MULTI-VIEW SHAPE ESTIMATION OF TRANSPARENT CONTAINERS

Alessio Xompero[1] *, Ricardo Sanchez-Matilla*[1]*, Apostolos Modas*[2]*, Pascal Frossard*[2]*, Andrea Cavallaro*[1]

[1]Centre for Intelligent Sensing, Queen Mary University of London, UK
[2] LTS4, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

The 3D localisation of an object and the estimation of its properties, such as shape and dimensions, are challenging under varying degrees of transparency and lighting conditions. In this paper, we propose a method for jointly localising container-like objects and estimating their dimensions using two wide-baseline, calibrated RGB cameras. Under the assumption of circular symmetry along the vertical axis, we estimate the dimensions of an object with a generative 3D sampling model of sparse circumferences, iterative shape fitting and image re-projection to verify the sampling hypotheses in each camera using semantic segmentation masks. We evaluate the proposed method on a novel dataset of objects with different degrees of transparency and captured under different backgrounds and illumination conditions. Our method, which is based on RGB images only, outperforms in terms of localisation success and dimension estimation accuracy a deep-learning based approach that uses depth maps.

***Index Terms*—** Object localisation, Dimension estimation, Transparency.

1. INTRODUCTION

Localising objects in 3D and estimating their properties (*e.g.* dimensions, shape), as well as their pose (location, orientation), is important for several robotic tasks, such as grasping [1, 2], manipulation [3] and human-to-robot handovers [4]. However, everyday objects can widely vary in shape, size, material, and transparency, thus making the vision-based estimation of their properties a challenging problem.

Existing methods for localising objects in 3D or estimating their 6 Degrees of Freedom (DoF) pose rely on databases of 3D object models (*e.g.* CAD) [5, 6, 7, 8] or motion capture systems [4, 9, 10]. To avoid using markers for motion capture, feature points [11, 12] can be localised in an image and matched against a 3D object model to estimate the object pose by solving a Perspective-n-Point (PnP) problem [13]. However, this strategy may fail when objects exhibit limited texture or are captured under unfavourable lighting conditions [7]. Approaches based on Deep Neural Network (DNN) learn to estimate the 6 DoF object pose quite accurately, but their training requires large amount of data, usually annotated only for the high-level object class [14], including depth information and/or *known* dense 3D models in addition to colour images [5, 6, 7, 15, 16]. For example, PoseCNN [17], DenseFusion [5], SegOPE [16] and PVNet [6] are evaluated only with objects whose high-quality 3D models and depth were available [17], discarding testing objects that are transparent as the segmentation may fail or be inaccurate. DenseFusion [5] combines features obtained from RGB-D data to handle occlusions and inaccurate segmentation. Pixel-wise Voting Network (PVNet) [6] estimates the pose of occluded or truncated objects with an uncertainty-driven PnP, learning a vector-field representation to localise a sparse set of 2D keypoints and their spatial uncertainty. Normalized Object Coordinate Space (NOCS) [15] formulates this coordinate space to jointly estimates the 6 DoF pose and the dimensions (in the form of a 3D bounding box) of an object not seen during the training of the DNN (*e.g.* intra-class variability for object shape, size, and appearance). As most of these works target object pose estimation, related comprehensive reviews can be found in [5, 6, 15, 16]. To handle textureless, translucent or reflective objects (*e.g.* wine glasses), whereas 3D reconstruction may perform poorly, Learning the Grasping Point (LGP) [2] uses supervised training on synthetic images with annotated grasping regions and learns to identify in two or more images a few points that are good for grasping unknown objects in 3D. Table 1 summarises relevant works based on their assumptions; their targeted tasks, especially the object dimensions estimation in addition to the localisation in 3D; and their capability to handle transparent objects. Nevertheless, estimating the dimensions of these objects in 3D is still challenging.

Table 1. Comparison of markerless methods for object localisation and dimensions estimation in 3D. KEY – Ref.: reference; n3D: no 3D object model; nD: no depth; HLC: known high-level object class; Loc.: object localisation in 3D; Dim.: object dimensions estimation in 3D; 3DM: dimensions given by the 3D model.

Ref.	Method	Assumptions			Tasks		Transparency
		n3D	nD	HLC	Loc.	Dim.	
[2]	LGP	✓	✓		✓		✓
[18]	DeepIM		✓		✓	3DM	
[7]	StoCS			✓	✓	3DM	
[6]	PVNet		✓	✓	✓	3DM	
[5]	DenseFusion			✓	✓	3DM	
[16]	SegOPE		✓	✓	✓	3DM	
[15]	NOCS	✓		✓	✓	✓	
	LoDE	✓	✓	✓	✓	✓	✓

In this paper, we propose LoDE (Localisation and object Dimensions Estimator)[1], a method that estimates the dimensions of container-like objects, such as cups, drinking glasses and bottles, using two calibrated RGB cameras, whose poses are known. LoDE localises the 3D centroid of the object from 2D centroids estimated from semantic segmentation masks. As most of these containers have a circular symmetry along their vertical axis, LoDE hypothesises an initial model with a set of circumferences sampled around the 3D centroid at different heights. Then, the model iteratively fits to the object by reducing the radius for sampling the circumferences until each circumference is verified within the object mask in each

This work is supported by the CHIST-ERA program through the project CORSMAL, under UK EPSRC grant EP/S031715/1 and Swiss NSF grant 20CH21_180444.

[1] http://corsmal.eecs.qmul.ac.uk/LoDE.html

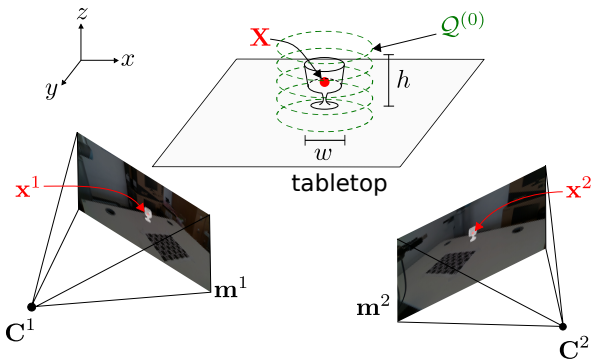

Fig. 1. Two cameras capture an object from different viewpoints. Given only RGB images and camera poses, we estimate the width w and height h of the object without relying on 3D object models, depth information, or markers. The proposed method, LoDE, localises the object centroid in 3D, $\mathbf{X}$, from the 2D centroids, $\mathbf{x}^1$ and $\mathbf{x}^2$, estimated on the segmented images, $\mathbf{m}^1$ and $\mathbf{m}^2$, and then samples a set of sparse 3D points, $\mathcal{Q}^{(0)}$, belonging to circumferences centred at the centroid location and at different heights, to fit the object shape with an iterative 3D-2D algorithm.

camera. We also collected a novel dataset with objects of different shapes and degrees of transparency, under varying lighting conditions and backgrounds.

2. LOCALISATION AND DIMENSION ESTIMATION

We propose a generative 3D sampling model to estimate the shape of an object and, as by-product, its dimensions, assuming the object to be circular symmetric with respect to its vertical axis. We represent the object as $\mathbf{O} = (x, y, z, w, h) \in \mathbb{R}^5$, where $\mathbf{X} = (x, y, z) \in \mathbb{R}^3$ is the location of its centroid in 3D, and h and w are the height and the largest width, respectively. Let I^c represent the camera views, where the object is observed, and $\mathbf{C}^c$ be the 3D pose of each camera whose calibration is modelled by the intrinsic parameters $\boldsymbol{\theta}^c$, consisting of focal length and principal point, with $c \in \{1, 2\}$.

As the object location and shape in 3D are unknown, we propose an iterative multi-view 3D-2D shape fitting via projective geometry [9] (see Fig. 1). The object is first detected in each image I^c via semantic segmentation:

$$D : \{0, \dots, 255\}^{W,H,C} \rightarrow \{0, 1\}^{W,H}, \tag{1}$$

where W, H, C are the image width, height and number of colour channels, respectively, and $\mathbf{m}^c = D(I^c) \in \{0, 1\}^{W,H}$ a binary feature map representing the segmented object.

Finding pixel correspondences between views when objects are textureless is challenging and ambiguities can lead to inaccurate estimations of the object location in 3D and its dimensions. We instead estimate the 2D centroid $\mathbf{x}^c$ of the segmented object with the intensity centroid method [19] through the definition of the moments within a local image area. Then, we triangulate the two 2D centroids to estimate the object centroid in 3D [9]:

$$\tilde{\mathbf{X}} = \tau(\mathbf{x}^1, \mathbf{x}^2, \mathbf{C}^1, \mathbf{C}^2, \boldsymbol{\theta}^1, \boldsymbol{\theta}^2), \tag{2}$$

where τ is the triangulation operator.

To estimate the object shape, we initialise around its estimated 3D centroid a cylindrical model that iteratively fits the object shape

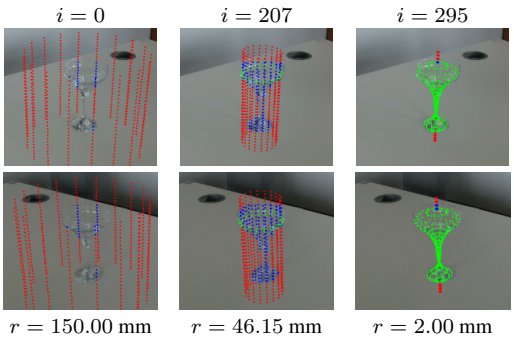

Fig. 2. Initialisation, sampled iteration and convergence of the 3D-2D shape fitting of a drinking glass (top: left camera, bottom: right camera). Legend: i iteration number, r radius of currently sampled circumference, ● projected points lying outside the segmentation mask, ● projected points lying inside the segmentation mask and ● projected points whose circumference fits the shape of the object (inside the segmentation mask of both cameras).

as observed by the cameras. For each iteration i, our approach samples L circumferences of radius $r^{(i)}$, centred at the estimated object 3D location $\tilde{\mathbf{X}}$ and with varying height z_l, $l = 1, \dots, L$,

$$\mathcal{C}^{(i)} = \{(r_l^i, z_l, \nu_l)\}_{l=1:L}, \tag{3}$$

where $\nu_l \in \{0, 1\}$ indicates whether a circumference lies within the object mask of both cameras. For each circumference l, we sample a set of N sparse 3D points,

$$\mathcal{Q}_l^{(i)} = \{\mathbf{Q}_{n,l}^{(i)} = (x_{n,l}, y_{n,l}, z_l)\}_{n=1:N}, \tag{4}$$

and the set of all sampled 3D points is $\mathcal{Q}^{(i)} = \{\mathcal{Q}_l^{(i)}\}_{l=1:L}$. We project the sampled 3D points onto the image of both cameras as

$$\mathbf{u}_{n,l}^c = \pi(\mathbf{Q}_{n,l}^{(i)}, \mathbf{C}^c, \boldsymbol{\theta}^c), \tag{5}$$

where $\pi(\cdot) : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ is the projection function [9]. Then, we verify if all the points belonging to circumference l, $\mathcal{Q}_l^{(i)}$, lie within the object mask of both cameras,

$$\eta = \sum_{n=1}^{N} \mathbf{m}^1(\mathbf{u}_{n,l}^1) + \mathbf{m}^2(\mathbf{u}_{n,l}^2), \tag{6}$$

and if the condition is satisfied (*i.e.* $\eta = 2N$), we set the corresponding flag as converged, *i.e.* $\nu_l = 1$. For iteration $i + 1$, we decrease the radius $r_l^{(i+1)}$ and re-sample the 3D circumference points, $\mathcal{Q}^{(i+1)}$. Points belonging to circumference l and with $\nu_l = 1$ are not re-sampled. This iterative 3D-2D shape fitting terminates when either all $\nu_l = 1$ or $r_l^{i+1} < \rho$, where ρ is the minimum radius that is sampled. Fig. 2 shows as example three iterations of the shape fitting for a transparent drinking glass.

Finally, to estimate the dimensions of the object, we select among the converged circumferences, $\mathcal{V} = \{(r_l, z_l, \nu_l) | \nu_l = 1\} \subset \mathcal{C}$, the one with the largest radius r^* and the ones with maximum and minimum heights, z^* and $\bar{z}$, respectively. The estimated largest object width is $\tilde{w} = 2r^*$ and the object height is $\tilde{h} = z^* - \bar{z}$.

3. THE CORSMAL CONTAINERS DATASET

We collect a set of images using 23 containers for liquids: 5 cups, 9 drinking glasses and 9 bottles (see Fig. 3). These objects are made

Fig. 3. Objects in the CORSMAL Container dataset. Objects 1 to 13 (transparent); 14 to 18 (translucent); 19 to 23 (opaque). Note that crops are taken from images acquired with the same camera view.

of plastic, glass or paper, with different degrees of transparency and arbitrary shapes. The dataset contains 3 objects that do not have circular symmetry, *e.g.* object 6 (diamond-shaped glass), object 16 (amaretto bottle) and object 20 (deformed water-bottle).

We placed each object on a table and we acquired RGB, depth and stereo infrared (IR) images (1280×720 pixels) with two Intel RealSense D435i cameras, located approximately at 40 cm from the object. RGB and depth images are spatially aligned. The cameras are calibrated and localised with respect to a calibration board. We acquired the images in two rooms with different lighting and background conditions. The first setup is an *office* with natural light from a window and objects placed on a table of size 160x80 cm and height 82 cm. The second setup is a *studio*-like room with no windows, where we used either ceiling lights or artificial studio-like lights to illuminate a table of size 60x60 cm and height 82 cm.

To acquire multiple images of the same object under different backgrounds, we capture data with the tabletop uncovered and then covered with two different tablecloths. We collected in total 207 configurations that are combinations of objects (23), backgrounds (3) and lighting conditions (3), resulting in 414 RGB images, 414 depth images and 828 IR images. We annotated the largest width and height of each object with a digital caliper (0-150 mm, ±0.01 mm) and a measuring tape (0-10 m, ±0.001 m).

4. EVALUATION AND RESULTS

We compare LoDE with NOCS [15], a state-of-the-art DNN-based approach that uses RGB-D data; and two baselines, which do not require 3D object models and can estimate object dimensions. One baseline uses segmentation on RGB-D data (SegDD). The other baseline is our approach applied to a stereo IR camera with narrow-baseline on a single device (LoDE-IR). SegDD partially replicates the initial part of several DNN-based approaches [5, 15, 17], by using semantic segmentation and then back-projecting in 3D the pixels belonging to the object of interest, using the distance estimation of the depth image. The object dimensions are estimated from the most external points along the x-axis and y-axis, respectively (camera coordinate system). Note that while LoDE is multi-view, NOCS, SegDD and LoDE-IR are single-view. Thus, we report the results of single-view methods as the concatenation of the results from the two cameras. Note that we do not compare with other approaches

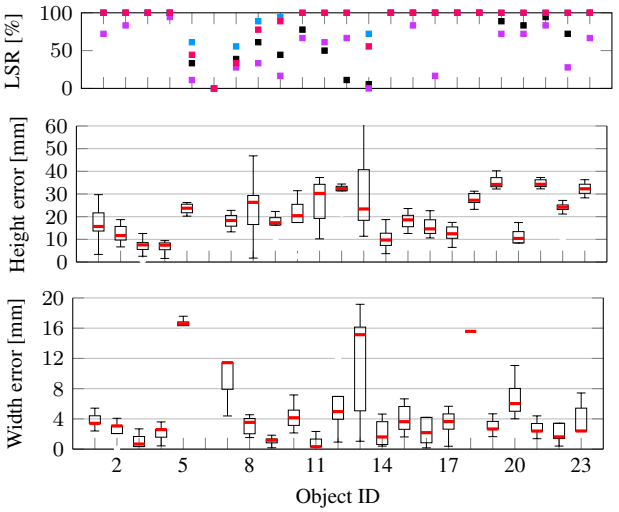

Fig. 4. Localisation success ratio (LSR) of all methods and errors for each dimension using LoDE for each object of the CORSMAL Container dataset, across all backgrounds and lighting conditions. Note the different scale of the y-axis. Legend: NOCS [15] ■, SegDD ■, LoDE-IR ■ and LoDE ■.

for 6 DoF pose estimation, (*e.g.* DenseFusion [5]), or 3D Object Detection, (*e.g.* FrustumNet [8]), as they require the exact 3D model of each object which is not the case of study of this work.

For the semantic segmentation, SegDD, LoDE-IR and LoDE adopt Mask R-CNN [20] trained on the MS COCO dataset [14] of which we consider the classes *cup*, *wine glass*, *bottle* and *vase*. For both LoDE-IR and LoDE, we set $L = 500$ circumferences, separated by 1 mm on height and composed of $N = 20$ points each (18° between point pairs) and we sample the radius of the circumferences, r, across iterations with the following schedule: $150.0, 149.5, \ldots, 1.5, \rho$ (mm), with a minimum circumference radius of $\rho = 1.0$ mm to fit the shape of objects that have a thin stem, (*e.g.* object 12, margarita glass, or object 8, plastic wine glass).

As performance measures, we compute the absolute error between the estimated and annotated width and height of the objects, and the Localisation Success Ratio (LSR), which measures the number of successful object localisations over the number of configurations (either the total number of configurations or a subset).

Fig. 4 shows the statistics (median, min, max, 25 percentile and 75 percentile) of the dimensions error of our approach for each object across all the background and lighting variations. LoDE accurately estimates the width of most of the objects with an error smaller than 20 mm and with small variations across the configurations. Objects 5 (juice glass), 7 (beer cup), 13 (champagne flute) and 18 (small white cup) are the least accurate cases, where the median error is larger than 10 mm. LoDE is less accurate in estimating the object height with the errors varying between ∼10 mm and ∼40 mm. This larger inaccuracy is due to the perspective on the image plane, as circumferences at lower/higher height than the real one are re-sampled with smaller radius to fit within the object masks. Objects 1 (bottle of water), 8 (plastic wine glass), 11 (rum glass) and 13 (champagne flute) show larger variations across configurations than other objects. As width and height are estimated independently, there is no correlation between the two dimensions. While LoDE localises most of the objects across all the configurations (100% LSR), there are some challenging cases, such as objects 5 (juice glass), 7 (beer cup) and 13 (champagne flute), where the LSR is below 60%. Note that cham-

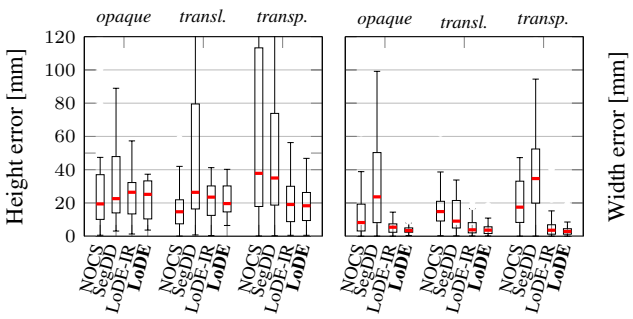

Fig. 5. Estimation error of height and width for opaque, translucent and transparent objects.

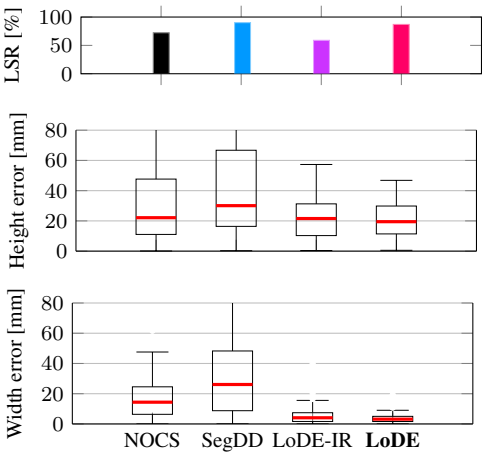

Fig. 6. Localisation success ratio (LSR) and dimension estimation error. Legend: NOCS [15] ■, SegDD ■, LoDE-IR ■ and LoDE ■.

pagne flute is not localised by NOCS and LoDE-IR. The most challenging case for all methods is object 6 (diamond-shaped glass) that is never detected through the semantic segmentation due to the high level of transparency and the unusual shape. Moreover, NOCS and LoDE-IR obtain a lower LSR than LoDE for most of the transparent glasses/cups (*e.g.* objs. 5–13) and the small cups (objs. 18 and 22).

Fig. 5 compares the methods under varying degrees of transparency, such as opaque, translucent and transparent. The error is computed only for the cases where the object is successfully localised. As previously observed for LoDE, we can observe even here that all methods estimate the width more accurately than the height. The top-down perspective of the cameras makes the segmentation treat different parts of the object as one and consequently affects the height estimation when back-projecting in 3D via depth map or triangulation, or projecting for circumference verification. SegDD is more inaccurate for both translucent and transparent objects, with large variations especially in the height, due to the inaccuracies of the depth maps, while NOCS is inaccurate for transparent objects if localised. However, NOCS and SegDD are more accurate in estimating the height for opaque objects, while LoDE-IR and LoDE estimate the dimensions with a median error smaller than 30 mm despite the object transparency.

Fig. 6 shows the success in localising the objects (LSR) and the error in estimating the height and width dimensions, across all the configurations. As previously observed, LoDE outperforms NOCS and SegDD obtaining 2.6 mm and 10.6 mm more accurate height estimations, and 11.2 mm and 22.9 mm more accurate width estimations comparing their medians, respectively, with a smaller standard deviation. LoDE also outperforms LoDE-IR in both height and width estimations. Furthermore, LoDE has a 25% LSR higher than LoDE-IR at similar dimension error. Although both LoDE and SegDD uses Mask R-CNN, the LSR of LoDE is slightly lower than SegDD, as LoDE considers the two views simultaneously, while SegDD works on each view individually.

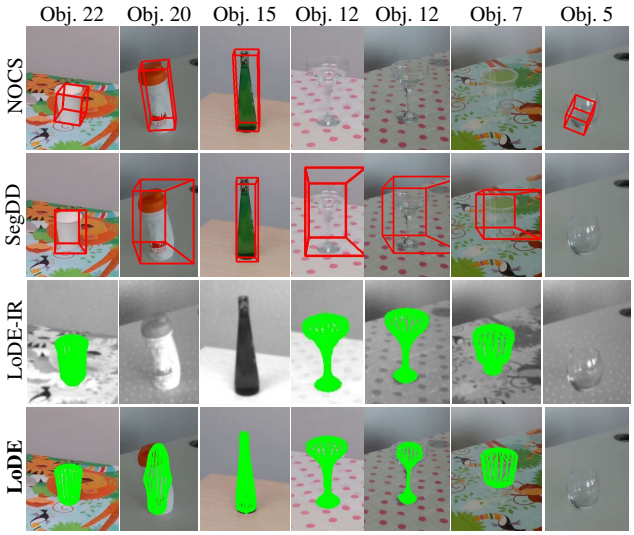

Fig. 7. Sample results for objects with varying transparency, backgrounds and lighting. Fourth and fifth columns correspond to the same object and background but different lighting (artificial and natural, respectively). KEY – Obj.: object.

Fig. 7 compares the results for one opaque and one transparent cup (objs. 22 and 7), one opaque and one translucent bottle (objs. 20 and 15), and two transparent drinking glasses (objs. 12 and 5) under different backgrounds and lighting conditions. All methods accurately estimate the dimensions of the opaque cup (obj. 22). While SegDD, LoDE-IR and LoDE fail to localise obj. 5 (juice glass) under natural light, the bounding box estimated by NOCS is inaccurate. Moreover, NOCS fails to localise two transparent objects (objs. 7 and 12). SegDD shows large inaccuracies for obj. 12 (margarita glass), obj. 7 (beer cup), and obj. 20 (deformed bottle), while LoDE-IR fails for objs. 20 and 15 (translucent bottle). LoDE is less accurate with non-symmetric objects (*e.g.* obj. 20) and under challenging lighting (last three columns), but successfully estimates transparent objects (*e.g.* obj.12, margarita glass).

5. CONCLUSION

We proposed LoDE, a method to estimate the dimensions of container-like objects with circular symmetric shape, without relying on depth information, markers, or 3D models. LoDE uses an iterative multi-view 3D-2D shape fitting algorithm of a generative 3D sampling model, verifying the model on the object image masks of two wide-baseline cameras. To better handle transparent objects, LoDE uses a DNN-based semantic segmentation approach re-trained on selected high-level object classes of containers. For the evaluation, we collected a dataset of containers with different degrees of transparency, and under varying lighting conditions and backgrounds. The object localisation success ratio of LoDE is 86.96% and its average error in estimating the objetc dimensions is smaller than 2 cm. As future work, we will generalise the approach to handle occlusions and generic object shapes under different poses.

6. REFERENCES

[1] T. Wang, C. Yang, F. Kirchner, P. Du, F. Sun, and B. Fang, "Multimodal grasp data set: A novel visual-tactile data set for robotic manipulation," *Int. J. Advanced Robotic Syst.*, vol. 16, no. 1, pp. 1–10, Jan. 2019.

[2] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *Int. J. Robot. Res.*, vol. 27, no. 2, pp. 157–173, 2008.

[3] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Yale-CMU-Berkeley dataset for robotic manipulation research," *Int. J. Robot. Res.*, vol. 36, no. 3, pp. 261–268, 2017.

[4] J. R. Medina, F. Duvallet, M. Karnam, and A. Billard, "A human-inspired controller for fluid human-robot handovers," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, Nov. 2016.

[5] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "DenseFusion: 6d object pose estimation by iterative dense fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 16–20 June 2019.

[6] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "PVNet: Pixel-wise voting network for 6DoF pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 16–20 June 2019.

[7] C. Mitash, A. Boularias, and K. E. Bekris, "Robust 6d object pose estimation with stochastic congruent sets," in *Proc. Brit. Mach. Vis. Conf.*, Newcastle, United Kingdom, 3–6 Sept. 2018.

[8] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 18–22 June 2018.

[9] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, Cambridge University Press, 2003.

[10] S. Kim, A. Shukla, and A. Billard, "Catching objects in flight," *IEEE Trans. Robotics*, vol. 30, no. 5, pp. 1049–1065, Oct. 2014.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[12] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-Net: A trainable CNN for joint description and detection of local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 16–20 June 2019.

[13] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An accurate O(n) solution to the PnP problem," *Int. J. Comput. Vis.*, vol. 81, no. 2, pp. 155–166, 2009.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 8–14 Sept. 2018.

[15] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 16–20 June 2019.

[16] Y. Hu, J. Hugonot, P. Fua, and M. Salzmann, "Segmentation-driven 6d object pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 16–20 June 2019.

[17] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6d object pose estimation in cluttered scenes," in *Proc. Robotics: Science and Syst.*, Pittsburgh, USA, 26–30 June 2018.

[18] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "DeepIM: Deep iterative matching for 6d pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 8–14 Sept. 2018.

[19] P. L. Rosin, "Measuring corner properties," *Comput. Vis. Image Understanding*, vol. 73, no. 2, pp. 291–307, Feb. 1999.

[20] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, 22–29 Oct. 2017.