ZePHyR: Zero-shot Pose Hypothesis Rating

Brian Okorn*, Qiao Gu*, Martial Hebert and David Held

***Abstract*—Pose estimation is a basic module in many robot manipulation pipelines. Estimating the pose of objects in the environment can be useful for grasping, motion planning, or manipulation. However, current state-of-the-art methods for pose estimation either rely on large annotated training sets or simulated data. Further, the long training times for these methods prohibit quick interaction with novel objects. To address these issues, we introduce a novel method for zero-shot object pose estimation in clutter. Our approach uses a hypothesis generation and scoring framework, with a focus on learning a scoring function that generalizes to objects not used for training. We achieve zero-shot generalization by rating hypotheses as a function of unordered point differences. We evaluate our method on challenging datasets with both textured and untextured objects in cluttered scenes and demonstrate that our method significantly outperforms previous methods on this task. We also demonstrate how our system can be used by quickly scanning and building a model of a novel object, which can immediately be used by our method for pose estimation. Our work allows users to estimate the pose of novel objects without requiring any retraining. Additional information can be found on our website https://bokorn.github.io/zephyr/**

I. Introduction

6D pose describes the position and orientation of an object, defined in a reference frame relative to a predefined model of the object. An object's 6D pose fully describes the state of a static rigid object and, as such, is commonly used as a representation for planning [6, 44]. A robot can use an estimate of an object's pose to perform complex manipulation interactions with the object [19, 34, 11, 5].

Current state-of-the-art methods for object pose estimation train a new model for each object they are being evaluated on [40, 38, 2]. This requires a large amount of annotated training data, either produced by capturing and annotating large datasets [40, 38] or through rendering the object in synthetically generated scenes [35, 7, 33]. Regardless of how this data is obtained, training new networks has a time and space cost, which makes the algorithm not scale well in cases where robots need to interact with many different types of objects. One approach to mitigate these issues is to use a non-learned geometry-based method [8, 37]. These methods, however, do not typically capture visual texture well, and they rely on hard-coded, rather than learned, invariances, which limits the potential accuracy of the system (based on our experiments in Section IV-D). A few recent learning-based approaches have attempted to perform zero-shot object pose estimation [28, 32] but these methods require instance segmentation masks to be provided as input, which limits their use in a "zero-shot" system, as such masks are typically trained per-object.

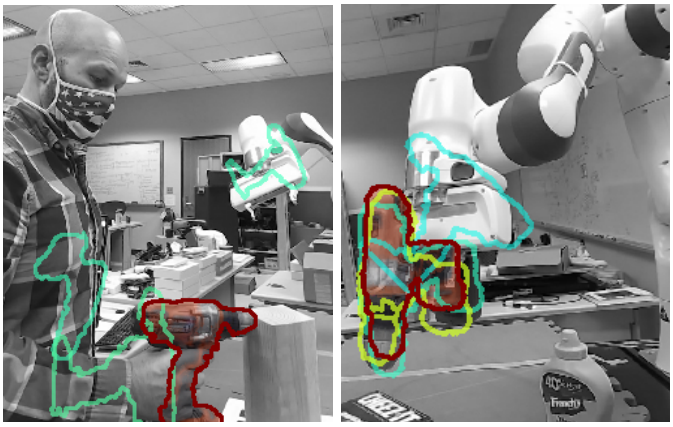

Fig. 1: Pose hypotheses scored using Zero-shot Pose Hypothesis Rating on novel drill object, reconstructed at test time. The highest scoring pose is rendered in color. Poses are outlined in color corresponding to score, with highly-rated poses in red and to lower ones in blue.

We seek to remove these limitations by developing a novel learning-based method for zero-shot object pose estimation that can handle both textured and untextured objects in cluttered scenes and does not require object masks as input. Our method uses the paradigm of pose hypothesis generation and evaluation: given a scene, a large number of candidate poses consistent with the observation are generated. The fitness of each hypothesis is then evaluated and the best-fit candidate is selected. Such an approach requires the hypothesis rating function to give appropriate weight to the features that most correlate with the correct pose. The variation between sensor data and the object model, caused by sensor noise or lighting changes, as well as partial occlusions, can make designing this scoring function challenging. Past approaches to hypothesis scoring have used voting over hypotheses or feature matching [10, 3, 8]; in contrast, this paper proposes a scoring function that learns to compare the observed images and rendered model points. Our learned scoring function demonstrates a significant improvement on zero-shot object pose estimation over a wide set of objects and environmental variations.

The key insight of our method is to use a learned scoring function that compares the sensor observation to a sparse rendering of each candidate pose hypothesis. This scoring

* indicates equal contribution.

This work was supported by NASA NSTRF, United States Air Force and DARPA under Contract No. FA8750-18-C-0092, National Science Foundation under Grant No. IIS-1849154, and LG Electronics

B. Okorn, Q. Gu, M. Hebert and D. Held are with the Robotics Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213. (bokorn@andrew.cmu.edu, qiaog@andrew.cmu.edu, hebert@cs.cmu.edu, dheld@andrew.cmu.edu)

function receives as input *an unordered set of point differences*, shown in Fig. 2, which we show is crucial to perform zero-shot generalization to novel objects not seen in the training set. Our method is trained over a disparate set of objects and then evaluated on novel objects not included in the training set.

We demonstrate that our Zero-shot Pose Hypothesis Rating method (ZePHyR) works on objects in clutter without requiring object masks as input, unlike past zero-shot methods [32, 28]. ZePHyR handles both untextured objects as well as objects with significant visual texture, not seen at training time. Therefore, ZePHyR achieves the goal of zero-shot object pose estimation mentioned earlier:

- We require no new human annotations or large-scale synthetic data generation to interact with novel objects.
- We require no retraining for novel objects.
- ZePHyR uses only a single set of network weights, rather than requiring new weights for each unique object, reducing the memory constraints.

We evaluate our method on YCB-Video and LineMOD-Occlusion, two challenging pose estimation datasets. Our method achieves state-of-the-art results over previous zero-shot pose estimation methods.

II. Related Work

A. Non-learned Zero-shot Pose Estimation

Zero-shot pose estimation is the task of estimating the pose of objects not seen at training time. Non-learning based approaches [1, 36, 13, 14, 26, 22, 25, 12, 43] are inherently zero-shot, leveraging robust features and the available object model at test time. Point Pair Features (PPF) [37, 8, 9, 18, 9, 15] use pairs of oriented points to generate geometrically consistent pose hypotheses and select the best hypothesis using voting and clustering. These are the top-performing zero-shot methods on the BOP leader board [16], when averaged over all datasets, but struggle to compete with deep learned methods on the highly textured YCB dataset due to the methods being exclusively based on depth.

B. Learned Zero-shot Object Pose Estimation

Several learned methods solve the zero-shot pose estimation problem using class-based pose estimation [24, 39] as opposed to instance-based pose estimation. These methods learn a pose estimator capable of generalizing among objects in the given class, but such methods are not intended to generalize to novel classes. While this is a step in the direction of zero-shot pose estimation, it still requires training a new network for each class.

A few recent zero-shot methods use a learned representation of the object in their pose estimation pipeline [41, 32, 28]. While these methods have been shown to generalize across objects, they require a bounding box for the target object, which is obtained using an object-specific learned detector (and hence not a zero-shot system) or the ground-truth bounding box. This requirement is avoided in the MOPED dataset [28], as there is only a single object in the scene, which greatly simplifies the task of estimating the object mask [42]. For the LineMOD-Occlusion dataset, ground truth object masks are used [28]. Our method does not require such bounding boxes or masks as input, making it truly zero-shot.

III. Method

A. Overview

The primary objective of this work is zero-shot object pose estimation in clutter. To achieve this, we train our pose estimation method on one set of objects and then evaluate on a set of novel objects, without requiring any re-training.

An overview of our method is shown in Figure 2. Given a set of 6D pose hypotheses, we first project each hypothesis into the scene. Our method learns to score each hypothesis by comparing differences in the projected object model point cloud to the RGB-D observation. For each projected model point, we extract the color and geometry information from both the model and the observation and compute the local differences of the extracted information. This yields a set of *point-differences*, one for each projected model point. Each element in this set encodes the local alignment between the model and the observation with respect to color and geometry. We adopt a point-based network [30, 31] to analyze this unordered set of point-differences and regress to an overall score for each pose hypothesis. Focusing on differences as well as adopting a point-based neighborhood structure helps us avoid overfitting to object-specific properties from the training set and allows us to generalize to unseen objects at test time.

In this work, our primary focus is the learned scoring function and we use a combination of Point Pair Features [8] and SIFT features [23] to generate our pose hypothesis set.

B. Learned Scoring Function

The main goal of our method is to score pose hypotheses by projecting them into the observed scene and learning to compare their local geometric and color differences. Suppose that we have a set of 6D pose hypotheses $\mathcal{H} = \{h_i\}_{i=1}^{m}$ that we wish to evaluate. We represent the object as a point cloud $\mathcal{M} = \{x_j\}_{j=1}^{n}$, sampled from the provided object mesh model, or obtained from a 3D reconstruction pipeline. Each point contains both geometric (depth and normal) and color information drawn from its local region on the object. Similarly the observation image I contains geometric and color values from the observation. To evaluate hypothesis h_i, we project each object point x_j onto the observation's image plane, using the known camera parameters. This projection gives a point at image coordinates y_{ij} with transformed point values $\tilde{x}_{ij}$ (the point depth and normal vector are transformed; the color of the projected point is unchanged). For each pose hypothesis, the difference between the projected values, $\tilde{x}_{ij}$, and their corresponding image values, $I(y_{ij})$, is computed according to a simple distance function, $d_{ij} = f(\tilde{x}_{ij}, I(y_{ij}))$ (see supplementary material for details).

The set $D_i = \{d_{ij}\}_{j=1}^{m}$ represents an unordered set of point differences for pose hypothesis h_i, each of which is associated with a given point x_j in the model and a location y_{ij} in the observation image. We train a deep neural network $g_\theta(D_i)$

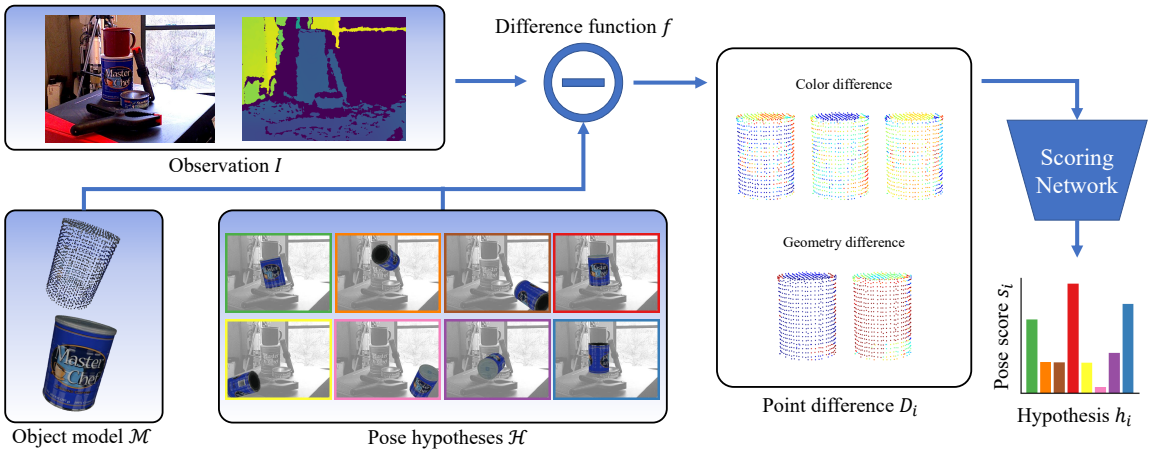

Fig. 2: System Pipeline. Our method first projects the sampled model points $\mathcal{M}$ onto the observation I according to a pose hypothesis h_i. Then D_i are extracted as the point-wise differences between the observation and the projected model points, describing the alignment of the pose hypothesis at each projected point. A network takes in D_i and regresses to a score s_i for each pose h_i which evaluates how well the pose matches the observation.

with parameters θ to analyze this difference set and regress to a pose fitness score, s_i. Our learned function can intelligently combine point differences on multiple parts of the object to robustly estimate the most likely pose hypothesis.

C. Loss Function

To train this hypothesis scoring function, we adopt the probabilistic selection loss proposed by DSAC [3], as it directly optimizes the expected pose error when hypotheses are sampled according to the predicted scores. For each pose hypotheses h_i with corresponding true pose error ϵ_i, we compute the expected pose error of sampling according to the softmax distribution induced by s_i, $\mathcal{L} = \sum_{i=1}^{m} softmax(s_i)\epsilon_i$.

In our experiment, ϵ_i is defined as the log of the average point distance (ADD) for non-symmetric objects and its symmetric analog (ADD-S) for symmetric ones [40]. Empirically, we find that the log of this error better dampens the effects of outliers. At test time, the highest-scoring pose hypothesis is selected. The inference pipeline is described in Algorithm 1.

Algorithm 1: Hypothesis Scoring Pose Estimation

Compute initial pose hypothesis set $\mathcal{H} = \{h_i\}_{i=1}^{m}$;
foreach h_i *in* $\mathcal{H}$ **do**
  Project all model points according to h_i onto the image plane to get projected model points $\tilde{x}_{ij}$ at projected image coordinates y_{ij};
  Get observation points $I(y_{ij})$;
  Compute point differences $d_i = f(\tilde{x}_{ij}, I(y_{ij}))$;
  Score point-differences $s_i = g_\theta(\{d_{ij}\}_{j=1}^{m})$;
end
Return hypothesis h_{i^*}, where $i^* = \arg\max_i s_i$;

D. Implementation details

Implementation details about hypothesis generation, network input construction and network structure can be found in the supplementary material.

IV. Experiments

A. Datasets

We evaluated our method on two of the most popular datasets in the BOP Challenge [16], the YCB-Video (YCB-V) dataset [40] and the LineMOD-Occlusion (LM-O) dataset [2]. In these experiments, we follow the evaluation protocol set up by the BOP Challenge, with the additional constraint that our method is not trained on the objects it is tested on. This allows us to test our ability to perform zero-shot generalization to novel objects.

YCB-Video dataset (YCB-V) [40] contains 92 RGB-D video sequences of 21 YCB objects [4] of varying shape and texture, annotated with 6D poses. This a particularly challenging dataset for object pose estimation due to its varying lighting conditions, occlusions, and sensor noise. We follow the dataset split in [40], and for the evaluation, we adopt the BOP testing set [16], where 75 images with higher-quality ground-truth poses from each of 12 test videos are used. To demonstrate the generalization ability, one half of the objects are used for training, and the other half are used for testing. Then, a second network is trained with train and test objects exchanged. When evaluating on YCB-V, we use hypotheses generated by both PPF and SIFT matching to handle the high degree of visual texture. We also adopt an ICP refinement step [1] for post-processing.

LineMOD-Occlusion dataset (LM-O) [2] adopted a single scene from the test set of the larger LineMOD (LM) dataset [14] and provides ground-truth 6D pose annotations for 8 low-textured objects. For training, we used the PBR-BlenderProc4BOP [17] training images provided by the BOP challenge. Our model is only trained on synthetic images of the 7 objects that are in the LM dataset but *not* in the LM-O dataset; we then evaluate on the LM-O objects, which were not seen at training time. When evaluating on LM-O, we only use hypotheses generated by PPF; we find that SIFT hypotheses are ineffective on this dataset since the objects do not contain much visual texture.

Input Image | Original PPF Results | Our Improved Results

Fig. 3: Qualitative results on image from YCB-V dataset showing the improved accuracy of our method.

	Zero-Shot Methods			Object Specific Methods	
	Drost [8]	Vidal [37]	ZePHyR + Drost (Ours)	CosyPose [20]	Pix2Pose [29]
YCB-V	0.344	0.450	**0.516**	0.861	0.675
LM-O	0.527	0.581	**0.598**	0.714	0.588

TABLE I: AR scores for methods of zero-shot and object specific pose estimation on object pose datasets (higher is better).

B. Metrics

As suggested by the BOP challenge, we report the average recall (AR) scores for the following metrics: Visible Surface Discrepancy (VSD), Maximum Symmetry-Aware Surface Distance (MSSD), and Maximum Symmetry-Aware Projection Distance (MSPD). For a detailed formulation of each metric, please refer to the supplementary material and [16].

C. Baselines

We compare our method to both zero-shot and object-specific methods. As we are most concerned with our performance as compared to other zero-shot methods, we compare to two variants of Point Pair Features, Drost [8] and Vidal [37]. An implementation of Drost's PPF [27] is used as the hypothesis generation algorithm in our work. Vidal had until recently been the top-performing method in the BOP challenge, and demonstrates the peak performance of PPF-only systems (although their code is not available). In addition to the zero-shot baselines, we report the current state of the art in object-specific methods as CosyPose [20] and Pix2Pose [29]. Both of these methods train a network on annotated instances of the test objects and have weights specifically associated with each object. While we are not attempting to match the performance of these systems, we report their results to illustrate the still remaining gap between zero-shot and object-specific methods.

D. Zero-shot Pose Estimation Results

In Table I, we find that our method outperforms all zero-shot methods, significantly improving over our initial pose hypotheses produced by Drost and outperforming the best PPF-only solution in Vidal [37]. We see the largest improvement on the YCB dataset, where PPF is unable to fully resolve the pose of the geometrically symmetric but textually asymmetric objects, seen in failure to match the cylindrical objects in Figure 3. Our method is able to leverage both color and geometry, selecting the most accurate pose hypothesis. Additionally, we find our method to be comparable to the object-specific results produced by Pix2Pose [29]. While DeepIM [21] is a local refinement method, and not directly comparable to ZePHyR, we do evaluate its performance based on PPF in the supplementary material.

E. Input Ablations

To determine the relative importance of each of our input channels, we retrain our networks without each dimension. We show results on YCB in Table II, training on the "Object Set 1" and testing on "Object Set 2". Additionally, this table shows the effects of concatenating observation and model inputs ("Model without Diff"), as opposed to computing their difference (as in our method). Unsurprisingly, the color information has the greatest effect on the accuracy of our system, as it is the most orthogonal to the information used by our PPF hypotheses.

	Model without				
	Color	Depth	Normal	Coords	Diff
Unseen Objects **(Zero-shot)**	-18%	-15%	-7.1%	-8.9%	-6.3%
Seen Objects (Training)	-24%	-4.2%	0.8%	1.1%	2.1%

TABLE II: Percent change in AR scores on YCB Video dataset caused by removal of each input to our method.

V. Conclusion

We propose a method for zero-shot object pose estimation, focusing on pose hypothesis scoring. By extracting point differences between the projected object points and the observation and imposing a loose neighborhood structure on these points, we learn a pose scoring function that generalizes well to novel objects. On the challenging YCB-Video and LineMOD-Occlusion datasets, our method achieves state-of-the-art performance for zero-shot object pose estimation in clutter, evaluated on both textured and untextured objects. We hope that our method paves the way for roboticists to obtain accurate pose estimates for novel objects without needing additional training or data annotation.

References

[1] P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. *TPAMI*, 14(2):239–256, 1992. doi: 10.1109/34.121791.

[2] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *ECCV*, pages 536–551, Cham, 2014.

[3] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *CVPR*, pages 6684–6692, 2017.

[4] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 International Conference on Advanced Robotics (ICAR)*, pages 510–517, 2015.

[5] Matei Ciocarlie, Kaijen Hsiao, Edward Gil Jones, Sachin Chitta, Radu Bogdan Rusu, and Ioan A Şucan. Towards reliable grasping and manipulation in household environments. In *Experimental Robotics*, 2014.

[6] Neil T Dantam, Zachary K Kingston, Swarat Chaudhuri, and Lydia E Kavraki. Incremental task and motion planning: A constraint-based approach. In *RSS*, 2016.

[7] Xinke Deng, Arsalan Mousavian, Yu Xiang, Fei Xia, Timothy Bretl, and Dieter Fox. Poserbpf: A rao-blackwellized particle filter for 6d object pose estimation. In *RSS*, 2019.

[8] B. Drost, M. Ulrich, N. Navab, and S. Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In *CVPR*, pages 998–1005, 2010.

[9] Bertram Drost and Slobodan Ilic. 3d object detection and localization using multimodal point pair features. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 9–16. IEEE, 2012.

[10] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[11] Corey Goldfeder, Matei Ciocarlie, Hao Dang, and Peter K Allen. The columbia grasp database. In *ICRA*, 2009.

[12] Z. Guo, Z. Chai, C. Liu, and Z. Xiong. A fast global method combined with local features for 6d object pose estimation. In *2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 1–6, 2019. doi: 10.1109/AIM.2019.8868409.

[13] S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, N. Navab, P. Fua, and V. Lepetit. Gradient response maps for real-time detection of textureless objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5): 876–888, 2012. doi: 10.1109/TPAMI.2011.206.

[14] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *ACCV*, pages 548–562, Berlin, Heidelberg, 2013.

[15] Stefan Hinterstoisser, Vincent Lepetit, Naresh Rajkumar, and Kurt Konolige. Going further with point pair features. In *ECCV*, pages 834–848, 2016.

[16] Tomáš Hodaň, Frank Michel, Eric Brachmann, Wadim Kehl, Anders Glent Buch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiří Matas, and Carsten Rother. BOP: Benchmark for 6D object pose estimation. *ECCV*, 2018.

[17] Tomáš Hodaň, Vibhav Vineet, Ran Gal, Emanuel Shalev, Jon Hanzelka, Treb Connell, Pedro Urbina, Sudipta Sinha, and Brian Guenter. Photorealistic image synthesis for object instance detection. *ICIP*, 2019.

[18] E. Kim and G. Medioni. 3d object recognition in range images using visibility context. In *IROS*, pages 3800–3807, 2011.

[19] Sung-Kyun Kim and Maxim Likhachev. Planning for grasp selection of partially occluded objects. In *ICRA*, 2016.

[20] Y. Labbe, J. Carpentier, M. Aubry, and J. Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[21] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *ECCV*, 2018.

[22] Joseph J. Lim, Aditya Khosla, and Antonio Torralba. Fpm: Fine pose parts-based model with 3d cad models. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 478–493, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10599-4.

[23] David G Lowe. Object recognition from local scale-invariant features. In *ICCV*, volume 2, pages 1150–1157. Ieee, 1999.

[24] Lucas Manuelli, Wei Gao, Peter Florence, and Russ Tedrake. kpam: Keypoint affordances for category-level robotic manipulation. *arXiv preprint arXiv:1903.06684*, 2019.

[25] E. Muñoz, Y. Konishi, C. Beltran, V. Murino, and A. Del Bue. Fast 6d pose from a single rgb image using cascaded forests templates. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4062–4069, 2016. doi: 10.1109/IROS.2016.7759598.

[26] E. Muñoz, Y. Konishi, V. Murino, and A. Del Bue. Fast 6d pose estimation for texture-less objects from a single rgb image. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5623–5630, 2016. doi: 10.1109/ICRA.2016.7487781.

[27] MVTec Software GmbH. Halcon. URL https://www.mvtec.com/products/halcon/documentation/release-notes-1911-0/.

[28] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Di-

eter Fox. Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In *CVPR*, 2020.

[29] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019.

[30] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, July 2017.

[31] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, pages 5099–5108, 2017.

[32] Martin Sundermeyer, Maximilian Durner, En Yen Puang, Zoltan-Csaba Marton, Narunas Vaskevicius, Kai O. Arras, and Rudolph Triebel. Multi-path learning for object pose estimation across domains. In *CVPR*, June 2020.

[33] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, and Rudolph Triebel. Augmented autoencoders: Implicit 3d orientation learning for 6d object detection. *IJCV*, 128(3):714–729, 2020.

[34] Garrett Thomas, Melissa Chien, Aviv Tamar, Juan Aparicio Ojea, and Pieter Abbeel. Learning robotic assembly from cad. In *ICRA*, 2018.

[35] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. *arXiv preprint arXiv:1809.10790*, 2018.

[36] M. Ulrich, C. Wiedemann, and C. Steger. Combining scale-space and similarity-based aspect graphs for fast 3d object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1902–1914, 2012. doi: 10.1109/TPAMI.2011.266.

[37] Joel Vidal, Chyi-Yeu Lin, Xavier Lladó, and Robert Martí. A method for 6d pose estimation of free-form rigid objects using point pair features on range data. *Sensors*, 18(8):2678, 2018.

[38] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *CVPR*, pages 3343–3352, 2019.

[39] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *CVPR*, pages 2642–2651, 2019.

[40] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *RSS*, 2018.

[41] Yang Xiao, Xuchong Qiu, Pierre-Alain Langlois, Mathieu Aubry, and Renaud Marlet. Pose from shape: Deep pose estimation for arbitrary 3d objects. *arXiv preprint arXiv:1906.05105*, 2019.

[42] Christopher Xie, Yu Xiang, Arsalan Mousavian, and Dieter Fox. The best of both modes: Separately leveraging rgb and depth for unseen object instance segmentation. In *CoRL*, pages 1369–1378, 2020.

[43] H. Yu, Q. Fu, Z. Yang, L. Tan, W. Sun, and M. Sun. Robust robot pose estimation for challenging scenes with an rgb-d camera. *IEEE Sensors Journal*, 19(6):2217–2229, 2019. doi: 10.1109/JSEN.2018.2884321.

[44] Matt Zucker, Nathan Ratliff, Anca D Dragan, Mihail Pivtoraiko, Matthew Klingensmith, Christopher M Dellin, J Andrew Bagnell, and Siddhartha S Srinivasa. Chomp: Covariant hamiltonian optimization for motion planning. *IJRR*, 2013.