Deforming Ellipsoids for Shape Estimation

Yorai Shaoul, Katherine Liu, and Nicholas Roy.

***Abstract*— The problem of object shape and pose estimation is integral to various robotic tasks ranging from manipulation to object-level localization and mapping. Given that only partial observations are often available due to self-occlusions or occlusions in the scene, past works have attempted to reconstruct object shapes from incomplete data. The majority of current shape estimation methods rely heavily on vast datasets to learn discrete object shapes from images. While effective for previously known and well defined objects (e.g. cars and bottles), these methods could fall short when fitting shapes of arbitrary objects (e.g. rocks or deformed packages).**

In this work we present a novel approach to estimating a continuous and differentiable shape estimate to partially observable objects. Leveraging past work on computing ellipsoid estimates to objects, our method improves on such coarse estimates by deforming the prior ellipsoid to tightly fit partial observations while retaining a reasonable volume. For a simulated dataset with available ground truth shape and pose information, we show that without relying on prior shape or semantic knowledge our method captures finer geometric details relative to other continuous shape estimates and produces a continuous and differentiable estimate that aids in the computation of grasp poses for the object.

I. Introduction

As roboticists work towards incorporating robots into unstructured environments, from our kitchens to unexplored planets, we wish to employ them with the skills needed to safely and accurately interact with objects in their space via manipulation. To this end, it is necessary for robots to have an understanding of the shape and pose of these objects such that they can reason about grasping them effectively. While the grasping task can be solved when perfect information about the geometry of an object is available [1], such knowledge is difficult to come by in the real world. For instance, one sensor cannot capture hidden parts of objects due to self-occlusions, and in dynamic and unstructured environments multiple objects can interact and occlude one another. Therefore, a manipulation system must be able to reason about objects using incomplete data alone.

To tackle the object-shape estimation challenge from imperfect information, prior works have attempted to learn the full shape of an object from partial data. Some employed learning methods to reconstruct the shape of a partially-observable object from visual observations [2] and others learned to adapt previously known prior shapes to the current observations [3]–[5] (e.g. using 3D models of cars to estimate novel automobile shapes). These methods, however, rely on the availability of massive amounts of semantically-relevant training data. This data is difficult to gather when the objects of interest are novel (e.g. rocks, items in unknown packagings, etc.). Such objects are often semantically ambiguous and do not fit a clear semantic class so it is not possible to adapt preexisting intra-class shape examples to observations.

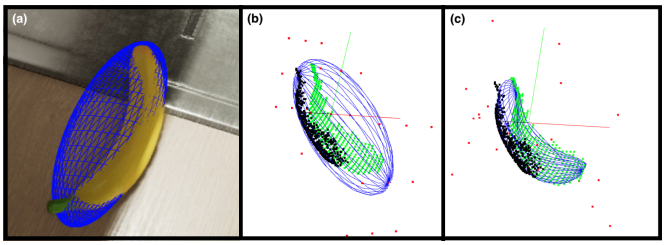

Fig. 1: 3D illustration of DELSE. (a) A primitive ellipsoid $\bar{S}$ (in blue) is fitted to a partial observation of an object O. (b) The partial observation $\mathcal{Z}$ (black points) is complemented with the support point cloud $\mathcal{Z}_s$ (green points). The free-form-deformation control points $\mathcal{P}$ are arranged in a regular grid and shown in red. (c) Starting with the shape of $\bar{S}$, the estimated shape $\hat{S}_{\mathcal{P}}$ is deformed to fit the ground truth shape S^* of the banana by adjusting the positions of the control points $\mathcal{P}$.

Among previous works that attempted to be independent from the semantics of the grasped objects, some sought to estimate a primitive shape (such as a superquadric or an ellipsoid) to fit the partial observations [6], [7]. These works have exploited the favorable mathematical properties of continuous and differentiable shapes to facilitate shape estimation and grasping. For instance, it is possible to recover a quadric representation (the generalized family of spheres and ellipsoids) for an object from planes that bound its observations [8], [9], to check if points are colliding with a quadric [10], [11], and to find grasps on an object [1] when it is represented parametrically. However, in the context of robotic manipulation, such coarse primitive representations of objects might not capture sufficient geometric features to allow effective grasping. See Fig. 1(a, b) for an example. Methods like [12] attempt to directly learn to rank grasps from partial point-cloud observations. While they do not suffer from the coarseness of the quadric representation, they also do not exploit the benefits of a continuous and differentiable representation that contain approximate curvature information for *every* point on the surface of the shape.

In this work, we provide a method for improving coarse shape estimates such that they capture finer geometric details in the estimated objects. Building on past work where

All authors are with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology in Cambridge, USA. {yorai, katliu, nickroy}@mit.edu

This research was sponsored by the MIT Quest for Intelligence and the Army Research Laboratory. It was accomplished under Cooperative Agreement Number W911NF-17-2-0181. Their support is gratefully acknowledged.

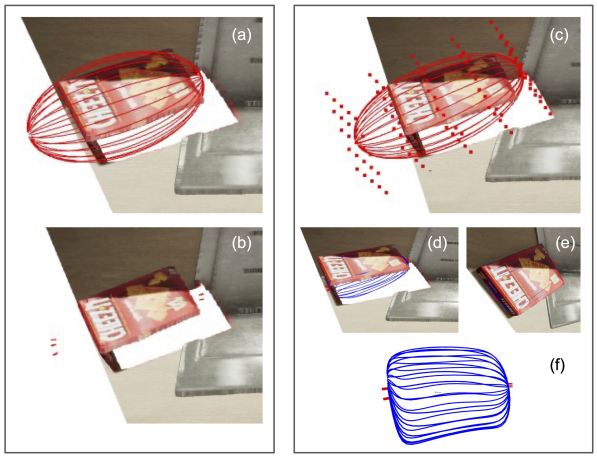

Fig. 2: An illustration of the performance improvements facilitated by deforming an ellipsoid prior $\bar{S}$ to fit a partially observed box object with true-shape S^* (the shape of the cracker box in this example). Image (a) shows a primitive ellipsoid fit to the box observation and image (b) shown antipodal vectors (short red lines) computed on the estimated surface. These vectors do not facilitate good grasps as they are far from the surface. DELSE can improve on this performance. Image (c) shows free-form-deformation control points overlaid on the ellipsoid prior and images (d, e, f) show the deformed ellipsoid with newly computed antipodal normal vectors. The estimated surface $\hat{S}_{\mathcal{P}}$ is now closer to the true box shape S^* and allows for the computation of better surface normals.

quadric representations for object shape and pose estimation were employed [8]–[10], [13], our method treats such primitive shapes as priors and adapts them to better approximate the true shape of a partially-observed object. To achieve this improvement, we compute a free-form-deformation [14], which is a 3D-sculpturing technique widely used in computer graphics, to parameterize a change for a coarse initial shape estimate that deforms it to fit the object better.

We demonstrate the advantages of our method in simulated tests. Evaluating DELSE on a variety of objects, our results show that DELSE successfully recovers good approximate object-shapes from partial point-cloud observations and qualitatively verify that the inferred shapes are sufficiently accurate for grasping. Our method's shape estimates are, on average, 66.5% better than the estimate given by an ellipsoid in terms of a Chamfer distance between the estimates and the true object shapes.

Moreover, our evaluation highlights the benefits of continuous and differentiable representations for shape estimate. In this formulation, we are able to solve for a family of antipodal grasps directly from the shape estimate. This benefit allows for a fast and purposeful search for antipodal grasps over the object shape estimate and, as we show, effective recovery of grasp points.

Our method "Deforming Ellipsoids for Shape Estimation" (DELSE) is an object shape-estimation method that recovers a continuous and differentiable shape representation for a partially-observable object while preserving the level of detail available in discretized observations. Our main contributions are:

- A novel method for optimizing free-form-deformations for ellipsoids.
- Our System DELSE for recovering continuous shape-estimates from partial point-cloud observations.
- An optimization framework for computing favorable grasps to perform object manipulation.
- An ablation study of the impact of noise in the initial ellipsoid estimates on the quality of the DELSE estimates.

In the following sections, we formulate the shape estimation and grasping problems (Section II), and discuss our shape approximation method (Section III) as well as the procedures for recovering primitive shape estimates and antipodal grasp points (Sections III-A and III-C). We describe our algorithm for computing shape estimates from partial point-cloud observations by adapting a coarse primitive estimate. Finally, we report experimental results (Section IV) and show that our method achieves improved performance when compared to other methods that employ continuous representations.

II. Problem Formulation

We are interested in the problems of shape estimation and robotic grasping of irregularly-shaped objects (e.g., rocks, bananas, defective parts, etc.). Given that it is possible to recover a rough initial shape estimate for an object of interest—in the form of a sphere, an ellipsoid or a superquadric for example — our aim in this work is to deform such a primitive initial shape-estimate such that it better represents the real shapes of an object. Fig. 1 illustrates this process. We further exploit the continuous and differentiable nature of our refined estimates to find favorable grasps for objects. We formally introduce the shape estimation problem in Section II-A and the antipodal grasping heuristic in Section II-B.

A. Object Shape Estimation

Given an observation $\mathcal{Z}$ of an object O with shape $S^* : [0, 2\pi)^2 \to \mathbb{R}^3$, our objective in the shape estimation problem is to find an estimate shape $\hat{S}_{\mathcal{P}} : [0, 2\pi)^2 \to \mathbb{R}^3$ that is "close" to the ground truth shape S^*. This estimated shape $\hat{S}_{\mathcal{P}}$ maps a pair of Euler angles $\theta, \phi \in [0, 2\pi)$ to a point in $\mathbb{R}^3$, and is parameterized by some collection of parameters $\mathcal{P}$. We measure "closeness" with a metric $\mathcal{D} : \mathbb{S}^2 \to \mathbb{R}$ with $\mathbb{S}$ being the set of all continuous, differentiable, and closed shapes in 3D. Formally, we wish to find

$$\min_{\mathcal{P}} \mathcal{D}(S^*, \hat{S}_{\mathcal{P}}). \tag{1}$$

In our formulation, the partial observations are point clouds $\mathcal{Z}$ with N_{pcl} points, such that $\mathcal{Z} = \{\mathbf{z_i} : \mathbf{z_i} \in \mathbb{R}^3 \text{ lies on } O,\ i \in \{1, 2, \dots N_{pcl}\}\}$.

B. Antipodal Grasps

We obtain points on the surface of $\hat{S}_{\mathcal{P}}$ that are thought to be effective for grasping O by leveraging the continuous and differentiable surface of the estimate shape $\hat{S}_{\mathcal{P}}$. The antipodal heuristic for computing such points is finding a pair of points $A = \{\mathbf{a_0}, \mathbf{a_1}\}$ with $\mathbf{a_i} \in \mathbb{R}^3$ and on $\hat{S}_{\mathcal{P}}$ for $i \in \{0, 1\}$

such that the normals to $\hat{S}_{\mathcal{P}}$ at points $\mathbf{a_i}$ are approximately collinear. Antipodal points are especially favorable for simple and widely used two-finger grippers. With grasp points being collinear and opposite, contact forces exerted from grippers on an object work against each other and do not induce rotation to the model–a detrimental behavior that can arise when forces are not exerted along one line. In Fig. 2, antipodal points are denoted in the rightmost image with red points.

III. DELSE

We seek to leverage Deformations of Ellipsoids to performs Shape Estimation (DELSE). After recovering an ellipsoid prior, DELSE computes a deformation to the ellipsoid to perform shape estimation for a partially observable object. In order to fit a continuous and differentiable shape $\hat{S}_{\mathcal{P}}$ to an object O that is partially observed through observations $\mathcal{Z}$ in 3-dimensions, we proceed in three steps. Section III-A briefly discusses fitting a primitive ellipsoid $\bar{S}$ to the partial point cloud observation $\mathcal{Z}$. Section III-B details our contribution, which is the optimization of the deformable shape $\hat{S}_{\mathcal{P}}$ originally taking on the shape of the prior $\bar{S}$, to better estimate the shape S^* of the object O. Finally, Section III-C presents our chosen method for extracting effective antipodal grasp points from the refined estimate $\hat{S}_{\mathcal{P}}$ to facilitate robotic manipulation of O.

A. Primitive Shape Prior

Our method DELSE leverages a primitive shape $\bar{S}$ for the optimization of $\hat{S}_{\mathcal{P}}$ to fit O. Building on work where ellipsoids and super-ellipsoids have been used as continuous shape estimates for an object [8], [10], [13], we chose ellipsoids as our primitive shape $\bar{S}$ as illustrated in Fig. 1(a,b). An ellipsoid can be compactly defined [9], [15], [16] with the implicit equation

$$(\mathbf{x}-\mathbf{c})^{\mathbf{T}}\mathbf{Q}(\mathbf{x}-\mathbf{c})-1=0, \tag{2}$$

where $\mathbf{Q} \in \mathbb{S}^3_{++}$ is a 3×3 positive definite matrix and $\mathbf{x}, \mathbf{c} \in \mathbb{R}^3$. The matrix $\mathbf{Q}$ fully determines the orientation and scale of the ellipsoid and $\mathbf{c}$ specifies its centroid. In this construction, for any point $\mathbf{x}$ in 3D space the expression $(\mathbf{x}-\mathbf{c})^{\mathbf{T}}\mathbf{Q}(\mathbf{x}-\mathbf{c})-1$ acts as an *inside-outside* function: taking on the value 1 if $\mathbf{x}$ is on the surface of $\bar{S}$. Otherwise, it will output an algebraic distance to the surface with a negative sign indicating $\mathbf{x}$ being *inside* the ellipsoid and positive otherwise.

The implicit representation makes it possible to find a primitive ellipsoid estimate by inferring the parameters in $\mathbf{Q}$ and $\mathbf{c}$ via solving the minimization

$$\begin{aligned} &\min_{\mathbf{Q},c} \det(\mathbf{Q}^{-1}) \\ &\text{s.t. } (\mathbf{x}_i-\mathbf{c})^T Q(\mathbf{x}_i-\mathbf{c}) \leq 1 \text{ for } i=1,2,\dots n \\ &\quad \mathbf{Q}>0 \end{aligned} \tag{3}$$

as formulated in [16]. The solution of this optimization problem is the minimum volume enclosing ellipsoid to the n observation points $\mathcal{Z}=\{\mathbf{x}_1,\dots\mathbf{x}_n\} \in \mathbb{R}^3$.

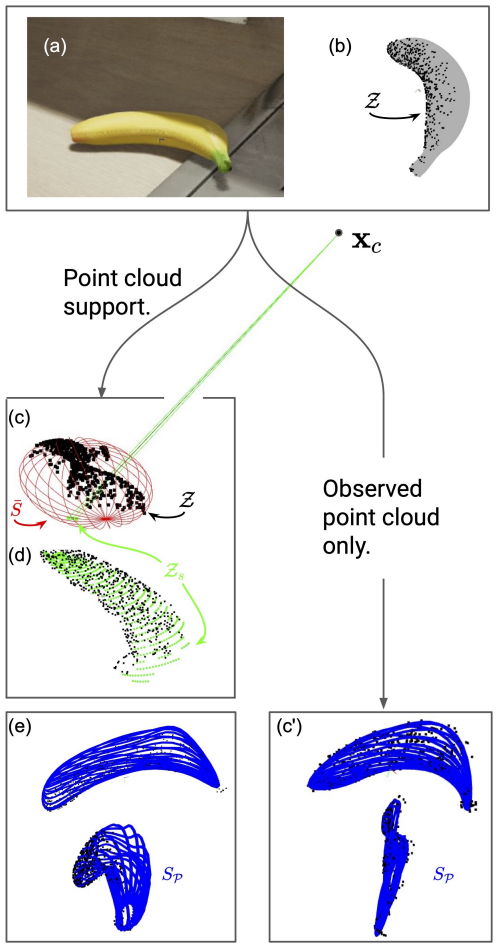

Fig. 3: DELSE constructs a *point cloud support* by using the primitive ellipsoid shape-estimate to provide approximate structure to the self-occluded portions of the observed object O. In this illustration O is a banana (a) and the dark points show the partial point cloud observation of $\mathcal{Z}$ (b). The observation provides incomplete information about the true shape of O, in gray (b). The construction of the point cloud support is illustrated in (c) and (d). The dark line from $\mathbf{x}_c$ connects the camera and an arbitrary observation point $\mathbf{x}_i \in \mathcal{Z}$. The green lines are drawn from the camera position to support points, which are the ellipsoid surface points that are approximately occluded by the observation point $\mathbf{x}_i$. In the absence of point-cloud support (c'), we show that the shape estimate $\hat{S}_{\mathcal{P}}$ is thin. It tightly fits the observation of the object and not the object itself. When point cloud support is use, in (e), the estimated shape assumes a reasonable volume.

B. Shape Optimization

Having recovered an initial shape estimate $\bar{S}$ for the partially observed object O, DELSE further refines the ellipsoid representation to better fit O. To find a well-fitting shape $\hat{S}_{\mathcal{P}}$ to the object O, DELSE infers the parameters $\mathcal{P}$ governing a free-form-deformation (FFD) of $\bar{S}$ such that $\hat{S}_{\mathcal{P}} = FFD(\bar{S}, \mathcal{P})$ approximates S^*, the true shape of O.

*1) **Free Form Deformation**:* A Free-Form-Deformation (FFD) [14] is a space-morphing function widely used in computer graphics for shape deformations. Parameterized by

a grid of $(l+1)\times(m+1)\times(n+1)$, $l,m,n\in\mathbb{N}^{>1}$ control points $\mathcal{P}$, as illustrated in Fig. 2, a point $\mathbf{x}\in\mathbb{R}^3$ is moved to the cartesian coordinate $FFD(\mathbf{x},\mathcal{P})$ according to

$$FFD(\mathbf{x},\mathcal{P})=\sum_{i=0}^{l}B_{l,i}(s)\sum_{j=0}^{m}B_{m,j}(t)\sum_{k=0}^{n}B_{n,k}(u)\mathcal{P}_{i,j,k}. \tag{4}$$

In this formulation, $B_{n,k}(p)=\binom{n}{k}p^k(1-p)^{n-k}$ is the binomial function, $[s\ t\ u]^T$ is the point $\mathbf{x}$ normalized to the grid frame, and $\mathcal{P}_{i,j,k}$ is the 3D control point at the $(i,j,k)^{\text{th}}$ index, which is potentially displaced from its original location in the initially regular lattice. $\mathcal{P}_{i,j,k}\in\mathbb{R}^3$ exists for combinations of $i\in\{0,1,...l\}, j\in\{0,1,\dots m\}$, and $k\in\{0,1,...n\}$. Since the free-form-deformation is a 3D extension of a Bezier curve [14], the offsets of control points $\mathcal{P}$ from their initial locations parameterizes a smooth deformation of points $\mathbf{x}$ according to a combination of points $\mathcal{P}$. Therefore, a smooth and differentiable input (e.g. an ellipsoid) yields a smooth and differentiable output.

*2) **Point Cloud Support**:* Given that our method operates on partial observations $\mathcal{Z}$ of an object O, self-occlusion would lead $\mathcal{Z}$ to only include points on the visible faces of O. Thus, simply deforming $\hat{S}_{\mathcal{P}}$ from the primitive shape $\bar{S}$ to fit the observed point cloud $\mathcal{Z}$ would yield unsatisfactory results (see Fig. 3(c')). To remediate this problem, we rely on our prior estimate $\bar{S}$ to provide a *point cloud support* $\mathcal{Z}_s$ to the observation $\mathcal{Z}$.

The supporting point cloud $\mathcal{Z}_s$ is a set of points sampled from $\bar{S}$ in the self-occluded regions of O, illustrated in Fig. 1(b) and in Fig. 3. To construct $\mathcal{Z}_s$, it is necessary to understand which areas on $\bar{S}$ are self-occluded and should be sampled. We proceed in the following steps, which are illustrated in Fig. 3

Aiming to find points on the ellipsoid that are "hidden" behind the current observation $\mathcal{Z}$, we construct vectors $\mathbf{v_i}=\mathbf{x}_i-\mathbf{x}_c$ between the camera location $\mathbf{x}_c$ and each point $\mathbf{x}_i\in\mathcal{Z}$. Additionally we construct a vector $\mathbf{u}_j=\mathbf{s}_j-\mathbf{x}_c$ between $\mathbf{x}_c$ and a queried point $\mathbf{s}_j\in\bar{S}$ on the ellipsoid surface. Any ellipsoid surface point $\mathbf{s}_j\in\bar{S}$ is said to be self-occluded by the observation point $\mathbf{x}_i$ if there exists a camera-to-observation vector $\mathbf{v}_i$ that is approximately collinear with the camera-to-ellipsoid vector $\mathbf{u}_j$ and $\mathbf{s}_j$ is farther from $\mathbf{x}_i$ with respect to the camera. Formally, the support point cloud takes on the values in the set

$$\begin{aligned}\mathcal{Z}_s=\{\mathbf{x}_i\in\mathcal{Z}:&\mathbf{v}_i^T\mathbf{u}_j>1-\epsilon \text{ and } \|\mathbf{v}_i\|<\|\mathbf{u}_j\|\\ &\forall\mathbf{u}_j \text{ with } s_j\in\bar{S}\}\end{aligned} \tag{5}$$

for some small and positive ϵ. When optimizing the initial ellipsoid to fit O, the reference point cloud is the union of the observation and the support

$$\mathcal{Z}\leftarrow\mathcal{Z}\cup\mathcal{Z}_s. \tag{6}$$

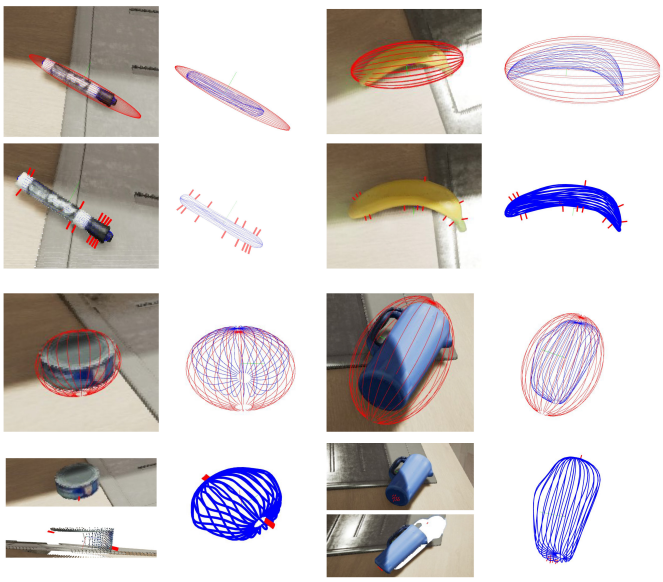

Fig. 4: Illustration of computed antipodal surface normals to a deformed shapes. Clockwise per object from the top left, the images show a partially observed object with a fitted ellipsoid prior (in red), the deformed ellipsoid shape estimate to the object computed with DELSE (in blue), a collection of feasible grasp points to the object, and the antipodal surface normals overlaid over the scene. Following our method outlined in Section III-C with the reference vector $\mathbf{k}$ set to the desktop normal vector, we visually verify that the computed antipodal normal-vector pairs are feasible. The Pitcher example on the bottom right shows a potential failure case, where normals are placed away from the self occluded surface since information regarding the depth of the object is not available. On the bottom left, the Tuna-Can example illustrates a success case where normals are computed in regions that are self-occluded.

*3) **Optimization**:* We formulate the process of fitting $\hat{S}_{\mathcal{P}}$ to a reference point cloud $\mathcal{Z}$ as an optimization. Given the continuous nature of $\hat{S}_{\mathcal{P}}$ and the discrete $\mathcal{Z}$, we choose to operate in the discrete domain. To this end, we sample points on $\hat{S}$ to obtain the point cloud $\hat{S}^d_{\mathcal{P}}$. We note that each point $\mathbf{s}_i\in\hat{S}^d_{\mathcal{P}}$ is governed by control points $\mathcal{P}$. For example, a choice of Euler angles $\theta,\phi\in[0,2\pi)$ samples a point $\mathbf{s}_i=FFD(\bar{S}(\theta,\phi),\mathcal{P})$ on $\hat{S}_{\mathcal{P}}$. Since $FFD(\cdot,\cdot)$ is continuous and differentiable, it is possible to perform gradient-based optimization for the positions of $\mathcal{P}$ by adjusting the sampled points $\mathbf{s}_i$.

Aiming to adjust the sampled points $\mathbf{s}_i$ on $\hat{S}$ such that they are close to the reference point cloud $\mathcal{Z}$, we use the Chamfer Distance $\mathcal{D}_{\text{Chamfer}}$ metric to measure the quality of our fit [17].

$$\begin{aligned}&\mathcal{D}_{\text{Chamfer}}(\mathcal{Z},\hat{S}^d_{\mathcal{P}})=\\ &\sum_{\mathbf{z}_i\in\mathcal{Z}}\|\mathbf{z}_i-m_{\hat{S}^d_{\mathcal{P}}}(\mathbf{z}_i)\|^2+\sum_{\mathbf{s}_i\in\hat{S}^d_{\mathcal{P}}}\|\mathbf{s}_i-m_{\mathcal{Z}}(\mathbf{s}_i)\|^2\end{aligned} \tag{7}$$

with $m_X(\mathbf{y})=\arg\min_{\mathbf{x}_i\in X}\|\mathbf{y}-\mathbf{x}_i\|^2$ denoting the nearest point to $\mathbf{y}$ in among points in set X.

Adding a regularization term $\lambda(\mathcal{P})$, composed of the L_2 distance between control points $\mathcal{P}$ and their original positions

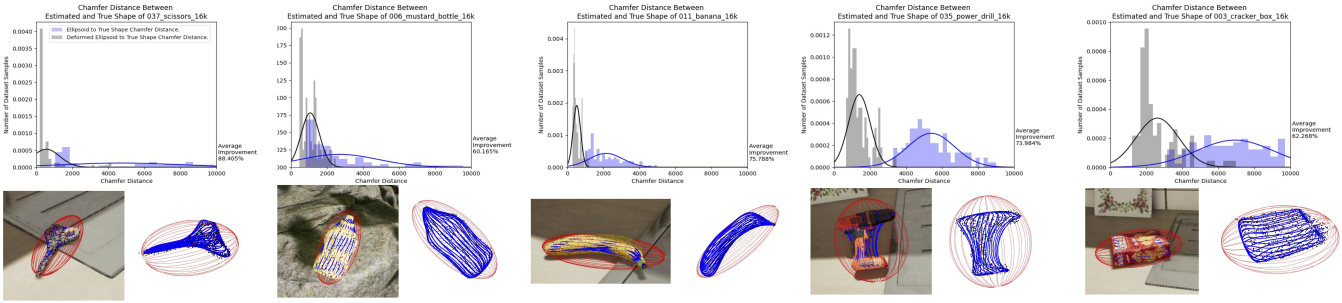

Fig. 5: Shape estimation evaluation with objects from the Falling Things dataset scenes. For each object (five of eight instances illustrated) in the dataset, we quantify estimation quality by computing a Chamfer Distance between points sampled from objects' true shape and points sampled from their respective estimated shape via DELSE (in red) and via ellipsoid estimate (in blue). Lower Chamfer Distance is better. The plots above include histograms (and Gaussians of best fit) of the Chamfer Distances obtained over 100 samples of each object. The blue fitted histograms correspond to ellipsoid estimates and the black ones correspond to ours. Our experiments show average Chamfer Distance reduction of at least 44%, pointing to the efficacy of DELSE in estimating shapes for partially observable objects.

(to keep control point deviation small) and a penalty on neighboring control points overlapping on any axis (to battle self-intersections), our optimization is

$$\min_{\mathcal{P}} \mathcal{D}_{\text{Chamfer}}(\mathcal{Z}, \hat{S}^d_{\mathcal{P}}) + \lambda(\mathcal{P}). \tag{8}$$

We recover the estimated shape $\hat{S}_{\mathcal{P}}$ from the optimized control points by applying a free-form deformation to the primitive ellipsoid. With slight abuse of the notation introduced in Section III-B, $\hat{S}_{\mathcal{P}} = FFD(\bar{S}, \mathcal{P})$.

C. Antipodal Points

We exploit the continuous and differentiable nature of our shape estimate $\hat{S}_{\mathcal{P}}$ to compute antipodal grasp points $A = \{\mathbf{a}_1, \mathbf{a}_2\} = \{\hat{S}_{\mathcal{P}}(\theta_1, \phi_1), \hat{S}_{\mathcal{P}}(\theta_2, \phi_2)\}$. One advantage of a continuous shape representation is the ability to directly compute the curvature and surface normal for any arbitrary point on the shape. Specifically, a normal vector at the point $\hat{S}_{\mathcal{P}}(\theta_0, \phi_0)$ takes on the form

$$\mathbf{n}(\theta, \phi) = \eta \left(\frac{\partial}{\partial \theta} \hat{S}_{\mathcal{P}}(\theta_0, \phi_0) \right) \times \left(\frac{\partial}{\partial \phi} \hat{S}_{\mathcal{P}}(\theta_0, \phi_0) \right) \tag{9}$$

with η being a normalizing factor to make $\mathbf{n} \in \mathbb{R}^3$ of unit length. In general, it is possible to analytically solve for pairs of global antipodal points (i.e. without restricting the orientation of the points) with methods such as [1]. However, since we are mainly concerned with shape estimation for robotic applications where physical constraints may limit the range of possible control points (e.g. a robotic arm cannot move through a table), we choose to leave analytical grasp computation for future work.

To compute grasping poses given a continuous shape we begin by samplng its surface in points parameterized by $(\theta_1, \phi_1), \ldots (\theta_m, \phi_m)$ and compute a collection of normal vectors $N = \{\mathbf{n}(\theta_1, \phi_1), \ldots \mathbf{n}(\theta_m, \phi_m)\}$. Given the normal vector collection N, the antipodal grasp problem reduces to the task of finding a pair of normals $\mathbf{n}(\theta_1, \phi_1), \mathbf{n}(\theta_2, \phi_2) \in N$, with associated points $\mathbf{a}_1 = \hat{S}_{\mathcal{P}}(\theta_1, \phi_1), \mathbf{a}_2 = \hat{S}_{\mathcal{P}}(\theta_2, \phi_2)$, that are approximately collinear, that point in opposite directions, and that produce a physically feasible grasp.

We find a pair of collinear normals $\mathbf{n}_1, \mathbf{n}_2$ that correspond to the surface points $\mathbf{a}_1, \mathbf{a}_2$ by solving

$$\underset{\mathbf{n}_1, \mathbf{n}_2 \in N}{\arg\min}\ \mathbf{n}_1(\mathbf{a}_1 - \mathbf{a}_2)^T + \mathbf{n}_2(\mathbf{a}_2 - \mathbf{a}_1). \tag{10}$$

This objective ensures that the normals are pointing in opposite directions and are collinear with the line connecting their start points.

To satisfy the physical feasibility requirement, we choose a vector $\mathbf{k}$ that is thought to be the direction along which the robotic end-effector can move. For example, in a setting where items are on a desktop, $\mathbf{k}$ could be the normal vector to the desktop. Thus, these physical feasibility constraint vector (or a collection of vectors) could be found automatically or encoded ahead of time using task-specific knowledge. To satisfy this condition, normal pairs should be perpendicular to $\mathbf{k}$. Our final optimization for finding antipodal points is therefore

$$\underset{\mathbf{n}_1, \mathbf{n}_2 \in N}{\arg\min}\ \mathbf{n}_1(\mathbf{a}_1 - \mathbf{a}_2)^T + \mathbf{n}_2(\mathbf{a}_2 - \mathbf{a}_1) + |\mathbf{n}_1\mathbf{k}^T| + |\mathbf{n}_2\mathbf{k}^T|. \tag{11}$$

with the last two components taking a value of zero when $\mathbf{k}$ and the surface normals $\mathbf{n}$ are perpendicular, i.e. the grasp pose can be executed by a robotic end-effector that moves along $\mathbf{k}$.

IV. Experimental Results

In this section we detail our simulated experiments. We show the efficacy of DELSE via

- our quantitative results on an object shape and pose estimation dataset,
- via qualitatively evaluating the performance of our grasp-pose computation pipeline,
- and with an analysis the the behavior of DELSE under varying configurations.

Fig. 5 visualizes our shape estimation evaluation results for five example objects.

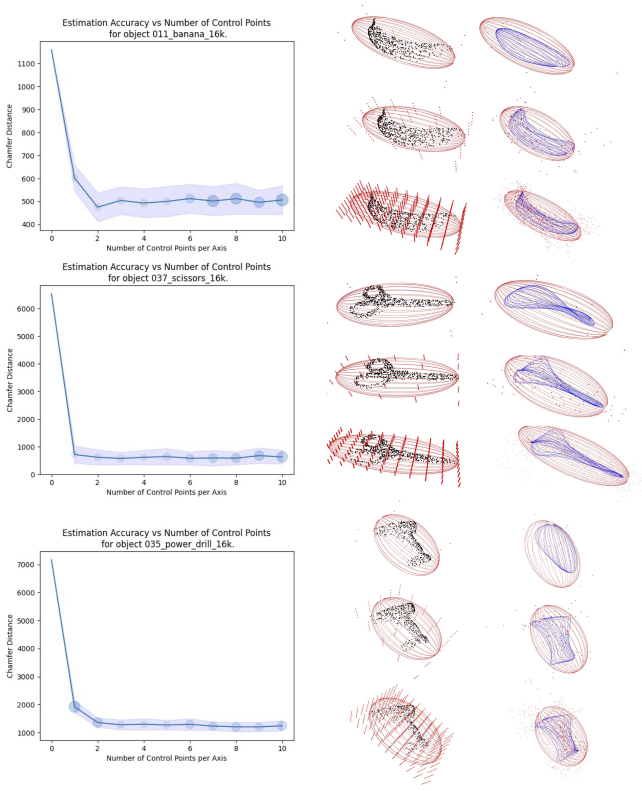

Fig. 6: Analysis of the impact of the number of control points on the quality of shape estimates. We varied the number of control points $l = m = n$ used to generate a shape estimate for three objects (banana, scissors, and drill) across 30 frames each. In the plots we compare the shape estimate to the ground truth shape via a Chamfer Distance and denote the computation time with the size and transparency of the circular dots. The shaded regions are the 99% confidence intervals. We note that after the performance improvement between 0 and 2 control points per axis (with 0 control points being the ellipsoid estimate), the later improvement is relatively small. We provide visualizations of the deformation given by 1, 4, and 10 control points on each axis for the three objects as well.

A. Dataset

We evaluate DELSE via the Falling Things (FAT) dataset [18]. This dataset provides depth images for objects in various settings alongside their true pose and shape. All images include segmentations for the shown objects as well.

B. Shape Estimation Evaluation

To evaluate the efficacy of our proposed method we compare our resulting shape estimates to an ellipsoid estimate baseline, which is a popular choice for continuous object shape representation [10], [13]. Our baseline ellipsoid shape estimate is computed with the method detailed in Section III-A.

In our shape-estimation experiments, we have used the Chamfer Distance (Eq. 7) metric to quantify the quality of a shape estimate given the true shape of the estimated object. For an object O, we sample points $\{\mathbf{s}_i^*\}$ from the surface of its true shape (transformed to its true pose in the evaluated scene), we sample points $\{\bar{\mathbf{x}}_i\}$ from the ellipsoid estimate $\bar{S}$, and sample points $\{\hat{\mathbf{s}}_i\}$ from the DELSE estimate $\hat{S}_{\mathcal{P}}$. We assign $\bar{S}$ the estimation quality

$$Q_{\text{Baseline}} = \mathcal{D}_{\text{Chamfer}}\left(\{\mathbf{s}_i^*\}, \{\bar{\mathbf{s}}_i\}\right)$$

and assign the DELSE output $\hat{S}_{\mathcal{P}}$ the quality score

$$Q_{\text{DELSE}} = \mathcal{D}_{\text{Chamfer}}\left(\{\mathbf{s}_i^*\}, \{\hat{\mathbf{s}}_i\}\right).$$

Finally, we compute the *improvement* in the estimate between the ellipsoid baseline and our method with the ratio $(Q_{\text{Baseline}} - Q_{\text{DELSE}})/Q_{\text{Baseline}}$ and interpret a higher *improvement* score as better.

Our results show that DELSE provides a substantial improvement to the baseline shape estimate provided by the ellipsoid baseline. Across 800 frames from the FAT dataset, where each of 8 objects appears alone in 100 frames, DELSE provided an average improvement of 66.5%. As shown in Fig. 5, both irregularly shaped objects and relatively smooth-shaped objects benefit from our method. For example, the DELSE estimate for scissors is 88.4% better than the ellipsoid baseline estimate and in the cracker box example DELSE provided an average improvement of 62.2%. Table I details our quantitative shape estimation on the FAT dataset.

	Chamfer Distance ↓ for method		
	DELSE	Ellipsoid	Improvement ↑
Marker	**261.83**	1398.0	81.27%
Pitcher	**2969.7**	6286.1	52.7%
Tune Can	**330.10**	662.64	50.1%
Bowl	**989.14**	1789.1	44.7%
Scissors	**545.01**	4700.4	88.4%
Drill	**1404.8**	5400.2	73.9%
Mustard	**1066.8**	2678.3	60.1%
Banana	**519.12**	2144.1	75.7%
Box	**2609.8**	6916.8	62.2%

TABLE I: Quantitative shape estimation results on the Falling Things (FAT) dataset comparing DELSE to an ellipsoid prior. We report the average Chamfer Distance between object ground truth shapes and estimated shapes via DELSE and an ellipsoid prior for 8 objects across 800 frames. The symbol ↓ denotes that a lower value is better, and ↑ that a larger score is preferable. For all evaluated objects, DELSE successfully improves upon the ellipsoid estimate.

C. Grasp Pose Evaluation

We qualitatively evaluate the performance of our method's ability to recover grasps from a given continuous and differentiable shape estimate, as illustrated in Fig. 4. On the FAT dataset, we show that the continuous and differentiable nature of our estimates can facilitate the computation of analytical surface derivatives and for the recovery of visually effective grasps. Robot experiments are left to future work.

D. Ablation Analysis

*1) **Number of Control Points**:* Free-form-deformations can, in theory, capture finer details when more control points are in use. An increase in the number of control point, however, makes the optimization of the Free-Form-Deformation more difficult. The analysis in Fig. 6 illustrates the relationship between a choice of the number of control points to the estimation quality and computation

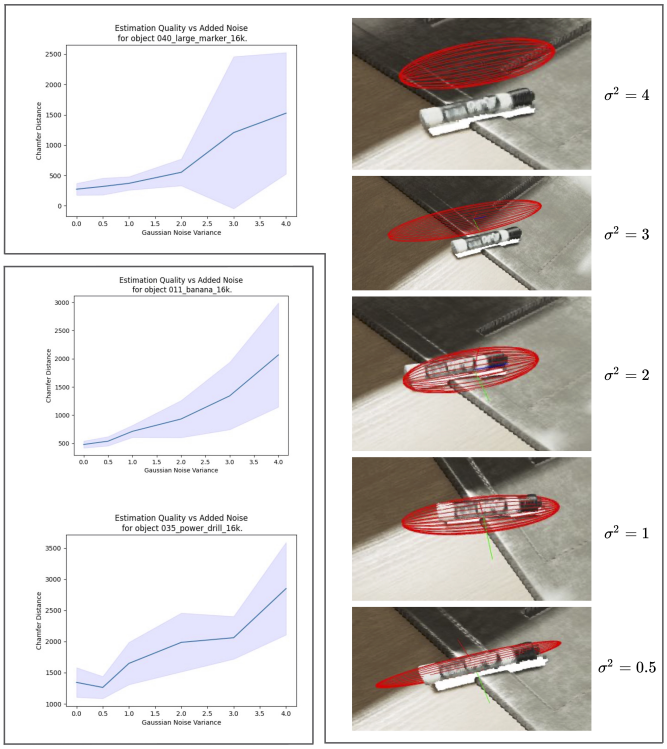

Fig. 7: Analysis of the impact of noise in the prior ellipsoid estimate on the quality of the deformed ellipsoid shape estimates. Applying zero-mean Gaussian noise on the shape and translation of the ellipsoids, we varied its variance $\sigma^2 \in \{0.5, 1, 2, 3, 4\}$. For the objects Marker, Banana, and Drill, the range $\sigma^2 \in [0, 2]$ seemed to induce little degradation while performance for $\sigma^2 \in [2, 4]$ was worse. The images on the right illustrate primitive ellipsoids generated with different noise levels. Data for this analysis was generated from 30 images for each of the objects and the shaded regions are 99% confidence intervals.

time. Perhaps un-intuitively, our results suggest that after the initial quality improvement achieved by performing a free-form-deformation on the primitive ellipsoid with a small number of control points, adding more control points does not offer significant quality improvements but does increase runtime. We have experimentally found that 4 control points per axis on a symmetric regular grid performed produced the best results. On an un-optimized CPU, 300 iterations of the DELSE optimization take between 2 and 5 seconds to execute.

*2) **Sensitivity to Ellipsoid Prior**:* Given that our method relies on an ellipsoid shape estimate prior to computing a suitable deformation, we provide an analysis of our estimation pipeline's response to noise in the ellipsoid prior parameters. As shown in Fig. 7, we applied zero-mean Gaussian noise on the shape and translation of the ellipsoids and varied its variance $\sigma^2 \in \{0.5, 1, 2, 3, 4\}$. We observed that performance generally degrades with increase in the values of σ^2 and hypothesize that this trend is induced by the change in the quality of the point cloud support $\mathcal{Z}_s$. An ellipsoid prior that does not cover the observation will not provide support to self-occluded regions of the object.

V. Conclusion

In this work we have presented a novel shape estimation method for recovering continuous and differentiable estimates for partially observable objects. Our method DELSE improves upon past work in continuous shape-representation by adding a refinement module that can be applied to any prior shape estimate. In our experiments, DELSE leveraged an ellipsoid shape prior for partially observed objects and further refined it to better fit the observed objects–yielding an average 66.5% improvement. Experimentally, we showed that the DELSE shape estimates fit observed objects better than ellipsoid estimates and that it is possible to compute physically feasible antipodal grasps from our estimates.

The benefits of continuous and differentiable shape estimates may extend beyond shape estimation and grasp computation. We believe that this representation choice could aid in producing novel solutions to problems such as instance-level object data-association and robot localization and mapping. We leave such inquiry to future work.

References

[1] I.-M. Chen and J. W. Burdick, "Finding antipodal point grasps on irregularly shaped objects," *IEEE transactions on Robotics and Automation*, vol. 9, no. 4, pp. 507–512, 1993.

[2] J. Gu, W.-C. Ma, S. Manivasagam, W. Zeng, Z. Wang, Y. Xiong, H. Su, and R. Urtasun, "Weakly-supervised 3d shape completion in the wild," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Springer, 2020.

[3] B. Wallace and B. Hariharan, "Few-shot generalization for single-image 3d reconstruction via priors," *CoRR*, vol. abs/1909.01205, 2019. arXiv: 1909.01205. [Online]. Available: http://arxiv.org/abs/1909.01205.

[4] D. Jack, J. K. Pontes, S. Sridharan, C. Fookes, S. Shirazi, F. Maire, and A. Eriksson, "Learning free-form deformations for 3d object reconstruction," in *Asian Conference on Computer Vision*, Springer, 2018, pp. 317–333.

[5] A. Kurenkov, J. Ji, A. Garg, V. Mehta, J. Gwak, C. Choy, and S. Savarese, "Deformnet: Free-form deformation network for 3d shape reconstruction from a single image," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2018, pp. 858–866.

[6] A. Makhal, F. Thomas, and A. P. Gracia, "Grasping unknown objects in clutter by superquadric representation," in *2018 Second IEEE International Conference on Robotic Computing (IRC)*, IEEE, 2018, pp. 292–299.

[7] F. Solina and R. Bajcsy, "Recovery of parametric models from range images: The case for superquadrics with global deformations," *IEEE transactions on pattern analysis and machine intelligence*, vol. 12, no. 2, pp. 131–147, 1990.

[8] K. Ok, K. Liu, K. Frey, J. P. How, and N. Roy, "Robust object-based slam for high-speed autonomous navigation," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 669–675.

[9] C. Rubino, M. Crocco, and A. Del Bue, "3d object localisation from multi-view image detections," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1281–1294, 2018. DOI: 10.1109/TPAMI.2017.2701373.

[10] G. Vezzani, U. Pattacini, G. Pasquale, and L. Natale, "Improving superquadric modeling and grasping with prior on object shapes," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2018, pp. 6875–6882.

[11] B. Odehnal, H. Stachel, and G. Glaeser, *The Universe of Quadrics*. Springer Nature, 2020.

[12] J. Mahler, M. Matl, X. Liu, A. Li, D. Gealy, and K. Goldberg, "Dex-net 3.0: Computing robust robot suction grasp targets in point clouds using a new analytic model and deep learning," *arXiv preprint arXiv:1709.06670*, 2017.

[13] K. Liu and N. Roy, "Volumon," in *Under review for the 2021 International Conference on Intelligent Robotic Systems (IROS)*, 2021.

[14] T. W. Sederberg and S. R. Parry, "Free-form deformation of solid geometric models," in *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, 1986, pp. 151–160.

[15] W. N. Greene and N. Roy, "Flame: Fast lightweight mesh estimation using variational smoothing on delaunay graphs," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4696–4704. DOI: 10.1109/ICCV.2017.502.

[16] N. Moshtagh *et al.*, "Minimum volume enclosing ellipsoid," *Convex optimization*, vol. 111, no. January, pp. 1–9, 2005.

[17] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *European Conference on Computer Vision (ECCV)*, 2014.

[18] J. Tremblay, T. To, and S. Birchfield, "Falling things: A synthetic dataset for 3d object detection and pose estimation," *CoRR*, vol. abs/1804.06534, 2018. arXiv: 1804.06534. [Online]. Available: http://arxiv.org/abs/1804.06534.