICRA)*, pp. 5237–5244, IEEE, 2017.