Usability Study of Learning-Based Pose Estimation of Industrial Objects from Synthetic Depth Data

Stefan Thalhammer, Timothy Patten, Markus Vincze

Usability Study of Learning-based Pose Estimation of Industrial Objects from Synthetic Depth Data

Stefan Thalhammer[0000-0002-0008-430X], Timothy Patten[0000-0003-1139-9451], and Markus Vincze[0000-0002-2799-491X]

Faculty of Electrical Engineering and Information Technology, TU Wien, 1040 Vienna, Austria
{thalhammer, patten, vincze}@acin.tuwien.ac.at

Abstract. For visual assistance systems deployed in an industrial setting, precise object pose estimation is an important task in order to support scene understanding and to enable subsequent grasping and manipulation. Industrial environments are especially challenging since mesh-models are usually available while physical objects are not or are expensive to model. Manufactured objects are often similar in appearance, have limited to no textural cues and exhibit symmetries. Thus, these are especially challenging for recognizers that are meant to provide detection, classification and pose estimation on instance level. A usability study of a recent synthetically trained learning-based recognizer for these particular challenges is conducted. Experiments are performed on the challenging T-LESS dataset due to its relevance for industry.

Keywords: pose estimation, deep learning, synthetic data, rendering, augmentation, depth data

1 Introduction

Typical tasks for visual assistance systems designed for industrial use cases are object detection and pose estimation. In order to minimize costs, training recognizers from synthetic data is desired since the data generation process can easily be automated. To automate these processes, objects models in the form of meshes are required.

While classical approaches are designed to strongly rely on the availability of mesh models of the objects of interest to train their recognizers, learning-based approaches require huge amounts of annotated images [2]. Object models are usually given since they are available in the company's resource planning system. If not, they can be easily acquired through reconstructing the physical object instances. This is a strong advantage over task-specific learning-based approaches that require real-world training images.

While traversing from synthetic training data to real-world test data without a decrease in performance is easy for classical approaches, learning-based

approaches do not traverse that easily. In order to circumvent this so called reality gap, methods need to be applied for the domain transition such as Generative Adversarial Networks (GAN), feature space mapping or randomized augmentations [3–8].

The authors of SyDPose [7] showed that using a randomized camera model combined with randomized augmentations on depth data leads to suitable recognizers for texture-less objects with considerable diversity in shape [9, 10]. However, the datasets used in the investigation are not designed to evaluate the challenges relevant in industry. Typical challenges of industrial use cases include little inter-class variations, little to no texture and symmetries. Recently the authors of [11] introduced the T-LESS dataset, which is designed to mimic the specifically challenging industrial object recognition setting. SyDPose is applied on the T-LESS dataset to evaluate the usability of depth data for learning-based object pose estimation using only synthetic data.

Handling symmetries is a general problem in pose estimation. All 30 objects in the T-LESS dataset exhibit either an axis of rotational symmetry or one to multiple planes of symmetry. Symmetry handling with learning-based approaches is usually performed during training data annotation, during training time or during the evaluation itself [1, 10, 12–14]. The first two approaches are often necessary to allow learning-based approaches to converge. Problems can occur when two or more distinct views visually appear similar but have different annotation, therefore, hindering the learning process and consequently slowing the convergence rate. Approaches that propose multiple hypotheses are agnostic to the problem, thus symmetry handling can be entrusted to the evaluation. In this work, symmetry handling during training data annotation is formalized. Although the training data has been generated without symmetry awareness, the symmetries of objects are known within the local reference frames and this information is exploited.

Our contributions are the following:

- We substantiate the usability of learning-based pose estimators trained using only synthetic depth data for challenging, industrially relevant, texture-less objects.
- We provide guidelines to train Convolutional Neural Networks (CNN) for pose estimation of objects with rotational symmetries or one plane of symmetry.

The remainder of the paper is structured as follows. Section 2 summarizes related work. Section 3 presents the approach. Section 4 gives insights regarding training and presents the results. Section 5 concludes with a discussion.

2 Related Work

Estimating object poses from single-shot images using learning-based approaches gained considerable attention recently [4–7, 12–18]. Many of these methods require annotated real-world data for training directly, or must learn the mapping

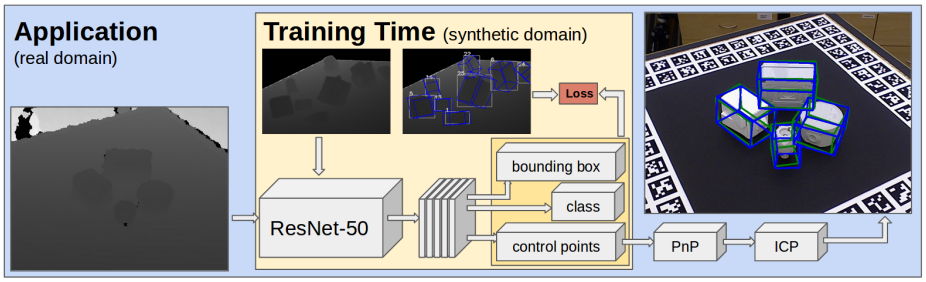

Fig. 1: The network is trained end-to-end in a multi-task fashion. For training, only augmented synthetic depth data is used. During deployment, estimations are made on real-world data with additional PnP for control point projection and ICP for refinement. RGB is used for visualization.

from synthetic to real world [8, 13, 15–18]. Some methods train using only synthetic data, either from RGB [4, 6] or depth data [7, 16].

Methods that are trained using only synthetic data are highly desired by industry because data acquisition and annotation is leveraged through processes that can easily be automated. While the reality gap is quite considerable for the RGB domain it is smaller when using depth data. Consequently, SyDPose is designed for depth data, trained from rendered and augmented synthetic data [7].

A promising stream of object pose estimation research focuses on the prediction of correspondences between object models and a given input image. These methods can be categorized by dense [8, 12, 17] or sparse [7, 13, 15, 18]. Common to all these methods is that they use the predictions to estimate the object poses by re-projecting to 3D. This is done using variants of the Perspective-n-Points algorithm (PnP).

Handling symmetries of objects of interest is especially challenging for learning-based approaches. This results from the necessity to incorporate symmetry handling already in the training process in order to guarantee successful convergence. Two main streams of symmetry handling are popular. Firstly, object poses are annotated such that ambiguous views with different poses are avoided [6, 13]. Secondly, symmetry handling is taken into account while calculating the loss during training time [12, 14]. For the task at hand, it is assumed that only mesh models are given. Thus, rules have to be formalized to disambiguate symmetric views of objects with rotational or discrete symmetries.

3 Approach

This section describes the approach for object detection and pose estimation using only synthetic depth data for training. An overview is given in Figure 1.

Synthetic depth images are rendered from a virtual scene. These images are augmented using a randomized camera model and used as training data. Object poses are presented to the network in the form of discrete feature locations projected to 2D control points. The network learns to regress potential object

locations in image space, the corresponding class and control points. The outputs are used to estimate the poses of the objects of interest in 3D space via re-projection using PnP. Additional pose refinement is performed using the Iterative Closest Point (ICP) algorithm [19]. RetinaNet[1] [20] is used with ResNet-50 [21] backbone, pretrained on ImageNet [22] as feature extractor and detector. An additional network head that is parallel to the classification and detection branches regresses 3D control point locations.

In the remaining of this section, the network architecture is described in more detail and the method for handling object symmetries is elaborated.

3.1 Network Architecture

RetinaNet consists of a backbone network that slides over the image computing convolutional feature maps over multiple scales. These feature maps are passed to subsequent networks for classification and bounding box regression. Translation invariant anchors are constructed using the parameters described in [20] to construct spatial positions. This procedure is coherent with sliding window approaches. For each of the anchor positions, the backbone network calculates features maps. The feature activation outputs of the third, fourth and fifth residual block of the ResNet-50 backbone are used to construct rich, multi-scale features by the Feature Pyramid Network [23]. These multi-scale features are passed to the classification and bounding box regression branches, which predict class probabilities and the bounding box offset from the corresponding anchor center when an object is present. The input feature map is additionally passed to a control point regression branch.

Control points are regressed in image space as proposed by [24]. Points can be chosen arbitrarily in the object's coordinate frame and must not represent actual meaningful features. Using the camera intrinsics and the calculated corresponding object pose these points are projected into image space.

The control point estimation branch consists of four convolutional layers used for pose-specific feature extraction, with Rectified Linear Units (ReLu) used as activations, and one final convolution layer with linear activation for coordinate prediction. The last convolutional layer predicts 16 values representing the x and y components of the eight defined control points location in image space. Additionally, *l2* regularization of every layer's weights with the hyperparameter set to 0.001 is used in order to help the network learn a more general representation.

3.2 Symmetry Handling

Symmetry handling in the context of deep learning is necessary to not hinder the learning process while training the network. This can be done by only showing symmetrical objects from certain views in order to obtain only unambiguous annotations. A different line of research incorporates the object's symmetry into

[1] https://github.com/fizyr/keras-retinanet

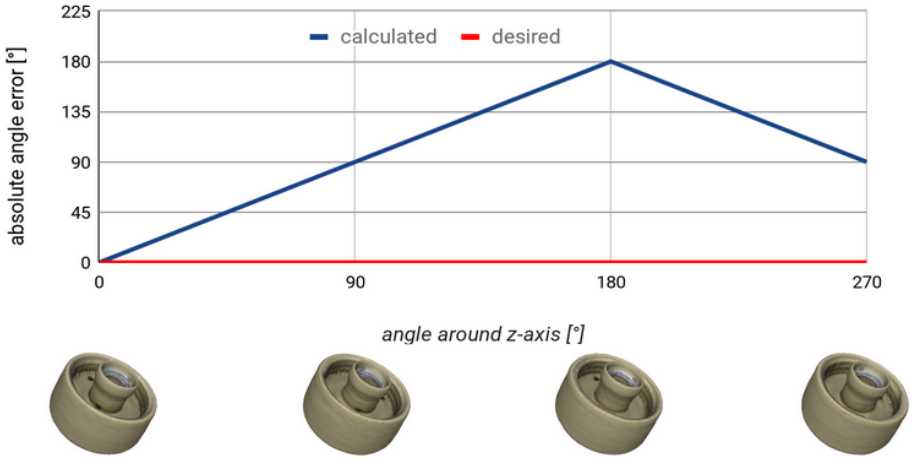

Fig. 2: Calculated versus desired angle error for loss calculation of object 17, of T-LESS, revolved around its axis of symmetry.

the loss calculation [13]. Disambiguating the loss calculation by calculating symmetry aware losses prevents the network training from being hindered [12, 14].

The case when object models and annotations regarding symmetries are given is considered.

Objects with continuous symmetries can be rotated arbitrarily around the axis of symmetry resulting in an infinite number of similar, thus ambiguous views. When training these cases without employing special measures to keep the angle error between the predicted and the ground truth pose near zero, the training process is hindered. Figure 2 shows the calculated angle error and the desired angle error for the loss calculation of objects with continuous symmetries. In order to disambiguate these views, the annotated object pose is transformed by rotating the object along its axis of symmetry. The rotation is stopped when the plane spanned by an arbitrary axis orthogonal to the plane of symmetry and the axis of symmetry passes through the camera center. As such, optimizing for the object rotation around its axis of symmetry is omitted. Leading to no hindrance of the network's convergence process.

When handling objects with one plane of symmetry, views that can be mirrored at the plane of symmetry can hinder the training process. Figure 3 compares the calculated and the desired angle error for the loss calculation of objects with a plane of symmetry. The desired absolute angle error vanishes for similar views. In order to disambiguate these cases, object pose annotations are mirrored such that the normal of the plane of symmetry always points to the same side of the image, i.e., to the positive or negative x-direction of the image frame. Consequently, the same control points are consistently either on the right or on the left side of the object during training. This results in a valid solution that does not hinder network convergence during training.

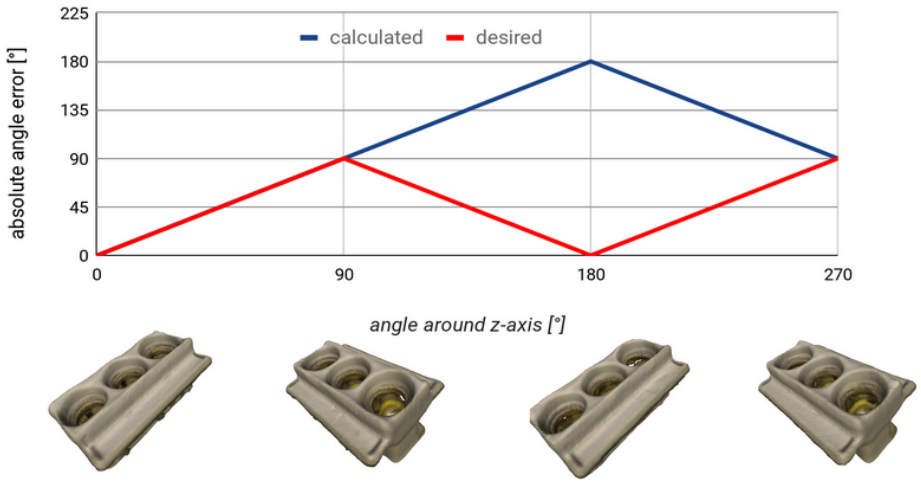

Fig. 3: Calculated versus desired angle error for loss calculation of object 9, of T-LESS, revolved around its z-axis.

The proposed transformations can already be applied during rendering. They can also be applied during training data annotation if at the time the data is rendered without the knowledge of symmetry annotation.

4 Experiments

In order to substantiate the usability of deep architectures for pose estimation of industrially relevant objects, SyDPose is trained only on synthetic data and the performance is evaluated on challenging texture-less objects. As a prerequisite for training, only non-textured meshes of the objects of interest are needed. Evaluation is done on the T-Less dataset. The dataset objects have no discriminative color, little texture, are often similar in shape and some objects are parts of other objects [11].

4.1 Network Training

Synthetic training images are rendered from a scene depicting the variations of real-world images. The generally applicable approach of domain randomization [25] is used for dataset creation and rendering.

Synthetic depth data of a virtual scene resembling the area of deployment is rendered using Blender[2]. These data are subsequently augmented and annotated using a randomized noise model and are used for supervised training as proposed by [7]. Domain randomization, i.e. diverse scene setups and various background information, produces training data with high variation regarding views and

[2] www.blender.org

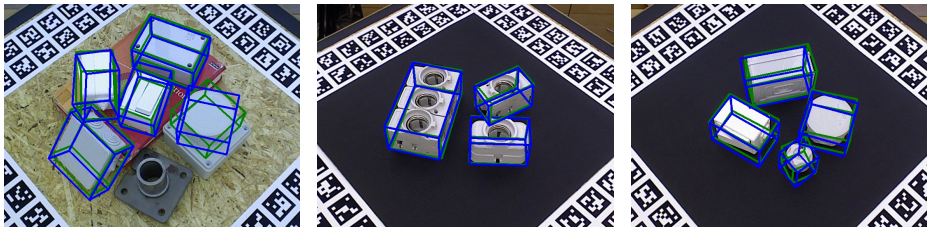

Fig. 4: Qualitative pose estimation results on T-LESS. The blue boxes indicate estimated poses and the green boxes indicate ground truth poses.

occlusion patterns. Additionally, noise heuristics are applied. An augmented domain X^{a} that creates a superset $X^{\mathrm{a}} \supseteq X^{\mathrm{r}}$ of the variations in real-world scans is created. Care is taken that X^{a} does not diverge far from X^{r} by choosing variations that do not violate the sampling theory. This has been shown to generate high-quality data to train deep architectures for object detection, classification and pose estimation in depth images.

Approximately 25,000 training images of virtual scenes exhibiting the expected variations regarding camera views, object poses and occlusion patterns are rendered. During training, annotations of objects with bounding boxes or control points outside of the image space are discarded. Annotations of objects with visibility of less than 50% are discarded as well.

The remaining objects are annotated with a bounding box, a class label and virtual control points representing the pose. Only one model is trained for all dataset objects combined in contrast to the state-of-the-art methods that train individual networks for each object [6, 13, 14, 18]. This is done in order to show industrial applicability by keeping the computational resources and time consumption low. As a consequence, SyDPose performs pose estimation with a frame rate of more than five frames per second. Model training is done for 100 epochs with the Adam optimizer with adaptive learning rate initialized at 10^{-5}.

4.2 Evaluation Metrics

For evaluation, the Visual Surface Discrepancy (VSD) metric proposed by Hodan et al. [1] is used. In order to calculate the score for an estimate, object models are rendered in the ground truth and the estimated pose to calculate distance maps. These are subsequently applied to the scene to obtain visiblity masks. For every visible object pixel the l_1 distance between ground truth and estimated distance map is calculated. If the distance is below a tolerance τ the pixel error is 0 otherwise it is 1. An estimated pose is consider correct if the average distance of all visible pixels is below the threshold θ. $\tau = 20mm$ and $\theta = 0.3$ is used as proposed by [1].

4.3 Results

Exemplary qualitative results for the estimated poses on the T-LESS dataset are provided in Figure 4. Blue boxes represent estimated and green boxes ground

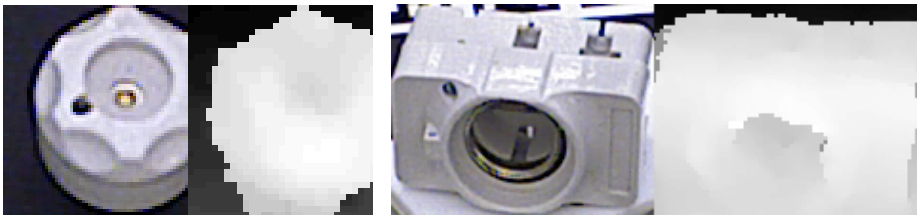

Fig. 5: RGB and depth images of the objects 2 and 6 from the T-LESS dataset, respectively. Many of the small cues are not observable in the depth modality while easily observable in RGB.

truth poses. For objects with rotational symmetries, poses can diverge along the axis of symmetry and still minimize the VSD-score. Such cases are visible in the left and the right image. The middle image shows accurate results for objects that are parts of others.

In order to evaluate the suitability of synthetic depth data for precision assembly tasks, a comparison is done to the state-of-the-art method for learning-based pose estimation from synthetic data [6]. Real-world data is used to train the detector. This approach uses RGB data in contrast to SyDPose that uses depth data. Only synthetic data is used for training.

Table 1 presents a comparison in terms of pose estimation recall using the VSD-metric. Since the detection network (not pose estimation) of Sundermayer et al. [6] is trained on real-world data, the approach exhibits a detection recall that is approximately almost twice as high as SyDPose that achieves only 47.49%. Consequently, when comparing the pose estimation recall from the network output the depth-based results are inferior to the RGB-based method. Since it is safe to assume that depth-based approaches have access to the raw depth input, refining results using ICP is viable. This is not the case for RGB-based approaches. Thus, comparing both methods under the assumption of availability of only one modality comparable results are achieved. Object 27 is excluded from evaluation since it exhibits three planes of symmetry.

One advantage of the depth modality compared to the RGB domain is that RGB is in general more challenging due to the broad variety of variations affecting the appearance of the captures scene. Thus, transitioning from synthetic to real-world data is easier in the depth domain due to the comparably limited influence of environmental conditions. Depth data also gives strong cues about the object's shape and scale. Both of these aspects are advantageous for pose estimation. However, a strong limitation is that the sampling noise is very high in the depth domain. Consequently, when using RGB, strong cues can be extracted for inference that cannot be captured by depth data. Figure 5 presents the crops of two T-LESS objects from the test set captured with the Primesense Carmine 1.09. The depth images are scaled to have 8 bit range. In the RGB images, cues that are relevant for classification and pose disambiguation are easily observable. While in the depth images these are mostly not observable due to

Table 1: Object pose estimation recall on T-LESS, using the VSD-metric.

Object	Sundermeyer et al. [6]	SyDPose	SyDPose + ICP
1	8.87	5.13	15.81
2	13.22	7.00	23.56
3	12.47	12.22	19.44
4	6.56	5.25	14.68
5	34.8	13.65	30.65
6	20.24	5.33	13.29
7	16.21	1.40	8.81
8	19.74	0.26	5.16
9	36.21	3.40	10.28
10	11.55	4.68	18.72
11	6.31	1.19	12.17
12	8.15	4.17	19.51
13	4.91	7.47	9.72
14	4.61	9.02	18.32
15	26.71	6.19	30.56
16	21.73	13.40	31.75
17	64.84	1.29	16.14
18	14.3	3.54	18.72
19	22.46	4.00	10.17
20	5.27	3.20	12.55
21	17.93	3.86	16.87
22	18.63	2.84	15.61
23	18.63	2.00	18.01
24	4.23	6.98	26.19
25	18.76	4.51	34.03
26	12.62	3.40	24.80
28	23.07	1.33	15.34
29	26.65	2.46	31.25
30	29.58	6.99	44.44
Mean	18.25	5.04	19.54

the sampling noise. These limitations, however, can be overcome by using depth imaging devices that induce less noise.

5 Conclusion

This work provides evidence for the usability of synthetic depth for learning-based pose estimation. However, methods working with RGB can potentially access cues that are not available for depth when targeting low cost sensor setups. RGB sensors come with comparably low cost relative to the sensors precision. Current state of the art for learning based pose estimation shows decent performance for the RGB and the depth modality when trained on synthetic data. Consequently, future work will address combining the RGB and the depth

modality in order to exploit the advantages of both domains and improve performance over each modality individually.

Acknowledgements

This work has been supported by the Austrian Research Promotion Agency in the program Production of the Future funded project MMAssist_II (FFG No. 858623), the Austrian Ministry for Transport, Innovation and Technology (bmvit) and the Austrian Science Foundation (FWF) under grant agreement No. I3969-N30 (InDex).

References

1. Hodaň, T., Michel, F., Brachmann, E., Kehl, W., Buch, A., Kraft, D., Drost, B., Vidal, J., Ihrke, S., Zabulis, X., Sahin, C., Manhardt, F., Tombari, F., Kim, T., Matas, J., Rother, C.: BOP: Benchmark for 6D Object Pose Estimation. Proceedings of European Conference on Computer Vision, pp. 19-35, 2018
2. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. pp. 1097-1105, 2012
3. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.; Unsupervised Pixel–Level Domain Adaptation with Generative Adversarial Networks. Proceedings of Conference on Computer Vision and Pattern Recognition. pp. 3722-3731, 2017
4. Kehl, W., Manhardt, F., Tombari, F., Ilic, S., Navab, N.: SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again. Proceedings of IEEE International Conference on Computer Vision. pp. 1521-1529, 2017
5. Rad, M., Oberweger, M., Lepetit, V.: Feature Mapping for Learning Fast and Accurate 3D Pose Inference from Synthetic Images. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2018
6. Sundermeyer, M., Marton, Z., Durner, M., Brucker, M., Triebel, R.: Implicit 3d orientation learning for 6d object detection from rgb images. Proceedings of European Conference on Computer Vision. pp. 699-715, 2018
7. Thalhammer, S., Patten, T., Vincze, M.: SyDPose: Object Detection and Pose Estimation in Cluttered Real-World Depth Images Trained using only Synthetic Data. Proceedings of International Conference on 3D Vision. pp. 106-115, 2019
8. Zakharov, S., Planche, B., Wu, Z., Hutter, A., Kosch, H., Ilic, S.: Keep it Unreal: Bridging the Realism Gap for 2.5D Recognition with Geometry Priors Only. Proceedings of International Conference on 3D Vision. pp. 1-11, 2018
9. Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J., Rother, C.: Learning 6d object pose estimation using 3d object coordinates. Proceedings of European Conference on Computer Vision. pp.536-551, 2014
10. Hinterstoisser, S, Holzer, S., Cagniart, C., Ilic, S., Konolige, K., Navab, N., Lepetit, V.: Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. Proceedings of International Conference on Computer Vision. pp. 858-865, 2011
11. Hodaň, T., Haluza, P., Obdržálek, S., Matas, J., Lourakis, M., Zabulis, X.: T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-less Objects. Proceedings of Winter Conference on Applications of Computer Vision, pp. 880-888, 2017

12. Park, K., Patten, T., Vincze, M.: Pix2Pose: Pixel-Wise Coordinate Regression of Objects for 6D Pose Estimation. Proceedings of International Conference on Computer Vision. 2019
13. Rad, M., Lepetit, V.: BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects Without Using Depth. Proceedings of IEEE International Conference on Computer Vision. 2017
14. Xiang, Y. Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. Proceedings of Robotics: Science and Systems. 2018
15. Oberweger, M., Rad, M., Lepetit, V.: Making Deep Heatmaps Robust to Partial Occlusions for 3D Object Pose Estimation. Proceedings of European Conference on Computer Vision, 2018
16. Park, K., Patten, T., Prankl, J., Vincze, M.: Multi-Task Template Matching for Object Detection, Segmentation and Pose Estimation Using Depth Images. Proceedings of International Conference on Robotics and Automation. pp. 7207-7213, 2019
17. Peng, S., Liu, Y., Huang, Q., Bao, H., Zhou, X.: PVNet:Pixel-wise Voting Network for 6DoF Pose Estimation. The IEEE Conference on Computer Vision and Pattern Recognition. pp. 4561-4570, 2019
18. Tekin, B.,Sinha, S., Fua, P.: Real-Time Seamless Single Shot 6D Object Pose Prediction. Proceedings of Conference on Computer Vision and Pattern Recognition. pp. 292-301, 2018
19. Besl, P., McKay, N.: Method for registration of 3-D shapes. Sensor Fusion IV: Control Paradigms and Data Structures. pp. 586-607, 1992
20. Lin, T., Goyal, P., Girshick, R., He, K., Piotr , P.: Focal Loss for Dense Object Detection. Proceedings of IEEE International Conference on Computer Vision. pp. 2999-3007, 2017
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 770-778, 2016
22. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., and others: Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115. pp. 211-252, 2015
23. Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117-2125, 2017
24. Crivellaro, A. Rad, M., Verdie, Y., Yi, K., Fua, P., Lepetit, V.: A Novel Representation of Parts for Accurate 3D Object Detection and Tracking in Monocular Images. Proceedings of IEEE International Conference on Computer Vision. pp. 4391-4399, 2015
25. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. International Conference on Intelligent Robots and Systems. pp. 23-30, 2017
26. Perlin, K.: Improving noise. ACM Transactions on Graphics, 21. pp. 681-682, 2002