UDA-COPE: Unsupervised Domain Adaptation for Category-level Object Pose Estimation [⋆]

Taeyeop Lee, Byeong-Uk Lee, Inkyu Shin, Jaesung Choe,
Ukcheol Shin, In So Kweon, and Kuk-Jin Yoon

KAIST

Abstract. Learning to estimate object pose often requires ground-truth (GT) labels, such as CAD model and absolute-scale object pose, which is expensive and laborious to obtain in the real world. To tackle this problem, we propose an unsupervised domain adaptation (UDA) for category-level object pose estimation, called **UDA-COPE**. Inspired by recent multi-modal UDA techniques, the proposed method exploits a teacher-student self-supervised learning scheme to train a pose estimation network without using target domain pose labels. We also introduce a bidirectional filtering method between the predicted normalized object coordinate space (NOCS) map and observed point cloud to not only make our teacher network more robust to the target domain but also to provide more reliable pseudo labels for the student network training. Our results demonstrate the effectiveness of our proposed method both quantitatively and qualitatively. Notably, without leveraging target-domain GT labels, our proposed method achieved comparable or sometimes superior performance to existing methods that depend on the GT labels.

Keywords: Object Pose Estimation · Unsupervised Domain Adaptation · Augmented Reality (AR) · Virtual Reality (VR) · Robotics

1 Introduction

Object pose estimation is one of the crucial tasks used in various robotics and computer vision applications for robot manipulation [33, 31, 28, 7] and augmented reality (AR) [23, 19, 20]. Using sensor data such as images or point clouds, this task aims to estimate the poses of target objects including 3D orientation, 3D location, and size information.

Previous 6D object pose estimation methods follow the instance-level pose estimation schemes [26, 32, 21, 22, 28, 11, 10] that rely on given 3D CAD model information (*e.g.*, keypoints, geometry) and the size of known objects. However, these methods typically have difficulty estimating the pose of unknown objects since they do not yet have 3D CAD models as priors. In contrast to the instance-level scheme, category-level object pose estimation [29, 25, 4, 17, 30, 5] approaches are more efficient in that a single network can infer multiple classes at once.

[⋆] This paper is the short version of CVPR'22 official publication.

In particular, Wang *et al.* [29] introduced a pioneering representation called Normalized Object Coordinate Space (NOCS), to align different object instances within one category in a shared 3D orientation. By estimating per-category NOCS maps, it is able to estimate the 6D pose of unseen objects without prior 3D CAD models. Its strengths have led to the use of NOCS representation in the following studies [25, 4, 17, 30, 5].

However, current object pose estimation research mostly relies on supervised learning, which requires expensive GT labels such as 3D object CAD models and absolute object pose. These labels are not only difficult to obtain in the real world but are also unreliable due to the human-annotation. Because of this difficulty, most of the training depends on synthetic datasets [24, 13, 26] and is usually not feasible in real-world applications due to domain gaps.

To cope with the real-world data scarcity problem, we take a look at unsupervised domain adaptation (UDA) methods [12, 16, 36]. UDA approaches often consider two types of datasets, the source domain (*i.e.* synthetic dataset) and the target domain (*i.e.* real-world dataset) dataset. The main goal of the UDA methods is to successfully make deep learning networks robust to the target domain using only the GT labels of the source domain. Various techniques exist, such as pseudo label generation [12, 16], teacher and student networks with momentum updates [1, 34], adversarial learning [3, 14, 2], and etc.

In this paper, we propose an Unsupervised Domain Adaptation for Category-level Object Pose Estimation (UDA-COPE). The proposed method effectively transfers task knowledge from a synthetic domain to a real domain by exploiting a multi-modal self-supervised learning scheme using pseudo labels. Our UDA-COPE concentrates on how to make high-quality pseudo-labels that are efficiently targeted for the category-level pose estimation task. To this end, we designed *bidirectional point filtering* to remove noisy and inaccurate points based on pose optimization. Moreover, our framework achieved better performance than the previous supervised methods [29, 25, 4, 30].

2 Proposed Method

Given an RGB image I, point cloud P, and segmentation labels S, our architecture aims to regress the 6D pose and size $s{\in}\mathbb{R}^3$ of objects. The 6D pose is defined as the rigid transformation of $[R|t]$: rotation $R{\in}SO(3)$, and translation $t{\in}\mathbb{R}^3$. Following previous studies [25, 4, 30, 5, 17], the segmentation labels S are used to crop the RGB images and point clouds. We leverage the NOCS representation to align different object instances within one category in a shared orientation. The categorical object pose $[R|t]$ and size s are estimated by Umeyama algorithm [27] with RANSAC [8], which optimizes $[R|t]$ and s by minimizing the distances between point cloud P and an estimated NOCS map N.

We first illustrate our network architecture (Sec. 2.1). Then, we introduce training methods of supervised learning using synthetic dataset (Sec. 2.2) and unsupervised domain adaption using real-world dataset (Sec. 2.3).

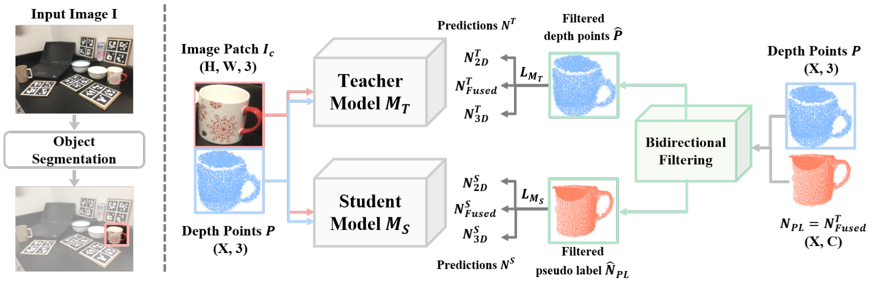


Fig. 1. Overview of unsupervised domain adaptation for category-level object pose estimation (UDA-COPE). UDA-COPE utilizes pseudo label-based teacher/student training scheme. Our proposed bidirectional point filtering method removes the noisy pseudo labels and gives reliable guidance to the student network. At the same time, filtered depth points gives additional self-supervision to the teacher network so that it can be robust to the domain gap between the synthetic and real dataset.

2.1 Network Architecture

Recent category-level object pose estimation methods [25, 4, 17] take an RGB-D input to extract the 2D/3D features. We designed separate 2D/3D branches to extract features from both modalities. We use PSPNet [35] with ResNet34 [9] for 2D feature extraction and the Mink16UNet34 [6] for 3D feature extraction. At this time, the 2D feature is extracted by sampling features that are validly matched with point cloud P from the feature volume. Finally, we have a fused branch that combines each feature from both branches. Every branch estimates a NOCS map (N) with a separate NOCS header, which consists of three multilayer perceptrons (MLP) layers. We designate the NOCS map estimation of each branch as N_{2D}, N_{3D}, N_{Fused}, according to respective feature property.

2.2 Pre-Training with Synthetic Data

Inspired by pseudo label (PL) based methods [12, 16], our method consists of a teacher and a student model. Fig. 1 shows the overview of our teacher and student model. The initial prediction of teacher model M_{T} becomes a pseudo label N_{PL} for a student model, and student model M_{S} learns from the pseudo label as a GT. Our teacher and student model have the same structure as was described in Sec. 2.1.

We first train our teacher model in a supervised manner using the labeled synthetic dataset. For the NOCS map prediction using the GT information, we utilize cross-entropy loss, as in, $H(N_{\text{gt}}, N^{\text{T}})$, where the supervision is given to all predictions from three branches. Additionally, to make our teacher network more robust, we apply the 2D image and 3D points augmentation and use consistency loss L_{C} so that each modality can output consistent results. Total loss for the teacher network on the synthetic dataset is formulated as follows:

$$\begin{aligned} L_{M_{\mathrm{T}}} &= \lambda_{\mathrm{N}} H(N_{\mathrm{gt}}, N^{\mathrm{T}}) + \lambda_{\mathrm{C}} L_{\mathrm{C}}, \\ L_{\mathrm{C}} &= H(N^{\mathrm{T}}, N^{\mathrm{T}}_{\mathrm{Aug}}) \end{aligned} \tag{1}$$

where $N^{\mathrm{T}}_{\mathrm{Aug}}$ is the NOCS map prediction from the augmented input, and λ_{N} and λ_{C} are weighting parameters. Notation for the modality of the predictions is discarded for better readability.

2.3 Pose-Aware Unsupervised Domain Adaptation

After training from the synthetic dataset, the most straightforward yet naive approach is to train the student network using the prediction of the teacher network. However, using the initial prediction from the teacher model as a pseudo-label can be risky. The risk is due to the lack of robustness of the teacher model itself, or more importantly, because of the insufficient knowledge that the teacher model holds with respect to real-world scenarios, due to the domain gap between the simulated and real worlds. Techniques such as data augmentation and momentum updates might help the feasibility but are still restricted. Therefore, we need additional guidance for the teacher model to estimate high-quality predictions, and more reliable pseudo labels for our student model to learn from.

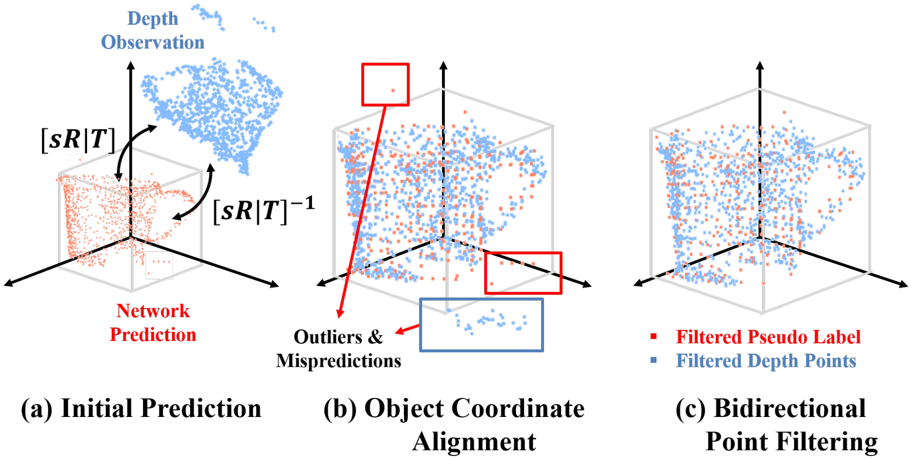


Fig. 2. Overview of bidirectional point filtering method. Given pseudo labels and depth points (a), we estimate the pose and size using the Umeyama [27] algorithm and RANSAC [8], and align the depth points to normalized object coordinate (b). The pseudo label (red) and aligned depth points (blue) have noisy and inaccurate points. After our bidirectional point filtering, the noisy points are removed to give more reliable supervision for both teacher and student (c).

Bidirectional Point Filtering To solve these problems, we propose the bidirectional point filtering method which simultaneously removes the noise of the pseudo labels for the student and filters noisy depth points P for a teacher network. Fig. 2 shows an overview of the proposed bidirectional filtering method. Our bidirectional filtering method uses the P and N_{PL} as input and initially estimates the pose $[R|t]$ and size s using the Umeyama algorithm [27] with RANSAC [8]. Then it aligns the depth points P to the NOCS coordinate by applying the inverse of the estimated pose, as in multiplying the matrix $[sR|t]^{-1}$. We denote aligned depth points as P'. And then we calculate the point-wise 3D distance d between the aligned depth points P' and pseudo label N_{PL} to filter out noisy points from both sides using ρ as the threshold. Finally, we get the refined pseudo label $\hat{N}_{\mathrm{PL}}$ and filtered aligned depth $\hat{P}$. Our bidirectional point filtering can be expressed as:

$$\begin{aligned} d(n) &= \|P'(n) - N_{\mathrm{PL}}(n)\| \quad \text{where} \quad \forall n \in [1|P'], \\ \hat{N}_{\mathrm{PL}} &= \{N_{\mathrm{PL}}(n) : d(n) < \rho\}, \\ \hat{P} &= \{P'(n) \quad : d(n) < \rho\}, \end{aligned} \tag{2}$$

Fig. 2 shows that our bidirectional filtering method removes outliers of pseudo label N_{PL} and depth P, and results in refined pseudo label $\hat{N}_{\mathrm{PL}}$ and filtered depth points $\hat{P}$.

Self-Supervised Learning After the bidirectional filtering, we jointly train the teacher network and student network using the filtered pseudo labels $\hat{N}_{\mathrm{PL}}$ and filtered aligned depth points $\hat{P}$. Noted that we only use the filtered points $\hat{P}$ for the teacher training, which may be a smaller subset of an original P. We use cross-entropy loss to train a student model using clean pseudo labels $\hat{N}_{\mathrm{PL}}$. The student model loss is defined as:

$$L_{M_{\mathrm{S}}} = -\frac{1}{|\hat{N_{\mathrm{PL}}}|} \sum_{n=1}^{|\hat{N}_{\mathrm{PL}}|} H(\hat{N}_{\mathrm{PL}}(n), N^{\mathrm{S}}(n)), \tag{3}$$

where N^S is the predictions of our student network. At the same time, the teacher learns real data knowledge from observation. We use cross-entropy loss by utilizing geometric consistency between our filtered aligned depth $\hat{P}$ and estimated pseudo labels N^T. The teacher model loss is defined as:

$$L_{M_{\mathrm{T}}} = -\frac{1}{|\hat{P}|} \sum_{n=1}^{|\hat{P}|} H(\hat{P}(n), N^{\mathrm{T}}(n)). \tag{4}$$

We train our teacher model with a small learning rate for stable teacher network training. For both teacher and student models, we compute the loss for all estimations, N_{2D}, N_{3D}, N_{Fused}, which shows better result than applying the loss to only N_{Fused}. We denote all estimation losses as all modality (AM) loss.

3 Experiments

3.1 Comparison with State-of-the-art

We compared our methods with state-of-the-art methods that were trained on different datasets and labels: 1) labeled synthetic dataset, 2) labeled synthetic and real datasets, 3) labeled synthetic and unlabeled real datasets. All methods were evaluated on the REAL275 dataset. Note that only the approaches with the ability to perform multi-class category-level pose estimation using a single network were considered. RGB, Depth and RGB-D denotes the modality of the network input, and most of the RGB based approaches utilize depth information in the pose optimization or refinement process.

Method	Input	Syn	Real	Real	mAP (↑)					
			w/ Label	w/o Label	$3D_{50}$	$3D_{75}$	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$
CPS++ [18]	RGB	✓			**72.6**	-	-	**25.8**	-	-
Metric Scale [15]	RGB	✓			68.1	32.9	2.2	5.3	10.0	24.7
NOCS [29]	RGB	✓			36.7	3.4	-	3.4	-	20.4
SPD [25]	RGB-D	✓			71.0	**43.1**	**11.4**	12.0	**33.5**	**37.8**
NOCS [29]	RGB	✓	✓		78.0	30.1	7.2	10.0	13.8	25.2
SPD [25]	RGB	✓	✓		75.2	46.5	15.7	18.8	33.7	47.4
SPD [25]	RGB-D	✓	✓		77.4	53.5	19.5	21.6	43.5	54.0
CASS [4]	RGB-D	✓	✓		77.7	-	-	23.5	-	58.0
CR-Net [30]	RGB-D	✓	✓		79.3	55.9	27.8	34.3	47.2	60.8
DualPoseNet [17]	RGB-D	✓	✓		79.8	**62.2**	29.3	35.9	50.0	66.8
SGPA [5]	RGB-D	✓	✓		**80.1**	61.9	**35.9**	**39.6**	**61.3**	**70.7**
CPS++ [18]	RGB	✓		✓	72.8	-	-	25.2	-	-
Ours	RGB	✓		✓	82.0	59.0	24.4	27.0	49.3	54.8
Ours	D	✓		✓	79.6	57.8	21.2	29.1	48.7	65.9
Ours	RGB-D	✓		✓	**82.6**	**62.5**	**30.4**	**34.8**	**56.9**	**66.0**

Table 1. Quantitative comparison with state-of-the art methods on the REAL275 dataset. Empty entries either could not be evaluated or were not reported in the original paper.

Method	Syn	Real	mAP (↑)							
		w/o Label	$3D_{25}$	$3D_{50}$	$3D_{75}$	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$	$10°10cm$
CPS++ (RGB)	✓		84.5	72.6	-	-	25.8	-	-	55.4
CPS++ (RGB)	✓	✓	84.6 (+0.1)	72.8 (+0.2)	-	-	25.2 (-0.6)	-	-	58.6 (+3.2)
Ours (RGB)	✓		83.3	79.9	49.7	15.4	18.3	37.6	46.7	48.9
Ours (RGB)	✓	✓	83.8 (+0.5)	82.0 (+2.1)	59.0 (+9.3)	24.4 (+9.0)	27.0 (+8.7)	49.3 (+11.7)	54.8 (+8.1)	56.9 (+8.0)
Ours (RGB-D)	✓	✓	84.0 (+0.7)	82.6 (+2.7)	62.5 (+12.8)	30.4 (+15.0)	34.8 (+16.5)	56.9 (+19.3)	66.0 (+19.3)	68.3 (+19.4)

Table 2. Quantitative comparison of unsupervised pose estimation approaches on the REAL275 dataset. Empty entries are either not able to be evaluated or not reported in the original paper. Performance margins are calculated compared to the synthetic-only results.

Supervised Pose Estimation methods. Table 1 summarizes the results of the state-of-the-art category-level object pose estimation methods. Obviously, supervised training with the real data annotation significantly improved the overall performance, as are revealed by comparing the results of NOCS [29] and SPD [25] on different training dataset conditions. However, our unsupervised method showed results superior to NOCS [29], SPD [25], CASS [4], and CR-Net [30]. Compared to two of the strongest previous approaches, SGPA [5] and DualPoseNet [17], ours still showed comparable performance. This indicates that our proposed filtered pseudo label based UDA-COPE is robust when estimating object pose in unseen real-world instances.

Unsupervised Pose Estimation methods. Table 2 summarizes the results of CPS++ and our method on source only, and source with unlabeled target training conditions. CPS++ [18] provides self-supervision by computing the consistency between the observed depth map and the rendered depth. The rendered depth is obtained by projecting an estimated 3D shape with the predicted pose. The results from row 1 and row 2 in Table 2 show that for CPS++, using unlabeled real data marginally improved performance, and sometimes even worsened it, as in $5°, 5$cm metric. We believe that their self-supervision is unreliable because of ambiguous 3D shape reconstruction using only a single-view RGB image.

Comparing row 3 and row 4, it can be seen that our proposed method shows improved results for every metrics, with some metrics showing notable margins such as an 8.7 mAP (48%) increase in $5°, 5$cm. Also, in the last row, our RGB-D result had better performance than the single modality based outputs. Therefore, we claim that our proposed algorithm is more effective by utilizing a pseudo label based learning scheme with modality and pose-aware self-supervision. The effectiveness of each components will be ablated thoroughly in the following sections.

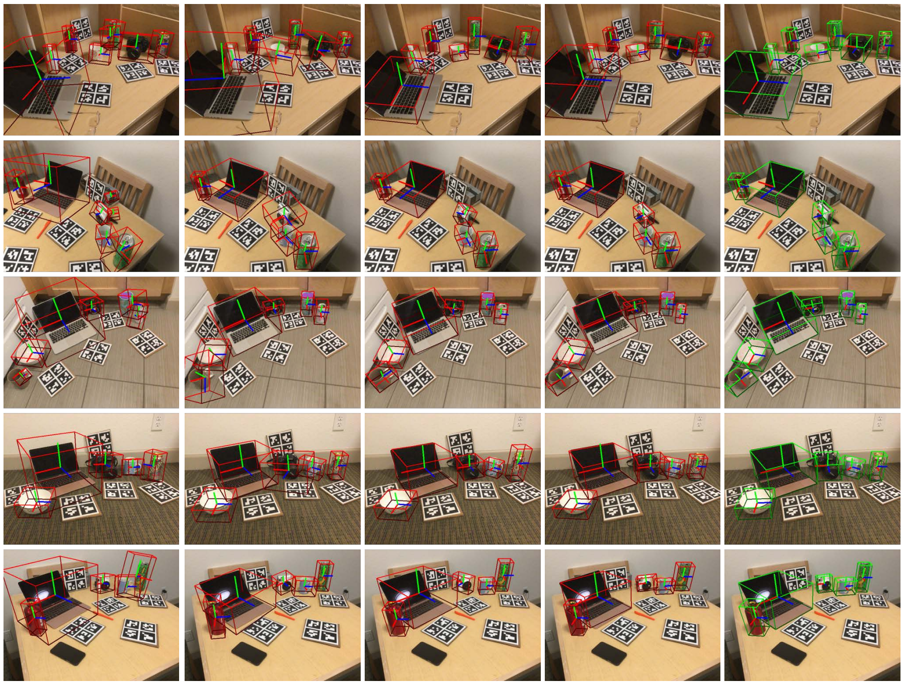
3.2 Qualitative Results

Fig. 3 shows qualitative results on the REAL275 dataset. We compare our results with some of the supervised methods, NOCS [29], SPD [25] and DualPoseNet [17]. Our method estimated pose and sizes more accurately than NOCS and SPD, especially on cameras and laptops. Compared to the state-of-the-art approach, DualPoseNet, ours exhibited comparable predictions, although it was not trained with the GT labels of the real dataset.

Fig. 4 visualizes some examples of the real training set with GT labels and our pseudo labels. 6D poses were obtained and visualized using the Umeyama algorithm [27], using GT NOCS map and our pseudo label NOCS map.

4 Conclusions

We propose UDA-COPE, unsupervised domain adaptation for category-level object pose estimation which addresses the real-world lack-of-label problem using multi-modality (RGB-D). Specifically, we designed a bidirectional point filtering method to filter noisy pseudo labels, and observed depth points, where the filtered depth points improve the robustness of the teacher network, and the filtered pseudo label helps efficient student network training. Both provide for better domain adaptation with real-world pose estimation. Experiments showed that our proposed pipeline and pose-aware point filtering results were comparable to or sometimes better than the performance of fully supervised approaches.



(a) NOCS [29] (b) SPD [25] (c) Dual [17] (d) Ours (e) GT

Fig. 3. Qualitative comparison on the REAL275 dataset. Compared to two of the strongest supervised approaches, SGPA [5] and DualPoseNet [17], ours still showed comparable performance even in unsupervised settings. This indicates that our proposed filtered pseudo label-based UDA-COPE is robust when estimating object pose in unseen real-world instances.

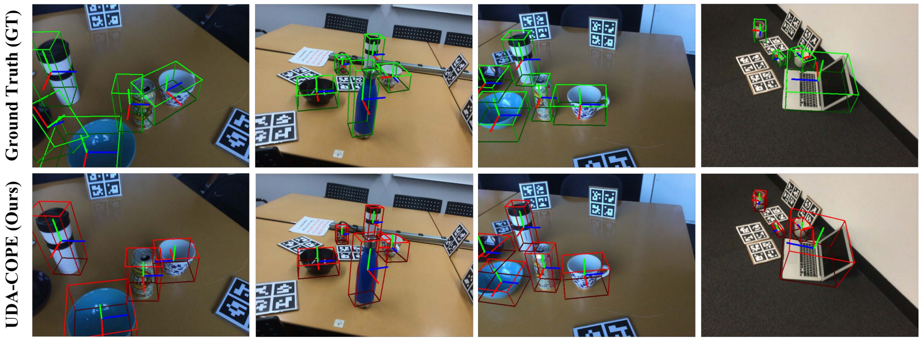


Fig. 4. Noisy GT label examples of the Real training dataset. Human-annotated GT pose labels on the real dataset (top row) are sometimes more inaccurate than our predicted pseudo labels (bottom row). The real data annotations were mainly performed automatically using aruco markers. For some of the failure cases, additional ICP or manual human annotations were needed. Therefore, frames with inaccurate labels exist, which might disrupt supervised training.

Remark

This paper is a re-publishing (summary presentation) of the paper, which has been published in CVPR 2022 by request of the IW-FCV2023 program committee to share the research results.

References

1. Araslanov, N., Roth, S.: Self-supervised augmentation consistency for adapting semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15384–15394 (2021)
2. Bousmalis, K., Irpan, A., Wohlhart, P., Bai, Y., Kelcey, M., Kalakrishnan, M., Downs, L., Ibarz, J., Pastor, P., Konolige, K., et al.: Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 4243–4250 (2018)
3. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3722–3731 (2017)
4. Chen, D., Li, J., Wang, Z., Xu, K.: Learning canonical shape space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11973–11982 (2020)
5. Chen, K., Dou, Q.: Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2773–2782 (2021)
6. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3075–3084 (2019)
7. Du, G., Wang, K., Lian, S., Zhao, K.: Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. Artificial Intelligence Review **54**(3), 1677–1734 (2021)
8. Fischler, M.A., Bolles, R.C.: 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)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
10. He, Y., Huang, H., Fan, H., Chen, Q., Sun, J.: Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3003–3013 (2021)
11. He, Y., Sun, W., Huang, H., Liu, J., Fan, H., Sun, J.: 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 (CVPR). pp. 11632–11641 (2020)
12. Jaritz, M., Vu, T.H., Charette, R.d., Wirbel, E., Pérez, P.: xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12605–12614 (2020)

13. Kehl, W., Manhardt, F., Tombari, F., Ilic, S., Navab, N.: Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1521–1529 (2017)
14. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 35–51 (2018)
15. Lee, T., Lee, B.U., Kim, M., Kweon, I.S.: Category-level metric scale object shape and pose estimation. IEEE Robotics and Automation Letters (RA-L) **6**(4), 8575–8582 (2021)
16. Li, Y., Yuan, L., Vasconcelos, N.: Bidirectional learning for domain adaptation of semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6936–6945 (2019)
17. Lin, J., Wei, Z., Li, Z., Xu, S., Jia, K., Li, Y.: Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3560–3569 (2021)
18. Manhardt, F., Wang, G., Busam, B., Nickel, M., Meier, S., Minciullo, L., Ji, X., Navab, N.: Cps++: Improving class-level 6d pose and shape estimation from monocular images with self-supervised learning. arXiv preprint arXiv:2003.05848 (2020)
19. Marchand, E., Uchiyama, H., Spindler, F.: Pose estimation for augmented reality: a hands-on survey. IEEE Transactions on Visualization and Computer Graphics (TVCG) **22**(12), 2633–2651 (2015)
20. Marder-Eppstein, E.: Project tango. In: ACM SIGGRAPH 2016 Real-Time Live!, pp. 25–25 (2016)
21. Park, K., Patten, T., Vincze, M.: Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 7668–7677 (2019)
22. Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H.: Pvnet: Pixel-wise voting network for 6dof pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4561–4570 (2019)
23. Runz, M., Buffier, M., Agapito, L.: Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In: IEEE International Symposium on Mixed and Augmented Reality (ISMAR). pp. 10–20 (2018)
24. Sundermeyer, M., Marton, Z.C., Durner, M., Brucker, M., Triebel, R.: Implicit 3d orientation learning for 6d object detection from rgb images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 699–715 (2018)
25. Tian, M., Ang Jr, M.H., Lee, G.H.: Shape prior deformation for categorical 6d object pose and size estimation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
26. Tremblay, J., To, T., Sundaralingam, B., Xiang, Y., Fox, D., Birchfield, S.: Deep object pose estimation for semantic robotic grasping of household objects. In: Conference on Robot Learning (CoRL) (2018), https://arxiv.org/abs/1809.10790
27. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on Pattern Analysis & Machine Intelligence (TPAMI) **13**(04), 376–380 (1991)
28. Wang, C., Xu, D., Zhu, Y., Martín-Martín, R., Lu, C., Fei-Fei, L., Savarese, S.: Densefusion: 6d object pose estimation by iterative dense fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3343–3352 (2019)

29. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2642–2651 (2019)
30. Wang, J., Chen, K., Dou, Q.: Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2021)
31. Wong, J.M., Kee, V., Le, T., Wagner, S., Mariottini, G.L., Schneider, A., Hamilton, L., Chipalkatty, R., Hebert, M., Johnson, D.M., et al.: Segicp: Integrated deep semantic segmentation and pose estimation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5784–5789 (2017)
32. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS) (2018)
33. Zeng, A., Yu, K.T., Song, S., Suo, D., Walker, E., Rodriguez, A., Xiao, J.: Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 1386–1383 (2017)
34. Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., Wen, F.: Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12414–12424 (2021)
35. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2881–2890 (2017)
36. Zou, Y., Yu, Z., Liu, X., Kumar, B., Wang, J.: Confidence regularized self-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5982–5991 (2019)