Touch if it's transparent!
ACTOR: Active Tactile-based Category-Level Transparent Object Reconstruction

Prajval Kumar Murali, Bernd Porr, and Mohsen Kaboli

***Abstract*— Accurate shape reconstruction of transparent objects is a challenging task due to their non-Lambertian surfaces and yet necessary for robots for accurate pose perception and safe manipulation. As vision-based sensing can produce erroneous measurements for transparent objects, the tactile modality is not sensitive to object transparency and can be used for reconstructing the object's shape. We propose *ACTOR*, a novel framework for *AC*tive tactile-based category-level *T*ransparent *O*bject *R*econstruction. ACTOR leverages large datasets of synthetic object with our proposed self-supervised learning approach for object shape reconstruction as the collection of real-world tactile data is prohibitively expensive. ACTOR can be used during inference with tactile data from category-level unknown transparent objects for reconstruction. Furthermore, we propose an active-tactile object exploration strategy as probing every part of the object surface can be sample inefficient. We also demonstrate tactile-based category-level object pose estimation task using ACTOR. We perform an extensive evaluation of our proposed methodology with real-world robotic experiments with comprehensive comparison studies with state-of-the-art approaches. Our proposed method outperforms these approaches in terms of tactile-based object reconstruction and object pose estimation.**

I. INTRODUCTION

Transparent objects such as cups, glasses, and bottles are ubiquitous around us and if robots are expected to work in unstructured scenarios such as household environments, it is essential to recognize and safely manipulate transparent objects. Reconstruction of the object shape is critical for detecting and identifying its pose and safely manipulating it [1]. While this is straightforward for opaque objects with off-the-shelf vision sensors, such sensors produce unreliable and erroneous data with transparent objects due to their non-Lambertian surfaces. Sophisticated custom calibrated setups with specialized scanners or modifying the transparent surface of objects are often necessary for accurate reconstruction [2, 3]. This is impractical for on-the-fly reconstruction of arbitrary unknown objects. On the contrary, high fidelity tactile sensing can be used for shape reconstruction of transparent objects as well as pose estimation and safe-manipulation [4–11].

Tactile perception is inherently action-conditioned as data depends on the type of contact action performed and local as only the local surface information around the contact area is extracted [12]. Hence, for reconstructing the surfaces of an object, multiple contact actions need to be performed by the robot. This leads to sparse information and prohibitively long data collection times. Early works have used offline methods to collect dense tactile data and used shape-fitting primitives such as superquadratics [13]. Aggregating contact points into a point cloud is often used to represent the shape of the objects. Some works have used Bayesian filtering techniques for defining a probabilistic model of the objects using the tactile point clouds and used them for other tasks such as classification [14]. Gaussian process implicit surface (GPIS) has been widely used for tactile object reconstruction [15–21]. The implicit surface

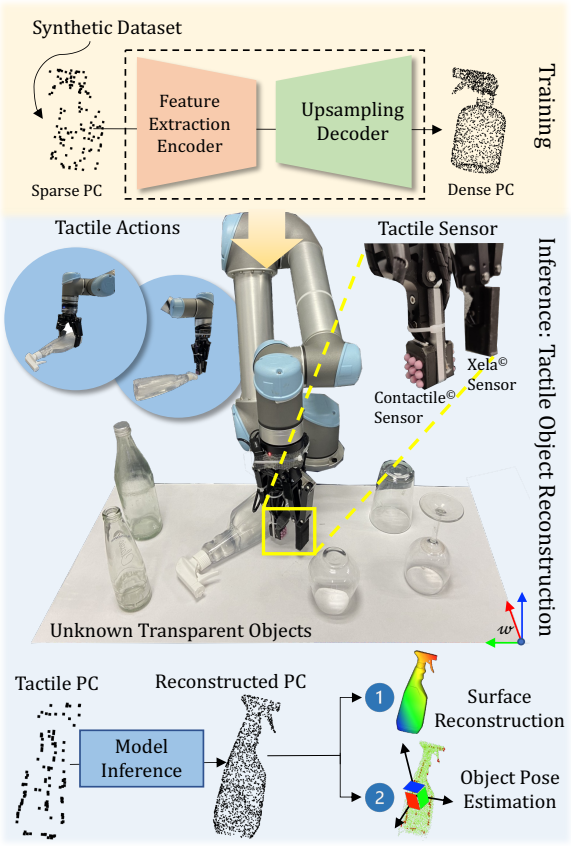


Fig. 1: Experimental Setup: A Universal Robots UR5 with sensorised Robotiq Gripper with 3-axis tactile sensor arrays for active tactile-based category-level unknown transparent object reconstruction.

P.K.Murali and M.Kaboli are with the BMW Group, Munich Germany. e-mail: name.surname@bmwgroup.com

P.K. Murali and B. Porr are with the University of Glasgow, Scotland

M. Kaboli is with the Donders Institute for Brain and Cognition, Radboud University, Netherlands

Funded in part by the BMW Group, EU H2020 INTUITIVE under Grant ID 861166 and EU Horizon PHASTRAC under Grant ID 101092096.

described by a Gaussian process describes the shape of an object through a function that decides for each point in space whether it is part of the object or not. It produces smooth surface manifolds with a reasonable number of tactile points as input and also provides probabilistic information to guide the tactile actions. However, for complex shapes it typically requires lots of points uniformly distributed on the object's surface for reconstruction [21]. Some works have also used tactile sensing with visual perception in order to perform shape completion with prior information observed with visual cameras [18, 22]. While these works focus on opaque objects, limited works exist for the reconstruction of transparent objects. Recently, deep learning methods have been used for point cloud based shape completion given partial or noisy input point clouds [23, 24]. Seminal works on PointNet [25] allowed using raw point clouds as inputs to deep networks for the task of classification and semantic segmentation. Prior works have worked towards point cloud completion using deep networks such as [26, 27] but are mainly evaluated on datasets derived from CAD models and rarely evaluated on real-world platforms with noisy and sparse sensors [23].

Using the constructed object shape for pose estimation of a transparent object through tactile sensing brings further challenges due to the nature of the tactile data. Typical pose-estimation methods for visual perception perform poorly with tactile data as they are sparse and extracted sequentially through contact probing [1, 4, 24, 28–30]. In summary, there are limitations in the state-of-the-art for the reconstruction and further applications such as pose estimation of transparent objects with tactile perception: (a) existing reconstruction strategies such as GPIS fail to capture fine shape details with sparse tactile input data, (b) directly deploying deep learning based strategies for shape completion with sparse input data is impractical as the collection of a large dataset of tactile data for training is prohibitively expensive, (c) existing tactile-based pose estimation techniques rely upon known object models or shape primitives but category-level tactile-based pose estimation wherein objects without *a priori* known CAD models but belong to a known category is necessary.

Contributions:

(I) We propose *ACTOR*, a novel framework for deep active tactile-based category-level perception of unknown transparent objects for reconstruction and pose estimation. Our proposed network is trained on a category-level synthetic dataset and tested on sparse tactile point clouds from real unknown transparent objects.

(II) Our proposed network consists of a feature-extraction encoder with self-attention and an upsampling decoder for accurate reconstruction of sparse input point clouds.

(III) We propose an autonomous and active tactile-based unknown object exploration strategy based on information gain.

(IV) We improve our previously presented novel Translation-Invariant Quaternion Filter (TIQF) [31] to category-level pose (6DoF) and scale (3DoF) estimation and relax the need of a prior known model of the object.

To validate our proposed framework, we perform extensive experiments on a real robotic setup and provide baseline comparisons with state-of-the-art methods for tactile-based object reconstruction and pose estimation.

II. METHODS

A. Problem Definition and Proposed Framework

The objective is to reconstruct a dense point cloud that precisely represents the shape of unknown transparent objects from sparse point clouds extracted with active tactile interactive perception. To this end, we propose a novel framework termed ACTOR shown in Fig. 2. In Fig. 2(a) we propose a self-surpervised learning approach with an autoencoder network that is trained on subsampled pointclouds from synthetic objects belonging to the same category but not identical as the real objects. In Fig. 2(b), we propose a novel active tactile-based unknown transparent object exploration strategy which is used for inference with our trained model to reconstruct a dense point cloud. We demonstrate downstream tasks such as tactile-based pose estimation.

B. Deep Self-Supervised Learning for 3D Object Reconstruction

We generate a dataset $\mathcal{D}$[1] of synthetic object models from the ShapeNet repository [32] in order to leverage the open-source datasets and avoid expensive real tactile-data collection. The synthetic object models belong to the same category but are different from the real unknown transparent objects. We uniformly sample $N_{in} = 2048$ points from the synthetic object meshes. These pointclouds are normalized and scaled to fit into a $[0,1]^3$ cube and added to the dataset, $\mathcal{P}_{in} \in \mathcal{D}$. In order to generate the input point clouds $\mathcal{P}^{\bullet}_{in}$ to the network, we randomly subsample the $\mathcal{P}_{in}$ by voxel-grid subsampling by the factor k i.e., $\mathcal{P}^{\bullet}_{in} \in \mathbb{R}^{\lceil \frac{1}{k} N_{in} \rceil \times 3}$. This creates a challenging task for reconstruction with higher values for k as simpler techniques based on interpolation with neighborhood points cannot be used.

Feature-Extraction Encoder: The network architecture shown in Figure 2(a) is proposed as an autoencoder (AE) that uses a self-supervised approach to reconstruct the original point cloud from a subsampled point cloud. The encoder takes subsampled point clouds as inputs and generates a high dimensional feature vector. The feature vector captures the global geometric shape information of the input point cloud. In general, any deep network that works on raw input point clouds to provide a high dimensional feature vector can be used as an encoder. In particular, we use a modified PointNet architecture [25] for the encoder. PointNet takes unordered point clouds and generates a global feature descriptor vector of size 1024. The network learns a set of optimization functions that select interesting or informative points of the

[1] https://www.robotact.de/tactile-reconstruction

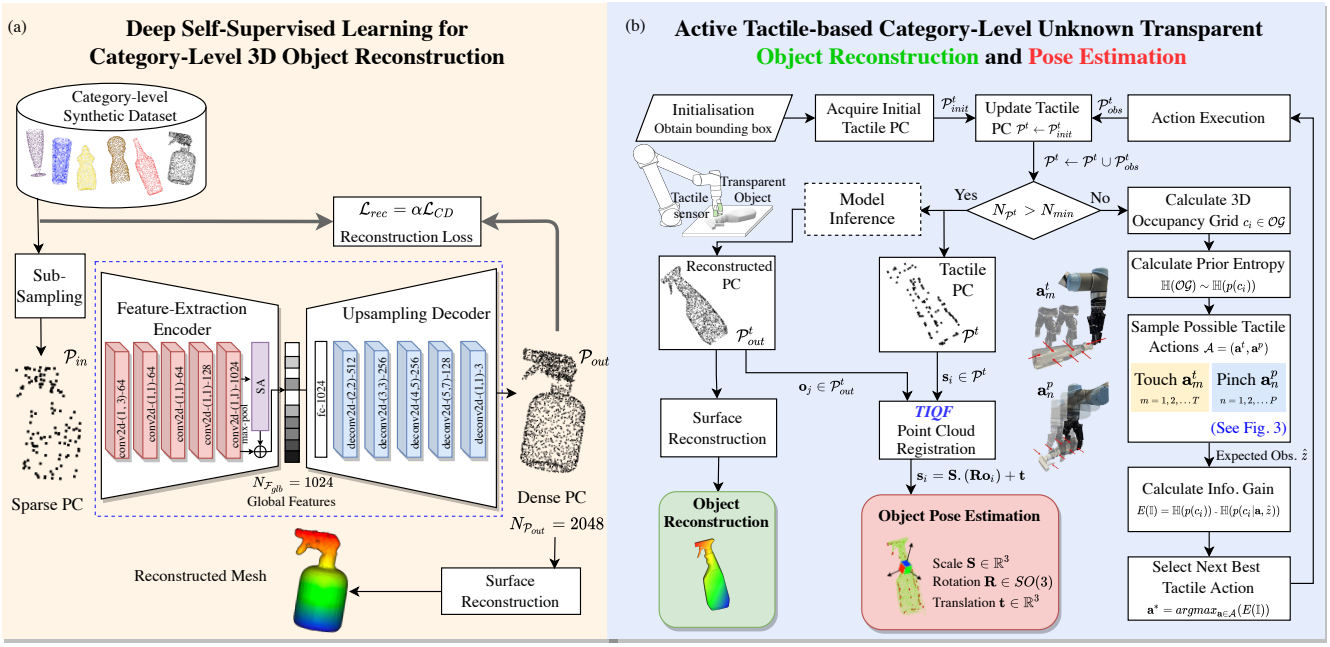


Fig. 2: Proposed framework ACTOR: Active Tactile-based Category-Level Transparent Object Reconstruction.

point cloud. The encoder consists of $[1 \times 1]$ convolutions with output channels size $(64, 64, 128, 1024)$ with the first convolutional layer with kernel size $[1 \times 3]$ to encode the input pointcloud of $N \times 3$ dimension. The convolution layers are aggregated by a max-pooling layer. We introduce a self-attention layer [33] whose outputs are aggregated with the max-pooled features to provide the global feature vector. We have summarized the encoder in Figure 2(a).

Self-Attention (SA) Layer: The SA layer is introduced as it can encode meaningful spatial relationships between features and focus on important local features. From the input layer ($\texttt{conv2d} - 1024$), two separate multi-layer perceptrons (MLPs) are used to get features $\mathbf{G}$ and $\mathbf{H}$ which are subsequently used to get the weights as $\mathbf{W} = softmax(\mathbf{G}^T\mathbf{H})$. The input features are transformed using another MLP to obtain $\mathbf{K}$ and multiplied with the weights as $\mathbf{W}^T\mathbf{K}$. These vectors are summed with the input vector to produce the output features.

Upsampling Decoder: We design an upsampling decoder that upsamples the input global feature vector to provide the reconstructed dense output point cloud $\mathcal{P}_{out}$. The upsampling decoder is composed by a fully connected layer with output dimension of 1024 and five deconvolutional layers with kernel sizes and output channels shown in Fig. 2(a). The decoder produces the output point cloud with point size set to 2048 while training as this is sufficiently dense for reconstruction purposes.

Loss Function: In order to encourage the upsampled point cloud to be in proximity to the original input point cloud and follow the underlying geometrical surface of the object, we use the Chamfer distance metric [34] as the loss. Given the input point cloud prior to subsampling, $\mathcal{P}_{in}$ and the reconstructed output point cloud $\mathcal{P}_{out}$, the loss is defined as:

$$\mathcal{L}_{CD}(\mathcal{P}_{in}, \mathcal{P}_{out}) = \frac{1}{|\mathcal{P}_{in}|} \sum_{p_1 \in \mathcal{P}_{in}} \min_{p_2 \in \mathcal{P}_{out}} ||p_1 - p_2||_2 + \frac{1}{|\mathcal{P}_{out}|} \sum_{p_2 \in \mathcal{P}_{out}} \min_{p_1 \in \mathcal{P}_{in}} ||p_2 - p_1||_2, \quad (1)$$

where $|\bullet|$ refers to the number of points in the point cloud and $||\bullet||_2$ refers to the L2 norm. The loss $\mathcal{L}_{CD}$ represents the average distance between the *closest* points in the two point clouds. We use the weighted loss for learning stability as the reconstruction loss $\mathcal{L}_{rec} = \alpha\mathcal{L}_{CD}$ with $\alpha = 100$ set empirically. For surface reconstruction from the dense reconstructed point cloud, we use the ball-pivoting algorithm [35].

C. Active Deep Tactile-based Unknown Transparent Object Reconstruction and Pose Estimation

1) Active Tactile-based Transparent Object Reconstruction: The model trained with only *synthetic data* as described in Sec. II-B is used during the inference with *real-world* transparent objects. The sparse tactile point cloud data is collected autonomously by the robot using an information gain-based active strategy. We define two types of tactile actions for data acquisition: touch and pinch actions as shown in Figure 3. The touch action is executed as a guarded horizontal straight-line motion wherein the object is not moved upon contact. The touch action is defined by a tuple $\mathbf{a}^t = \{\mathbf{s}^t, \overrightarrow{\mathbf{d}^t}\}$ where $\mathbf{s}^t \in \mathbb{R}^3$ is the start point of the tactile-sensorised gripper and $\overrightarrow{\mathbf{d}^t} \in \mathbb{R}^3$ is the direction of the gripper-motion defined in the world-coordinate frame $\mathcal{W}$. During the pinch action the robot approaches the object in a vertical straight-line motion with a completely open gripper and performs an antipodal enclosure grasp on the object. The

fingers of the gripper are closed until the force on the tactile sensors exceeds a predefined threshold. The pinch action is characterized by $\mathbf{a}^p = \{\mathbf{s}^p\}$ where $\mathbf{s}^p \in \mathbb{R}^3$ is the start position of the gripper motion vertically above the object at a predefined height as shown in Figure 3. Given the 2D bounding box of the object (a priori known or through a RGB camera), a probabilistic occupancy grid $\mathcal{OG}_i$ of preset height and resolution og_{res} is defined. Each cell of the occupancy grid c_i is represented by an occupancy probability $p(c_i)$ which is initially set to 0.5. During exploration, if a cell is discovered to belong to the object, the probability is set to 1 and similarly, if the cell belongs to free space, the probability is set to 0. The probabilities are updated through ray intersections based on the virtual sensor model. We define a virtual sensor model of the tactile sensor which casts a set of rays $\mathcal{R} = \{r_1, r_2, \ldots, r_{n_{taxel}}\}$ where n_{taxel} refers to the number of taxels in the sensor array. The independence assumption of the probability of each grid cell with one another allows us to calculate the overall entropy of the $\mathcal{OG}$ as the summation of the entropy of each cell. The Shannon entropy of the overall occupancy grid is calculated as:

$$\mathbb{H}(\mathcal{OG}) = \sum_{c_i \in \mathcal{OG}} p(c_i) log(p(c_i)) + (1 - p(c_i))(1 - log(p(c_i))). \tag{2}$$

Monte-Carlo sampling of possible tactile actions N_{nbt} are performed for computing the next best tactile (NBT) action. The actions space $\mathcal{A}_{nbt}$ is comprised of an equal number of touch and pinch respectively as $\mathcal{A}_{nbt} = \{a^p, a^t\}_{N_{nbt}}$. The expected measurements $\hat{\mathbf{z}}_t$ for each action $a_t \in \mathcal{A}$ is computed using ray-traversal algorithms [36]. Given the observed grid cell c and the measurement from sensor observation z, the log-odds is updated as $L(c|z) = L(c) + l(z)$ wherein $L(c) = log \frac{p(c)}{1-p(c)}$ and

$$l(z) = \begin{cases} log \frac{p_h}{1-p_h} & \text{if } z \widehat{=} \textit{ hit} \\ log \frac{p_m}{1-p_m} & \text{if } z \widehat{=} \textit{ miss} \end{cases} \tag{3}$$

where p_h and p_m are the probabilities of hit and miss which are user-defined values set to 0.7 and 0.4 respectively as in [36]. The posterior probability $p(c|z)$ can be computed by inverting $L(c|z)$. The expected information gain by taking an action $a_t \in \mathcal{A}_{nbt}$ with expected measurement $\hat{\mathbf{z}}_t$ is provided by the Kullback-Liebler divergence of the posterior entropy and the prior entropy as:

$$E[\mathbb{I}(p(c_i|\mathbf{a}_t, \hat{z}_t))] = \mathbb{H}(p(c_i)) - \mathbb{H}(p(c_i|\mathbf{a}_t, \hat{z}_t)) \tag{4}$$

Therefore, the action that maximizes the expected information gain is considered as the NBT action:

$$\mathbf{a}_t^{nbt*} = \arg\max_{\mathbf{a} \in \mathcal{A}} (E[\mathbb{I}(p(c_i|\mathbf{a}_t, \hat{z}_t))]) \tag{5}$$

Each tactile action extracts contact positions in 3D space and contact forces. The direction of the normal force is used to extract the normal direction $\hat{n}$ of the object surface. The contact points are aggregated into the tactile point cloud $\mathcal{P}^t$. In order to initialize the NBT action calculation, an initial point cloud (with $N_{\mathcal{P}^t} = 20$) is required, which is extracted by

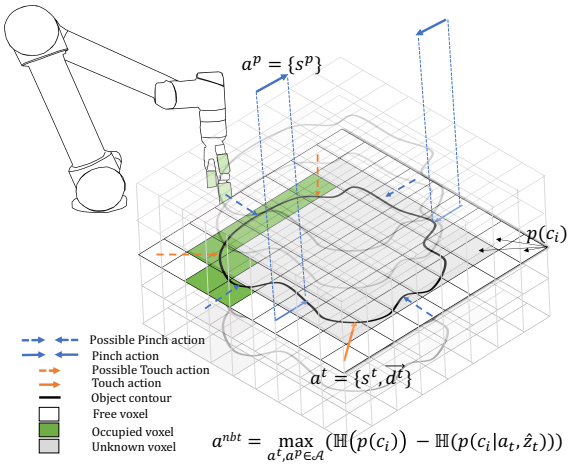


Fig. 3: Action selection voxelised probabilistic occupancy grid.

randomised touch actions. Further points are collected in an active manner using the NBT criteria. A minimum number of points in the tactile point cloud is required to perform model inference $N_{\mathcal{P}^t} > N_{min}$ which is tuned empirically. The tactile point cloud is provided as input to the trained network and the reconstructed point cloud $\mathcal{P}_{out}$ is obtained .

2) Tactile-Based Object Pose Estimation: We perform the 6D pose estimation through dense to sparse point cloud registration. The sparse scene point cloud $\mathbf{s}_i \in \mathcal{S}$ is represented by the tactile points and the dense object point cloud $\mathbf{o}_i \in \mathcal{O}$ is represented by the reconstructed point cloud in II-B without the need for the object model. Point cloud registration problem with M known correspondences can be formulated as:

$$\mathbf{s}_i = \mathbf{S} \cdot (\mathbf{R}\mathbf{o}_i) + \mathbf{t} \quad i = 1, \ldots M, \tag{6}$$

where $\mathbf{S} \in \mathbb{R}^3$ represents scale, $\mathbf{R} \in SO(3)$ represents rotation and $\mathbf{t} \in \mathbb{R}^3$ represents translation which are unknown and to be estimated and $\cdot$ is the element-wise product.

We perform the point cloud registration using our novel translation-invariant Quaternion filter (TIQF) presented in [31] to determine $\mathbf{R}$, $\mathbf{S}$ and $\mathbf{t}$. The scale, rotation and translation are decoupled by finding the relative vectors between corresponding points, i.e., $\forall o_i, o_j \in \mathcal{O}, s_i, s_j \in \mathcal{S}$ the relative vectors are $\mathbf{s}_{ji} = \mathbf{s}_j - \mathbf{s}_i$ and $\mathbf{o}_{ji} = \mathbf{o}_j - \mathbf{o}_i$. Equation (6) is reformulated as:

$$\mathbf{s}_j - \mathbf{s}_i = (\mathbf{S} \cdot \mathbf{R}\mathbf{o}_j + \mathbf{t}) - (\mathbf{S} \cdot \mathbf{R}\mathbf{o}_i + \mathbf{t}), \tag{7}$$

$$\mathbf{s}_{ji} = \mathbf{S} \cdot \mathbf{R}\mathbf{o}_{ji} \quad . \tag{8}$$

We note that equation (8) is independent of translation. Taking the L2-norm on both sides for Eq. (8) and recalling that norm is rotation invariant we get:

$$||\mathbf{s}||_{ji} = ||\mathbf{S}|| \cdot ||\mathbf{o}||_{ji} \quad . \tag{9}$$

The scale $\mathbf{S}$ is estimated by taking the ratio of the axis aligned bounding box (AABB) of the scene and object point clouds, i.e., if $\mathcal{X}_{AABB} = \{(x_{min}, x_{max}), (y_{min}, y_{max}), (z_{min}, z_{max})\}$

represents the AABB for a point cloud $\mathcal{X}$, then:

$$\mathbf{S} = \{\frac{|x_{max}-x_{min}|_{\mathcal{S}}}{|x_{max}-x_{min}|_{\mathcal{O}}}, \frac{|y_{max}-y_{min}|_{\mathcal{S}}}{|y_{max}-y_{min}|_{\mathcal{O}}}, \frac{|z_{max}-z_{min}|_{\mathcal{S}}}{|z_{max}-z_{min}|_{\mathcal{O}}}\} \quad (10)$$

Using the estimated scale and using $\tilde{\mathbf{o}}_{ji} = \mathbf{S}\mathbf{o}_{ji}$ for convenience we are left with a pure rotation to estimate:

$$\tilde{\mathbf{s}}_{ji} = \mathbf{R}\tilde{\mathbf{o}}_{ji} \quad . \quad (11)$$

We cast the rotation estimation problem into a recursive Bayesian estimation framework and derive a linear state and measurement model. Reformulating Eq.(11) using quaternions we get:

$$\bar{\mathbf{s}}_{ji} = \mathbf{x} \odot \bar{\mathbf{o}}_{ji} \odot \mathbf{x}^*, \quad (12)$$

where $\mathbf{x}$ is the quaternion form of $\mathbf{R}$, $\odot$ is the quaternion product, $\mathbf{x}^*$ is the conjugate of $\mathbf{x}$, and $\bar{\mathbf{s}}_{ji} = \{0, \tilde{\mathbf{s}}_{ji}\}$ and $\bar{\mathbf{o}}_{ji} = \{0, \tilde{\mathbf{o}}_{ji}\}$. Using the matrix form of quaternion product, we can rewrite Eq.(12) as:

$$\begin{bmatrix} 0 & -\tilde{\mathbf{s}}_{ji}^T \\ \tilde{\mathbf{s}}_{ji} & \tilde{\mathbf{s}}_{ji}^{\times} \end{bmatrix} \mathbf{x} - \begin{bmatrix} 0 & -\tilde{\mathbf{o}}_{ji}^T \\ \tilde{\mathbf{o}}_{ji} & -\tilde{\mathbf{o}}_{ji}^{\times} \end{bmatrix} \mathbf{x} = \mathbf{0} \quad (13)$$

$$\underbrace{\begin{bmatrix} 0 & -(\tilde{\mathbf{s}}_{ji} - \tilde{\mathbf{o}}_{ij})^T \\ (\tilde{\mathbf{s}}_{ji} - \tilde{\mathbf{o}}_{ji}) & (\tilde{\mathbf{s}}_j + \tilde{\mathbf{s}}_i + \tilde{\mathbf{o}}_j + \tilde{\mathbf{o}}_i)^{\times} \end{bmatrix}_{4\times 4}}_{\mathbf{H}_t} \mathbf{x} = \mathbf{0} \quad , \quad (14)$$

where $(\ )^{\times}$ denotes the skew-symmetric matrix formulation. Equation (14) is of the form $\mathbf{H}_t\mathbf{x} = 0$ where $\mathbf{H}_t$ is the pseudo-measurement matrix [37]. We note that Eq. (14) represents a noise-free state estimation where $\mathbf{H}_t$ depends only on sparse and dense point correspondences which are $\tilde{\mathbf{s}}_{ji}$ and $\tilde{\mathbf{o}}_{ji}$. We design a pseudo-measurement model as $\mathbf{H}_t\mathbf{x} = \mathbf{z}^h$ and set $\mathbf{z}^h = 0$. Since we have a static process model, the object does not move and $\mathbf{x}$ and $\mathbf{z}_t$ are Gaussian distributed, the state $\mathbf{x}_t$ and covariance matrix $\Sigma_t^{\mathbf{x}}$ at each timestep t are computed through a linear Kalman filter. The Kalman filter equations are skipped for brevity and a in-depth derivation is provided in our prior work [31]. As the Kalman filter does not implicitly ensure the constraints on the quaternion as $||\mathbf{x}|| = 1$, we normalise the state and uncertainty after each update step as $\bar{\mathbf{x}}_t = \frac{\mathbf{x}_t}{||\mathbf{x}_t||_2}$ $\quad, \bar{\Sigma}_t^{\mathbf{x}} = \frac{\Sigma_t^{\mathbf{x}}}{||\mathbf{x}_t||_2^2}$. We convert the estimated rotation $\bar{\mathbf{x}}_t$ to its equivalent rotation matrix $\mathbf{R}$. It used to estimate the translation using the following relation: $\mathbf{t} = \frac{1}{N}\sum_{i=0}^{N}(\bar{\mathbf{s}}_i - \mathbf{R}\bar{\mathbf{o}}_i)$. At each iteration, a rotation and translation estimate is found which is used to transform the object point cloud and the process is repeated by re-estimating the correspondence points. The convergence criteria are set by (a) maximum number of iterations or (b) the relative change in estimated pose parameters is less than a predefined threshold ($0.1mm$ and 0.1^o).

III. EXPERIMENTAL RESULTS

A. Experimental Setup

The experimental setup is shown in Fig. 1 consists of a set of 9 unknown transparent objects belonging to six categories and a Universal Robots UR5 equipped with a sensorised Robotiq 2F140 gripper. The tactile sensor array of the two-finger gripper are sourced from XELA robotics© and Contactile©. The outer and inner side of each finger are sensorised and comprise of 3×3 sensor array from the Contactile sensors and 4×4 sensor array from XELA sensors respectively. The fingertip of the finger sensorised with the XELA sensors also has 6×1 array. Each taxel of the sensor array provides 3-axis force measurements. The normalised force values of the tactile sensors are measured and contact is established when the force exceeds the baseline threshold $f_{ts} \geq \tau_f$ where $\tau_f = 1.1$. All operations involving point clouds use the Point Cloud Library[2], occupancy grid computations uses Octomap library[3], and the overall setup uses a ROS-based framework[4]. All robot experiments are run on a workstation using Ubuntu 18.04 with Intel©Xeon(R) Gold 5222 CPU. The object exploration and reconstruction time is between 5-7 minutes on average as the robot's maximum speed is limited to 250 mm/s for safety regulations.

Network Implementation Details: Our proposed network is implemented using the Tensorflow framework and training/ inference are performed on Nvidia Quadro RTX 4000 GPU. We used the ADAM optimiser, learning rate set to 10^{-4}, momentum 0.9 and batch size 8. All layers of the encoder-decoder uses batch normalisation and the decay rate initialized at 0.5 and gradually increased to 0.99 with decay step size 2×10^5. During training with our synthetic dataset $\mathcal{D}$, random voxel-grid subsampling is done to have input point clouds with point size between 40 and 120.

Object List: We use the following widely-available transparent objects as unknown objects: bottle 1, bottle 2, can, detergent, cup 1, cup 2, cup 3, wineglass and spray as shown in Tab. I.

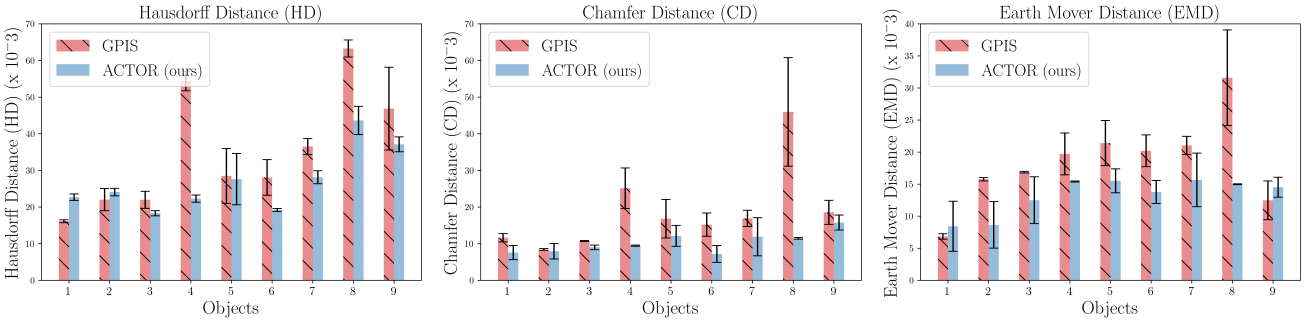
B. Active Tactile-based Deep Self-Supervised Category-level Transparent Object Reconstruction

The height of the occupancy grid is set constant for every object at 0.4m which is larger than the biggest object. Reconstruction with acceptable accuracy is obtained with 100 points or more as input. For each object, ten tactile point clouds with point number between 100 and 120 points are extracted using the active exploration strategy and used for reconstruction. The ground-truth point cloud and CAD mesh are obtained by spray-painting the objects and using a scanning device. For evaluation, we use the following performance metrics: Hausdorff distance (HD), Chamfer distance (CD) and Earth Mover distance (EMD). The CD is described in Sec. II-B. Given two points S_1 and S_2, the Hausdorff distance is defined as [38]:

$$HD(S_1, S_2) = \max\{\max_{x\in S_1}\min_{y\in S_2}\{||x-y||_2\}, \max_{y\in S_2}\min_{x\in S_1}\{||y-x||_2\}\} \quad (15)$$

The HD represents the maximum distance between the two point sets and can be affected by extreme outliers during the reconstruction. The EMD finds a bijection $\phi : S_1 \to S_2$ to minimise the average distance between corresponding points

[2] https://pointclouds.org/
[3] https://octomap.github.io/
[4] https://www.ros.org/



(a) Hausdorff distance (HD) (b) Chamfer distance (CD) (c) Earth mover distance (EMD)

Fig. 4: Quantitative reconstruction results. Object numbered as follows: {1: Bottle 1, 2: Bottle 2, 3: Can, 4: Detergent, 5: Cup 1, 6: Cup 2, 7: Cup 3, 8: Wineglass, 9: Spray }

TABLE I: Qualitative reconstruction results of our proposed method in comparison with Gaussian process implicit surfaces for unknown real test objects. (Best viewed on screen in color).

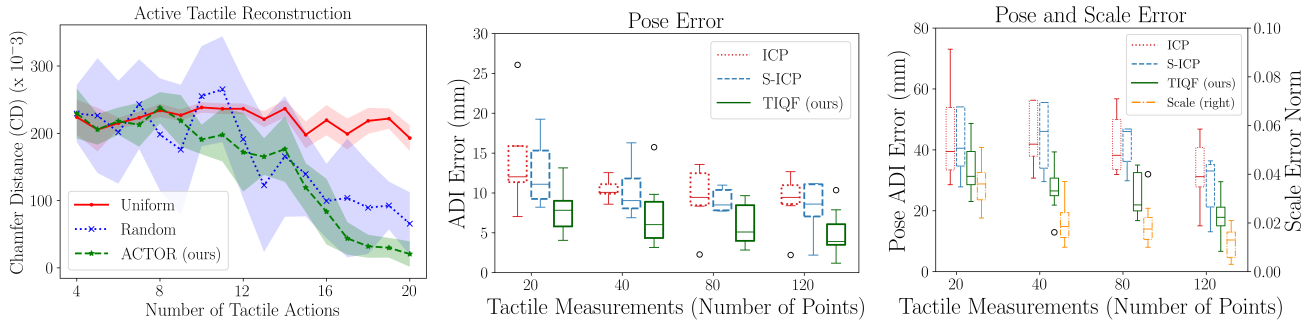
Object	Tactile PC	Ground Truth		GPIS		ACTOR (ours)	
	N ∼120	PC	Surface	Recon. PC	Recon. Surf.	Recon. PC	Recon. Surf.
Bottle 1							
Bottle 2							
Can							
Detergent							
Cup 1							
Cup 2							
Cup 3							
Wineglass							
Spray							

in the point clouds as:

$$EMD(S_1,S_2) = \min_{\phi:S_1\to S_2} \frac{1}{|S_1|}\sum_{x\in S_1} ||x-\phi(x)||_2 \quad . \tag{16}$$

A perfect reconstruction will yield $\{CD, HD, EMD\} \to 0$ and lower values signify better reconstruction.

We use Gaussian Process Implicit Surfaces (GPIS) as baseline as it is widely used in the literature for tactile-based object reconstruction [15–21]. For implementation, we utilise the GP for machine learning toolbox [39] in MATLAB and the Matérn kernel.



(a) (b) (c)

Fig. 5: (a) Active tactile reconstruction accuracy evaluated using the chamfer distance with ground-truth, (b) Pose estimation with ground truth object point cloud from CAD mesh, (c) Pose and scale estimation with reconstructed point cloud as object point cloud

The quantitative results of tactile-based reconstruction using our method and baseline GPIS method are shown in Fig. 4 and qualitative reconstruction results are presented in Tab. I. From Fig. 4, we note our proposed approach yields lower CD values for all objects. For HD and EMD, apart from the bottle and spray, our method performs better than the baseline approach. On average, our approach is 45%, 23.5% and 28% lower in CD, HD and EMD values compared to baseline GPIS. While the quantitative results focus on local point-distances between the reconstructed and ground-truth point cloud, the qualitative results in Tab. I demonstrate the differences in reconstruction accuracy at the object level. GPIS produces warped reconstructed surfaces due to the low number of tactile points. Whereas our method, with the help of the learned model over the category-level synthetic objects, is able to reconstruct the object to an acceptable accuracy even with sparse input data.

Active Tactile Reconstruction: Using our proposed framework ACTOR, we can achieve accurate reconstruction with fewer tactile actions in comparison to the baselines as shown in Fig. 5a. We define an uniform object exploration and random object exploration strategy as baselines as follows: the bounding box around the object is transformed into a grid with each grid cell of size 3cm $\times$ 3cm (size of the sensor patch). The grid does not encode the probabilistic occupancy as in our ACTOR approach. The robot explores each grid cell in a sequential manner in the uniform strategy. In contrast, for the random strategy, the robot picks a grid cell at random for exploration. In order to have an unbiased comparison between the exploration methods, a maximum of 20 actions are chosen as on average it takes 20 actions to extract atleast 100 tactile points. We begin the model inference from the 4th action onwards to have a minimum of 20 points in the tactile point cloud. We note that the uniform strategy requires a large number of tactile actions to completely explore the object in order for reconstruction. The random strategy has high variance in terms of reconstruction accuracy and stems from the stochastic nature of the exploration while ACTOR deterministically improves reconstruction accuracy with the increasing number of tactile actions.

C. Tactile-based Transparent Object Pose Estimation

As the error in reconstruction propagates to downstream tasks, we perform two experiments: firstly, instance-level estimation using the ground-truth model point cloud as the object point cloud (Fig. 5b) and secondly, category-level pose estimation using the reconstructed point cloud as the object point cloud (Fig. 5c). For category-level pose estimation, norm scale error is also reported in addition to rotation and translation. As our proposed TIQF method is a local registration method, we chose the standard Iterative Closest Point (ICP) [40] and Sparse Iterative Closest Point (S-ICP) [41] as baselines. S-ICP is chosen as it demonstrates higher robustness to outliers and incomplete data as typically found in tactile point clouds. We use the Average Distance of model points with Indistinguishable views metric (ADI) [42] as a combined measure of the rotation and translational error as we have multiple objects with axis of symmetry. The ADI metric is defined as:

$$\texttt{err}_{adi} = \frac{1}{|\mathcal{O}|} \sum_{\mathbf{p}_1 \in \mathcal{O}} \min_{\mathbf{p}_2 \in \mathcal{O}} ||(\mathbf{R}_{gt}\mathbf{p}_1 + \mathbf{t}_{gt}) - (\mathbf{R}_{est}\mathbf{p}_2 + \mathbf{t}_{est})||, \tag{17}$$

where $(\mathbf{R}_{gt}, \mathbf{t}_{gt})$ and $(\mathbf{R}_{est}, \mathbf{t}_{est})$ refers to ground-truth and estimated rotation and translation respectively. As seen from Fig. 5b,5c, our proposed approach outperforms the baseline approaches for all input tactile point clouds with varying point numbers demonstrating robustness to point sparsity. The median $\texttt{err}_{adi} < 1cm$ for our proposed approach even with sparse point clouds with $N_{P'} = 20$ and improves with increasing the number of points. The category-level pose estimation errors are higher than instance-level due to the errors in the reconstructed point clouds. However, the accuracy improves by reducing scale error with median $\texttt{err}_{adi} < 2cm$ for $N_{P'} = 120$ with our proposed method.

D. Discussion

Our proposed approach, ACTOR outperforms the GPIS strategy by all our evaluation metrics. We also note the qualitative reconstruction results in Tab. I, wherein GPIS fails to capture the shape details of the object while our approach captures the global and local shape accurately (see object spray and wineglass). Our network implicitly

learns important feature points and is able to reconstruct the object accurately given few sparse inputs. Our active exploration strategy converges faster to reconstruct the object shape thus improving the sample efficiency. Furthermore, our proposed category-level pose estimation method outperforms the baseline methods by $\geq 25\%$ ADI error.

A limitation of the work is the need for category-wise object models for training. A possible future work includes using neural radiance fields (NeRFs) [43] to generate synthetic models of objects from images that can be used for training.

IV. CONCLUSIONS

In this work we proposed ACTOR, a novel framework for active tactile-based category-level transparent object reconstruction. By learning with only synthetic object models, ACTOR is capable of performing real-world transparent object reconstruction through sparse tactile data. Our approach outperforms state-of-the-art GPIS method in terms of reconstruction accuracy. Furthermore, we demonstrated category-level pose estimation with the reconstructed object model and our approach outperforms baseline ICP and S-ICP methods. As future work, we would like to extend ACTOR for safely manipulating transparent objects in unstructured scenarios with possible deformability and dynamic center of mass [11, 44, 45].

REFERENCES

[1] Q. Li *et al.*, "A review of tactile information: Perception and action through touch," *IEEE Trans. on Rob.*, 2020.

[2] I. Ihrke *et al.*, "Transparent and specular object reconstruction," in *Computer Graphics Forum*, 2010.

[3] Z. Li *et al.*, "Through the looking glass: neural 3d reconstruction of transparent shapes," in *IEEE/CVF Conf. on Comp. Vis. and Pat. Rec.*, 2020, pp. 1262–1271.

[4] P. K. Murali *et al.*, "Active visuo-tactile point cloud registration for accurate pose estimation of objects in an unknown workspace," in *2021 IEEEInt. Conf. on Int. Rob. and Sys.* IEEE, 2021.

[5] P. K. Murali *et al.*, "Intelligent in-vehicle interaction technologies," *Adv. Int. Sys.*, vol. 4, no. 2, p. 2100122, 2022.

[6] M. Kaboli *et al.*, "Tactile-based active object discrimination and target object search in an unknown workspace," *Auto. Rob.*, 2019.

[7] M. Kaboli *et al.*, "Active tactile transfer learning for object discrimination in an unstructured environment using multimodal robotic skin," *Int. Jour. of Hum. Rob.*, vol. 15, no. 01, p. 1850001, 2018.

[8] M. Kaboli *et al.*, "A tactile-based framework for active object learning and discrimination using multimodal robotic skin," *IEEE Rob. and Aut. Let.*, vol. 2, no. 4, pp. 2143–2150, 2017.

[9] M. Kaboli *et al.*, "Robust tactile descriptors for discriminating objects from textural properties via artificial robotic skin," *IEEE Trans. on Rob.*, 2018.

[10] F. Liu *et al.*, "Neuro-inspired electronic skin for robots," *Science Robotics*, vol. 7, no. 67, p. eabl7344, 2022.

[11] M. Kaboli *et al.*, "Tactile-based manipulation of deformable objects with dynamic center of mass," in *IEEE-RAS Int. Conf. on Hum. Rob.* IEEE, 2016.

[12] M. Kaboli *et al.*, "Humanoids learn touch modalities identification via multi-modal robotic skin and robust tactile descriptors," *Adv. Rob.*, 2015.

[13] A. Bierbaum *et al.*, "Haptic exploration for 3d shape reconstruction using five-finger hands," in *IEEE-RAS Int. Conf. on Hum. Rob.* IEEE, 2007.

[14] M. Meier *et al.*, "A probabilistic approach to tactile shape reconstruction," *IEEE Trans. on Rob.*, 2011.

[15] S. Dragiev *et al.*, "Gaussian process implicit surfaces for shape estimation and grasping," in *IEEE Int. Conf. on Rob. and Aut.* IEEE, 2011.

[16] Z. Yi *et al.*, "Active tactile object exploration with gaussian processes," in *IEEE/RSJ Int. Conf. on Int. Rob. and Sys.* IEEE, 2016.

[17] M. Björkman *et al.*, "Enhancing visual perception of shape through tactile glances," in *IEEE/RSJ Int. Conf. on Int. Rob. and Sys.* IEEE, 2013.

[18] G. Z. Gandler *et al.*, "Object shape estimation and modeling, based on sparse gaussian process implicit surfaces, combining visual data and tactile exploration," *Rob. and Auto. Sys.*, 2020.

[19] W. Martens *et al.*, "Geometric priors for gaussian process implicit surfaces," *IEEE Rob. and Aut. Let.*, 2016.

[20] S. Suresh *et al.*, "Tactile slam: Real-time inference of shape and pose from planar pushing," in *IEEE Int. Conf. on Rob. and Auto.* IEEE, 2021.

[21] N. Jamali *et al.*, "Active perception: Building objects' models using tactile exploration," in *IEEE-RAS Int. Conf. on Hum. Rob.* IEEE, 2016.

[22] E. Smith *et al.*, "3d shape reconstruction from vision and touch," *Adv. in Neu. Inf. Proc. Sys.*, 2020.

[23] B. Fei *et al.*, "Comprehensive review of deep learning-based 3d point cloud completion processing and analysis," *IEEE Trans. on Int. Trans. Sys.*, 2022.

[24] P. K. Murali *et al.*, "Deep active cross-modal visuo-tactile transfer learning for robotic object recognition," *IEEE Rob. and Auto. Let.*, 2022.

[25] C. R. Qi *et al.*, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proc. of the IEEE conf. on comp. vis. and pat. rec.*, 2017.

[26] L. Yu *et al.*, "Pu-net: Point cloud upsampling network," in *IEEE conf. on comp. vis. and pat. recog.*, 2018.

[27] W. Yuan *et al.*, "Pcn: Point completion network," in *Int. conf. on 3D vis.* IEEE, 2018.

[28] N. A. Piga *et al.*, "Maskukf: An instance segmentation aided unscented kalman filter for 6d object pose and velocity tracking," *Fr. in Rob. and AI*, 2021.

[29] P. K. Murali *et al.*, "An empirical evaluation of info. gain criteria for active tactile action selection for pose estimation," in *2022 IEEE Int. Conf. on Flex. and Print. Sens. and Sys.* IEEE, 2022, pp. 1–4.

[30] P. K. Murali *et al.*, "Towards robust 3d object recognition with dense-to-sparse deep domain adaptation," in *2022 IEEE Int. Conf. on Flex. and Print. Sens. and Sys.* IEEE, 2022, pp. 1–4.

[31] P. K. Murali *et al.*, "Active visuo-tactile interactive robotic perception for accurate object pose estimation in dense clutter," *IEEE Rob. and Aut. Let.*, 2022.

[32] A. X. Chang *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[33] H. Zhang *et al.*, "Self-attention generative adversarial networks," in *Int. conf. on mach. learn.* PMLR, 2019.

[34] G. Borgefors, "Distance transformations in digital images," *Comp. vis., grap., and im. proc.*, 1986.

[35] F. Bernardini *et al.*, "The ball-pivoting algorithm for surface reconstruction," *IEEE trans. on vis. and comp. graph.*, 1999.

[36] A. Hornung *et al.*, "Octomap: An efficient probabilistic 3d mapping framework based on octrees," *Autonomous robots*, 2013.

[37] D. Choukroun *et al.*, "Novel quaternion kalman filter," *IEEE Trans. on Aero. and Elec. Sys.*, 2006.

[38] M. Berger *et al.*, "A benchmark for surface reconstruction," *ACM Trans. on Gra.*, 2013.

[39] C. E. Rasmussen *et al.*, "Gaussian processes for machine learning (gpml) toolbox," *The Jou. of Mach. Learn. Res.*, 2010.

[40] P. J. Besl *et al.*, "Method for registration of 3-d shapes," in *Sensor fusion IV*, 1992.

[41] S. Bouaziz *et al.*, "Sparse iterative closest point," in *Computer graphics forum*. Wiley Online Library, 2013.

[42] S. Hinterstoisser *et al.*, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *ACCV 2012*. Springer, 2013.

[43] Z. Wang *et al.*, "Nerf–: Neural radiance fields without known camera parameters," *arXiv preprint arXiv:2102.07064*, 2021.

[44] M. Kaboli *et al.*, "In-hand object recognition via texture properties with robotic hands, artificial skin, and novel tactile descriptors," in *2015 IEEE Int. Conf. on Hum. Rob.* IEEE, 2015, pp. 1155–1160.

[45] K. Yao *et al.*, "Tactile-based object center of mass exploration and discrimination," in *2017 IEEE Int. Conf. on Hum. Rob.* IEEE, 2017, pp. 876–881.