Hindawi
Complexity
Volume 2020, Article ID 8820500, 7 pages
https://doi.org/10.1155/2020/8820500

WILEY | Hindawi

Research Article

Symmetry-Aware 6D Object Pose Estimation via Multitask Learning

Hongjia Zhang, Junwen Huang, Xin Xu, Qiang Fang, and Yifei Shi

National University of Defense Technology, Changsha, Hunan 410073, China

Correspondence should be addressed to Yifei Shi; yifei.j.shi@gmail.com

Received 6 August 2020; Revised 21 September 2020; Accepted 29 September 2020; Published 21 October 2020

Academic Editor: Zhile Yang

Although 6D object pose estimation has been intensively explored in the past decades, the performance is still not fully satisfactory, especially when it comes to symmetric objects. In this paper, we study the problem of 6D object pose estimation by leveraging the information of object symmetry. To this end, a network is proposed that predicts 6D object pose and object reflectional symmetry as well as the key points simultaneously via a multitask learning scheme. Consequently, the pose estimation is aware of and regulated by the symmetry axis and the key points of the to-be-estimated objects. Moreover, we devise an optimization function to refine the predicted 6D object pose by considering the predicted symmetry. Experiments on two datasets demonstrate that the proposed symmetry-aware approach outperforms the existing methods in terms of predicting 6D pose estimation of symmetric objects.

1. Introduction

6D object pose estimation is of remarkable importance to a variety of industrial applications, ranging from robotic manipulation [1, 2] and autonomous navigation [3] to augmented reality [4]. Serving as the base for the perception of objects in the environment, it concerns the acquisition of the 6D pose (location and orientation) information [2, 5]. The ultimate goal is to achieve speedy real-time 6D object after estimation with robust performance regardless of varying shape, texture, occlusion, illumination, or sensor noise.

The recent development of 6D object pose estimation methods has been promising, thanks to the advancement of economical depth sensors. Existing work has explored the 6D object pose estimation for both household furniture (e.g., chair or table) [6] and table-top objects (e.g., box or book) [7]. These methods are able to generate accurate object pose in many real-world scenarios by utilizing the 3D information from the depth image. Moreover, when integrating with the color image, the RGB-D image-based 6D object pose estimation methods [8] achieved competitive results on objects with complex geometry and moderate occlusion.

However, previous pose estimation approaches are still hardly satisfactory when it comes to dealing with symmetric objects. The reason lies in the fact that a symmetric object might correspond to multiple poses, leading to ambiguity in the training of the neural networks. On the contrary, the symmetry-related feature of the object has been proved to be one of the most informative geometric clues for a variety of applications [9] and has the potential to facilitate the pose estimation task as a complementary element [10]. In this work, we tackle the problem of 6D object pose estimation for objects with reflectional symmetry, which plays a crucial role in a variety of applications [11–13]. Specifically, the symmetry is predicted jointly with the symmetry axis, thus making these two relevant tasks boost each other.

To this end, this paper proposes an approach for 6D object pose estimation that is aware of and regulated by the symmetry axis and the key points of the to-be-estimated objects. During training, the proposed approach learns to predict the 6D object pose, the object symmetry, and the key points in a unified network. In particular, the network contains a multiscale feature extraction module to fuse the appearance feature and the geometric feature at multiple scales. The ground-truth object symmetry of the training

data is generated in a self-supervised manner, so no manual data labelling is required. During testing, we propose to use an optimization function to determine the final prediction on the 6D object pose with the object symmetry. This decision-level optimization boosts the performance on the prediction of 6D object pose by the predicted object symmetry.

We evaluate the proposed method on two datasets: YCB-Video and ShapeNet. Experimental results demonstrate that our method outperforms the state-of-the-art methods on most of the symmetric objects. Also, we provide a qualitative comparison to the baseline method to demonstrate the effects of our symmetry-aware pose estimation approach.

The contributions of this paper are as follows: (1) we introduce a multitask network to estimate the 6D object pose, the symmetry axis, and the key points at the same time; (2) we propose a multiscale feature extraction module to fuse the features from the color image and the depth image; (3) we devise an optimization function to refine the predicted 6D object pose by the predicted symmetry; (4) we show that our approach outperforms the existing methods on 6D object pose estimation of the symmetries objects.

2. Related Work

2.1. 6D Pose Estimation. Given a single RGB image, previous methods estimate the 6D object pose by using either the template matching techniques or end-to-end data-driven neural networks. These methods are limited by various factors, such as occlusion or the existence of ambiguity along the depth direction, and are inadequate for 3D data reasoning [14]. Another type of the 6D pose estimation approach is based on data from range sensors, such as depth camera or LIDAR. Existing approaches typically address this problem by first establishing rough pose candidates by using point features and then performing an iterative closest point (ICP) algorithm to refine and select the optimal pose [15]. Recently, Xiang, Song and Xiao, and Li [5, 7, 8] integrated the features from both the RGB image and the depth map by leveraging the feature fusion techniques. Wada [16] proposed an object-level volumetric fusion to reason 6D pose of multiple objects. These approaches have proven to be fairly robust for scenarios with poor lighting conditions or heavy occlusions. Although we utilize a similar feature fusion approach, our method particularly improves these approaches by introducing a symmetry detection module. We demonstrate how our method outperforms the previous works for 6D pose estimation on symmetric objects.

2.2. 3D Symmetry Detection. 3D symmetry detection has received significant research attention in computer vision and graphics communities for both synthetic and real-world applications. Conceptually, symmetry is well defined in mathematics and is geometrically measurable. Conventional symmetry detection methods [17, 18] mostly use point clustering to detect symmetries of complete geometries (such as CAD models). However, 3D data acquired from sensors are possibly coupled with noise, occlusion, or complex lighting condition. This makes the traditional symmetry detection method incapable. To tackle this problem, Ecins et al. proposed to detect an object from incomplete point cloud [19]. This method can detect symmetries for objects with simple geometry in occluded tabletop scenes, but it is still limited by its inferior generality so cannot be extended to more general object types. Another 3D symmetry detection approach is to first predict the complete geometry of the input data [20] followed by a conventional symmetry detection [21]. The drawback is that it requires the shape completion method to make point-level predictions with high accuracy, which is nontrivial as the training data collection and network training procedures are both effort-intensive. More recently, Shi proposed an end-to-end deep neural network, which is able to predict both reflectional and rotational symmetries from RGB-D images [22]. Our method is inspired by their work. However, the output of our method is not only the symmetry but also the 6D pose.

2.3. Key Point Detection. Efforts have been made to compute 6D pose parameters based on the detected key points via deep neural networks [14, 23–25]. Previously, key point detection on texture-less objects was proven challenging [26–28]. With the recent progress of deep learning, Rad and Lepetit, Tekin et al, and Hu [24, 25, 29] proposed to obtain the coordinates of the 2D key points via direction regression. Methods mentioned above are designed to minimize the 2D projection errors on the objects. However, small projection errors might still be large when it comes to the 3D world. 3D poses are obtained via 3D key points from two views of perspective provided by synthetic RGB images [30]. However, the depth information is missing with only RGB images. The nowadays economical depth sensors allow us to construct, compute, and detect key points in the real 3D world, thanks to the captured depth information.

2.4. Multitask Learning. Multitask learning refers to the approach where multiple objectives corresponding to different tasks with a common representation are learned in parallel simultaneously [31, 32]. It features the advantages of improved efficiency and accuracy respectively in terms of learning and prediction due to the fact that commonalities and differences across tasks are exploited [33]. In addition, it is effective in the avoidance of overfitting on a specific task since the network model is regularized [34]. Wang et al. managed to improve the 6D object pose estimation performance, especially under the condition of occlusion via a multitask learning network combining object recognition with pose estimation [35]. The issue of 6D pose estimation of multiple instances in the bin-picking situation was studied by Sock [36]. He demonstrated outstanding performance of the multitask network which learns depth, 2D detection, and 3D object pose estimation jointly as three subtasks. Xiang et al. proposed PoseCNN where the extracted feature maps are shared by three subtasks, namely, 3D rotation regression, 3D translation estimation, and semantic labelling [5].

3. Method

A 6D pose consists of a position and an orientation both of which are defined based on the camera coordinate frame in this paper. Specifically, a pose is defined by a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$. The representation of a pose is therefore a homogeneous transformation $\mathbf{T} = [\mathbf{R}, \mathbf{t}]$. A reflectional symmetry plane is defined by a point on the plane and its plane normal, i.e., $S = [\mathbf{p}, \mathbf{n}]$, where $\mathbf{p}$ is the location of the point and $\mathbf{n}$ is the plane normal.

3.1. Overview. We propose to estimate 6D object pose and object symmetry in a multitask network (see Figure 1). In particular, the network consists of a multiscale feature extraction module to fuse the features from the RGB image and the depth map. During training, 6D object pose estimation, symmetry prediction, and key point detection components are coupled with each other and trained by a multitask learning strategy.

During testing, we first predict the 6D object pose and object symmetry by a network inference. The predicted pose is then refined by an optimization process which considers the constraints provided by the predicted symmetry and the detected key points.

3.2. Network Architecture

3.2.1. Multiscale Feature Extraction. The input to our method is an RGB-D image which contains at least one object. In our problem setting, the segmentation is precomputed by a segmentation algorithm [5]. For the segmented object, we crop the pixels in the RGB-D image and compute the point cloud by using the intrinsic parameters of the camera.

Our network is derived from the pixelwise dense feature extraction network introduced in [8]. First, the point cloud is fed into a geometric feature extraction network. Different from [8], which uses PointNet as its backbone, we opt to use PointNet++ [37] because of its superior ability on feature extraction for objects with complex geometry. For the RGB image, we use a Resnet-based U-Net to extract pixel-level feature. The difference to [8] is that we enlarge the dilation in the convolution layers so that the network could perceive more context information. We found that this adjustment is of great significance for symmetry prediction.

The multiscale features from the point cloud and the features from the RGB image are subsequently concatenated before being fed to another network to obtain the global feature by using an average pooling layer. The pixel-level feature is then concatenated with the global feature to form the overall pixelwise features which are in the end used to predict the 6D object pose and the symmetry as well as the key points.

3.3. Loss Function. The multitask learning network comprises the pose predictor, the symmetry predictor, and the key point predictor whose losses are embedded into the overall loss function so that the symmetry and key point information can serve as additional regulations to the learning process for the pose prediction. In the end, the results of the 6D pose estimation and the symmetry estimation are output in the format of $\mathbf{T}$ and S. We define the symmetric transformation of the predicted symmetry S as $\mathbf{T}_s = [\mathbf{R}_s, \mathbf{t}_s]$.

The overall loss of the network training is the sum of the loss of point-level predictions. For each point, the loss consists of a pose estimation loss, a symmetry prediction loss, and a key point detection loss:

$$\mathscr{L} = \frac{1}{N}\sum_{i}^{N}\mathscr{L}_i = \frac{1}{N}\sum_{i}^{N}\left(\mathscr{L}_i^{\text{pose}} + \mathscr{L}_i^{\text{symmetry}} + \mathscr{L}_i^{\text{keypoint}}\right), \tag{1}$$

where N is the total number of the points. The 6D pose estimation loss $\mathscr{L}_i^{\text{pose}}$ is defined as the average distance between the sampled points on the object transformed by the ground-truth pose and by the predicted pose of the i-th point:

$$\mathscr{L}_i^{\text{pose}} = \frac{1}{M}\sum_{j}^{M}\left\|\left(\mathbf{R}\mathbf{x}_j + \mathbf{t}\right) - \left(\widehat{R}\mathbf{x}_j + \widehat{t}\right)\right\|, \tag{2}$$

where M is the number of the sampled points, $\mathbf{x}_j$ is the j-th point of the sampled points, and $\widehat{T} = [\widehat{R}, \widehat{t}]$ is the ground-truth pose.

The symmetry prediction loss $\mathscr{L}_i^{\text{symmetry}}$ is the average distance between the sampled points on the object transformed by the ground-truth symmetric transformation and the predicted symmetric transformation:

$$\mathscr{L}_i^{\text{symmetry}} = \frac{1}{M}\sum_{j}^{M}\left\|\left(\mathbf{R}_s\mathbf{x}_j + \mathbf{t}_s\right) - \left(\widehat{R}_s\mathbf{x}_j + \widehat{t}_s\right)\right\|, \tag{3}$$

where $\widehat{T}_s = [\widehat{R}_s, \widehat{t}_s]$ is the ground-truth symmetric transformation.

Similar to [38], the key point detection loss $\mathscr{L}_i^{\text{keypoint}}$ is the sum of the offset distances between the sampled points and the key points:

$$\mathscr{L}_i^{\text{keypoint}} = \frac{1}{M}\sum_{j}^{M}\sum_{p}^{K}\left\|d_j^p - \widehat{d}_j^p\right\|, \tag{4}$$

where d_j^p and $\widehat{d}_j^p$ are respectively the offset distance and the corresponding ground-truth between the j-th point and the p-th key point. Note that our key point detection module is different from [38], as our key points not only contain the points selected by the farthest point sampling but also their symmetric counterparts.

3.4. Multitask Network Training. The three subtasks, i.e., pose prediction, symmetry prediction, and key point detection, share the same pixelwise feature maps extracted in prior and are trained in parallel jointly. The symmetry prediction task serves as an additional metric to reveal the quality of the pointwise features, hence aiding to boost the accuracy of the overall pose estimation task.

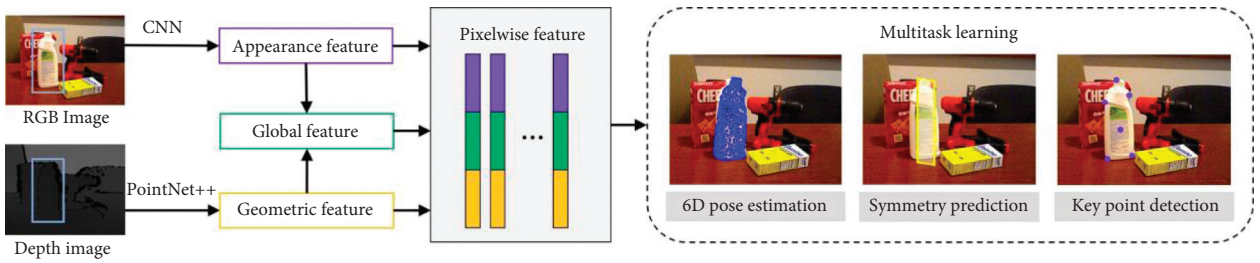


Figure 1: Architecture of the symmetry-aware object pose estimation network.

Figure 2: Examples of images from YCB-Video and ShapeNet datasets used for the experiments.

3.5. Inference. During inference, we first extract the point-level features and make point-level 6D object pose and symmetry predictions. By averaging all the predictions, the ultimate prediction is generated. The design of multitask learning on the 6D object pose and the symmetry has made the two subtasks regulated by each other. However, we observe from the experiments that (1) the 6D object pose prediction and the symmetry prediction are not perfectly consistent with each other; (2) the error of symmetry prediction is noticeably smaller than the error of 6D object pose prediction, illustrating that the predicted symmetry could be further used to refine the predicted pose. To this end, we introduce an optimization function as follows to refine the predicted pose by considering the constraints provided by the predicted symmetry:

$$\arg\min_{T} \frac{1}{M} \sum_{j}^{M} \left\| \mathbf{T}_s\left(\mathbf{T}\left(\mathbf{x}_j\right)\right) - \hat{T}_s\left(\hat{T}\left(\mathbf{x}_j\right)\right) \right\|, \tag{5}$$

where $\mathbf{T}(\mathbf{x}_j)$ represents the location of transformed $\mathbf{x}_j$ by $\mathbf{T}$:

$$\mathbf{T}\left(\mathbf{x}_j\right) = \mathbf{R}\mathbf{x}_j + \mathbf{t}. \tag{6}$$

We use Ceres Solver [39] to optimize the above function. We consider the $\boldsymbol{T}$ after the optimization as the final 6D object pose of our method.

4. Results

4.1. Benchmark. We create a benchmark to evaluate our method. The benchmark is built based on two datasets: YCB-Video [40] and ShapeNet [41]. YCB-Video consists of 92 RGB-D videos captured in indoor scenes with 21 different table-top objects. The images in the dataset are annotated with object pose. We compute the ground-truth for each object by using an offline symmetry detection method [21]. ShapeNet is a large-scale CAD model dataset with category-label annotations. To generate the training and testing data, we first perform a virtual scanning on the CAD model from random viewpoints around the object and then compute the ground-truth object pose and the symmetry. Note that,

Table 1: Quantitative comparison of 6D object pose estimation on the YCB-Video dataset.

Datasets	AUC	<2 cm
PointFusion	78.4	71.8
PoseCNN	87.5	88.0
DenseFusion	90.3	91.6
Ours	90.6	91.6

Table 2: Quantitative comparison of 6D object pose estimation on the ShapeNet dataset.

Datasets	AUC	<2 cm
PointFusion	72.4	68.8
PoseCNN	74.9	73.5
DenseFusion	77.2	79.6
Ours	80.8	82.3

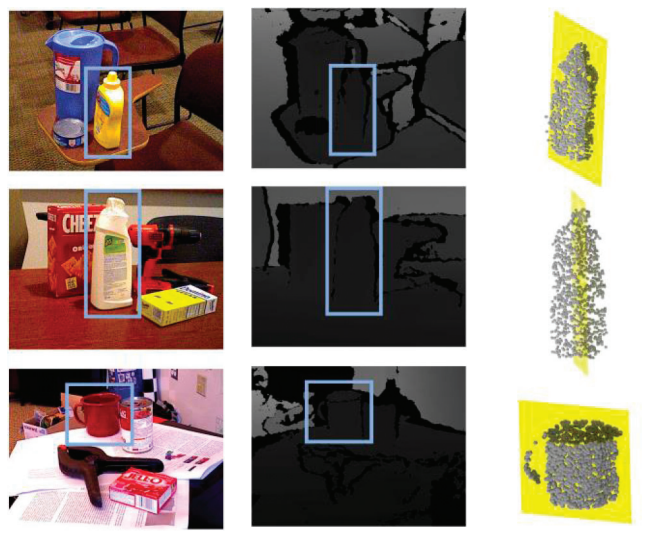
DenseFusion

(a)

Ours

(b)

Figure 3: Qualitative comparison on 6D pose estimation performance between the proposed approach and previous work [8] with the YCB-Video dataset. Our method achieves more accurate pose estimation on a variety of objects.



RGB image | Depth image | Predicted symmetry

(a) (b) (c)

Figure 4: Qualitative results of the predicted symmetry on the YCB-Video dataset.

different from [22], we only use the objects with one reflectional symmetry.

In order to generate images with the occlusions and background clutter, we randomly select objects from both images of other ShapeNet objects or images from the real-world scenes [42]. The examples of the two datasets are shown in Figure 2.

4.2. Evaluation Metrics. We evaluate the 6D object pose using the average closest point distance (ADD-S) proposed in [5]. Specifically, we report the area under the ADD-S curve (AUC). Given the ground-truth pose and the predicted one, ADD-S measures the mean distance between each sample point on the object transformed by the ground-truth pose and its closest neighbouring point among the sample points transformed by the predicted pose. We set the AUC threshold as 0.1 m. We also evaluate the percentage of predictions whose ADD-S is smaller than 2 cm.

4.3. Comparison to Baselines. We compare our method with three baselines: PointFusion [43], PoseCNN [5], and DenseFusion [8]. The quantitative results are shown in Table 1 (YCB-Video) and Table 2 (ShapeNet). It is clear that our method demonstrates the best results on both YCB-Video and ShapeNet. In particular, our method outperforms all the baselines on ShapeNet by a large margin. Given the fact that most of the objects in ShapeNet are symmetric, we therefore reckon that our proposed method is especially suitable for symmetric objects.

4.4. Qualitative Results. To demonstrate the advantages of our method, we show the qualitative results of our method and DenseFusion on YCB-Video in Figure 3. It shows that our method is able to successfully produce accurate 6D object pose on cases where DenseFusion fails. We also visualize the predicted symmetry in Figure 4.

5. Conclusion

In this paper, we focus on the problem of boosting 6D object pose estimation by leveraging object symmetry. We propose a network that predicts 6D object pose, object symmetry, and key points through multitask learning. The predicted 6D object pose is then refined by the predicted object symmetry via an optimization function. We evaluate our method using both quantitative and qualitative comparisons to the state-of-the-art approaches. Experimental results show that our method outperforms the three baseline approaches, particularly by a large margin in the case of ShapeNet where most objects are symmetric. For future work, we are interested in integrating other relevant geometry clues into the pose estimation network [22, 44]. It is possible to reduce the size of the network and improve accuracy simultaneously, by considering relevant geometric mechanisms [44].

Data Availability

Data that support the findings of this study are available in the website https://http://www.ycbbenchmarks.com/https://www.shapenet.org/.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grant nos. 62002379 and 61973311), by the College of Intelligence Science and Technology, National University of Defense Technology (Youth Innovation Project nos. 2020006 and 2020008), and by the National Key R&D Program of China (Grant no. 2018YFB1305105).

References

[1] M. Zhu, "Single image 3D object detection and pose estimation for grasping," in *Proceedings of 2014 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, Hong Kong, China, May 2014.

[2] A. Collet, M. Martinez, and S. S. Srinivasa, "The MOPED framework: object recognition and pose estimation for manipulation," *The International Journal of Robotics Research*, vol. 30, no. 10, pp. 1284–1306, 2011.

[3] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Providence, RI, USA, June 2012.

[4] E. Marchand, H. Uchiyama, and F. Spindler, "Pose estimation for augmented reality: a hands-on survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 12, pp. 2633–2651, 2015.

[5] Y. Xiang, "Posecnn: a convolutional neural network for 6d object pose estimation in cluttered scenes," 2017, https://arxiv.org/abs/1711.00199.

[6] S. Song and J. Xiao, "Deep sliding shapes for amodal 3D object detection in rgb-d images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 2016.

[7] Y. Li, "Deepim: deep iterative matching for 6D pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, September 2018.

[8] C. Wang, "Densefusion: 6D object pose estimation by iterative dense fusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, CA, USA, June 2019.

[9] J. Zhu, "Body symmetry and part-locality-guided direct nonparametric deep feature enhancement for person reidentification," *IEEE Internet of Things Journal*, vol. 7, no. 3, pp. 2053–2065, 2019.

[10] S. Dieleman, J. De Fauw, and K. Kavukcuoglu, "Exploiting cyclic symmetry in convolutional neural networks," 2016, https://arxiv.org/abs/1602.02660.

[11] S. J. Lederman and A. M. Wing, "Perceptual judgement, grasp point selection and object symmetry," *Experimental Brain Research*, vol. 152, no. 2, pp. 156–165, 2003.

[12] D. Schiebener, "Heuristic 3D object shape completion based on symmetry and scene context," in *Proceedings of 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, Daejeon, South Korea, October 2016.

[13] J. Varley, "Shape completion enabled robotic grasping," in *Proceedings of 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, Vancouver, Canada, September 2017.

[14] S. Peng, "PVNet: pixel-wise voting network for 6D of pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Singapore, Singapore, May 2019.

[15] B. Drost, "Model globally, match locally: efficient and robust 3D object recognition," in *Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, San Juan, PR, USA, June 2010.

[16] K. Wada, "MoreFusion: multi-object reasoning for 6D pose estimation from volumetric fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, September 2020.

[17] Y. Lipman, "Symmetry factored embedding and distance," in *Proceedings of ACM SIGGRAPH 2010 Papers*, pp. 1–12, Los Angeles, CA, USA, July 2010.

[18] N. J. Mitra, L. J. Guibas, and M. Pauly, "Partial and approximate symmetry detection for 3D geometry," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 560–568, 2006.

[19] A. Ecins, C. Fermüller, and Y. Aloimonos, "Seeing behind the scene: using symmetry to reason about objects in cluttered environments," in *Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, Madrid, Spain, October 2018.

[20] M. Liu, "Morphing and sampling network for dense point cloud completion," 2019, https://arxiv.org/abs/1912.00280.

[21] B. Li, "Efficient view-based 3D reflection symmetry detection," in *Proceedings of SIGGRAPH Asia 2014 Creative Shape Modeling and Design*, pp. 1–8, New York, NY, USA, December 2014.

[22] Y. Shi, "SymmetryNet: learning to predict reflectional and rotational symmetries of 3D shapes from single-view RGB-D images," 2020, https://arxiv.org/abs/2008.00485.

[23] G. Pavlakos, "6D of object pose from semantic keypoints," in *Proceedings of 2017 IEEE International Conference on Robotics*

and Automation (ICRA), IEEE, Singapore, Singapore, May 2017.

[24] M. Rad and V. Lepetit, "BB8: a scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth," in *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, October 2017.

[25] B. Tekin, S. N. Sinha, and P. Fua, "Real-time seamless single shot 6D object pose prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, UT, USA, June 2018.

[26] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, IEEE, Kerkyra, Greece, September 1999.

[27] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, "3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints," *International Journal of Computer Vision*, vol. 66, no. 3, pp. 231–259, 2006.

[28] H. Bay, T. Tuytelaars, and L. Van Gool, *Surf: Speeded Up Robust Features: European Conference on Computer Vision*, Springer, Berlin, Germany, 2006.

[29] Y. Hu, "Segmentation-driven 6d object pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[30] S. Suwajanakorn, "Discovery of latent 3D keypoints via end-to-end geometric reasoning," in *Proceedings of the Advances in Neural Information Processing Systems*, Montréal, Canada, December 2018.

[31] R. Caruana, "Multitask learning: a knowledge-based source of inductive bias ICML," in *Proceedings of the Tenth International Conference on Machine Learning*, Amherst, MA, USA, July 1993.

[32] A. Rastogi and S. Sampath, "Multi-task learning via linear functional strategy," *Journal of Complexity*, vol. 43, pp. 51–75, 2017.

[33] R. Caruana, "Multitask Learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[34] K. Park, "Multi-task template matching for object detection, segmentation and pose estimation using depth images," in *Proceedings of 2019 International Conference on Robotics and Automation (ICRA)*, IEEE, Montreal, QC, Canada, May 2019.

[35] Y. Wang, S. Jin, and Y. Ou, "A multi-task learning convolutional neural network for object pose estimation⋆," in *Proceedings of 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, IEEE, Dali, China, December 2019.

[36] J. Sock, "Multi-task deep networks for depth-based 6d object pose and joint registration in crowd scenarios," 2018, https://arxiv.org/abs/1806.03891.

[37] C. R. Qi, "Pointnet++: deep hierarchical feature learning on point sets in a metric space," in *Proceedings of Advances in Neural Information Processing Systems*, Long Beach, CA, USA, December 2017.

[38] Y. He, "PVN3D: a deep point-wise 3D keypoints voting network for 6DoF pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, WA, USA, June 2020.

[39] S. Agarwal and K. Mierle, "Ceres solver," 2010, http://ceres-solver.org/.

[40] B. Calli, "Benchmarking in manipulation research: the ycb object and model set and benchmarking protocols," 2015, https://arxiv.org/abs/1502.03143.

[41] A. X. Chang, "Shapenet: an information-rich 3D model repository," 2015, https://arxiv.org/abs/1512.03012.

[42] A. Dai, "Scannet: richly-annotated 3D reconstructions of indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 2017.

[43] D. Xu, D. Anguelov, and A. J.. Pointfusion, "Deep sensor fusion for 3D bounding box estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.

[44] X. Tang, "Model migration neural network for predicting battery aging trajectories," *IEEE Transactions on Transportation Electrification*, vol. 6, no. 2, pp. 363–374, 2020.