CroCPS: Addressing Photometric Challenges in Self-Supervised Category-Level 6D Object Poses with Cross-Modal Learning

Pengyuan Wang[1]
pengyuan.wang@tum.de
Lorenzo Garattoni[2]
lorenzo.garattoni@toyota-europe.com
Sven Meier[2]
sven.meier@toyota-europe.com
Nassir Navab[1]
nassir.navab@tum.de
Benjamin Busam[1]
b.busam@tum.de

[1] Technical University of Munich
Munich, Germany
[2] Toyota Motor Europe
Brussel, Belgium

Abstract

Estimating 6D object poses for everyday household objects is a crucial and challenging task for robotic applications. Recent advances in category-level object pose estimation show great potential in this direction. Since the training of the networks relies heavily on ground truth 6D poses, which are expensive to annotate in real environments, self-supervised methods become a realistic approach to overcome the domain gap between synthetic and real images. However, these methods work poorly on photometrically-challenging objects because of the missing depth or artifacts in RGBD data.

We propose to use the polarization clues to overcome the drawbacks of RGBD images and improve the detection performance for objects with specular surfaces in the self-supervision stage. To this end, we generate a synthetic dataset containing cutlery of various shapes and sizes, and a markerless real dataset with accurate 6D pose annotations. We introduce several novel losses for self-supervision based on inputs of multiple modalities which fully utilize the polarization information. The experiment result shows that the proposed method improves both 2D detection and 3D IoU of the predicted bounding boxes over SOTA methods without usage of annotated ground truth. This work constitutes the first solution for self-supervision on challenging reflective objects and explores the usage of polarization images. We evaluate the effectiveness of the proposed pipeline by proposing synthetic and real data and thorough evaluations.

1 Introduction

Category-level object pose estimation is a key task in computer vision. The problem is challenging as it requires real images with accurate 6D object pose annotations which are intricate to acquire [3, 35]. Since 6D poses of objects are difficult to be annotated, training on synthetic images constitutes a valid alternative, while the domain gap between synthetic and real images remains a challenge. To overcome the domain gap, self-supervision methods such as CPS++ [23] are proposed.

Current self-supervision methods rely heavily on depth sensors and perform poorly on photometrically challenging objects due to missing values and artifacts in the depth images. Depth sensors such as time-of-flight (ToF) cameras measure the time that the light signal takes to bounce back from the object surface, and calculate the depth based on this elapsed time which is largely influenced by reflection and refraction properties and leads to incorrect results for photometrically challenging objects [12]. However, these object categories are very common in household environments. Although active sensors fail on reflective surfaces, passive sensors like polarization cameras capture light information related to surface normals [29]. We make use of polarization images to complement RGB-D images and overcome the photometric challenges for the domain adaptation, while only requiring RGB images for inference.

Category-level 6D object pose estimation datasets, such as NOCS [33], contain synthetic and real images of several object categories, but typically do not feature common photometrically challenging objects such as cutlery. Therefore, we create a synthetic dataset with 10k images containing cutlery of various shapes and sizes for training, and a multimodal real dataset containing cutlery in daily scenarios with accurate 6D pose annotations. We then propose a self-supervision pipeline to overcome the domain gap between. Our pipeine consists of two steps. The first step is the self-supervision of the 2D detections. Due to the domain gap between the training (synthetic) images and the real images, there are false positives and missing detections in the real images. The light becomes polarized on specular surfaces of metallic objects, therefore leading to higher values of degree of linear polarization (DOLP) than the surrounding environment. The polarization information is leveraged to determine the false positives in the detections and preserve correct detections for finetuning the network. The result shows that both the average precision and recall improve after self-supervision. In the second step, multiple novel losses are proposed for self-supervision of the 9D object bounding boxes, which includes the rotation, translation and scales of the object. In the 3D lifting module, the normalized object shape, object scales, along with rotations and translation of the object is predicted. Then, the predicted object model in 3D space is fed into a differentiable renderer to generate the rendered mask and depth image, where the normal map is also calculated. The rendered mask should be consistent with the mask extracted from DOLP image within the predicted bounding box, and the mask loss is calculated with focal loss between the masks. The normal image derived from the predicted depth image is compared with the possible normal directions from the polarization image, as a guidance for the object shape and poses. To better utilize the depth map as a complement, artifacts in the real depth map are analysed and removed, while the remaining part is utilized for self-supervision to reveal the scaling of objects in the image.

In summary, we provide three main contributions in this paper:

1. Our work is the first to investigate self-supervision of category-level object pose estimation networks for photometrically challenging objects. To this end, we introduce a dataset including both synthetic and real images of reflective objects, as a complement for the pre-

vious dataset [33], which did not contain photometrically challenging categories.
2. We introduce the use of polarization images in the self-supervision stage of the training to reveal the material property and normal angles, which active sensors such as depth cameras fail to highlight. At inference, the network works on RGB images only, which simplifies the deployment and scalability.
3. In the training stage, we leverage polarization clues to verify success detection and improve the 2D detection results. For self-supervision of the 6D pose and scales of the objects, we propose novel losses combining the polarization information and valid depth measurements. The evaluations show the effectiveness of the pipeline and the proposed losses.

2 Related Work

2.1 Instance-level 6D Object Pose Estimation

6D object pose estimation networks for instance-level objects have achieved great advances recently [17, 27, 37]. The networks can be categorized into three types, direct regression methods, keypoint-based methods or with latent representations. The direct regression methods extract the bounding box features and directly regress the translation and rotation of the objects. [36] estimates the translation of object by predicting the center point and depth of center, while estimating the rotation as quaternion. [14] [21] directly predicts the 3d coordinates of the object. [2] extends 2d EfficientNet [28] for 6d pose and propose 6d augumentation methods which greatly boosts the performance. [30] combines features from both rgb and depth images to regress the object pose. [15] utilizes object candidates from multiple images for the global scene refinement.

2.2 Category-level 6D Object Pose Estimation

The need for an instance-specific network can be a limitation which is difficult to overcome [26]. For this task, researchers designed category-level methods for 6D object pose estimation. These can be divided into monocular or rgb-d based approaches. CPS++ [23] estimates the normalized object point cloud as well as the scales, rotation and translations of the objects from the 2d features. [6] utilizes implicit representations to representing the appearance, shape and pose of category-level objects which is utilized in the inference time. [9] predicts the pseudo depth image and nocs representation from the monocular image and estimates the object poses by alignment.
Most of the works leverage depth images for estimating category-level object poses. NOCS [33] estimates the normalized coordinate space of the object and lifts to 3d with corresponding depth map. [4] explores intra-class variation based on the category priors and detects object key-points in the point cloud, to get the deformation and pose of the objects. [22] utilizes spherical convolutions to better extract the features from the point cloud. [5] introduces graph convolutions to extract features from rgb-d inputs and use decoupled heads to predicts the translation and rotations. [8] further leverages a voting-based method in the point cloud to better get the poses of the objects. DualPoseNet [22] uses two parallel pose decoders on top of a shared pose encoder and performed spherical convolution on the point cloud to fully utilize RGBD images there. [20] estimates the pose considering the articulated objects.

2.3 Domain Adaptation for 6D Object Pose Estimation

Getting accurate 6D object poses takes a lot of effort and training with few [34] or weak labels [18, 19] as well as synthetic images is a popular solution. However, with the domain gap, it is hard to deploy the trained networks in the real environments. Therefore, domain adaptation has been an attractive topic in the field. For the self-supervision of the instance-level 6D pose estimation networks, [31] employs both visual alignments and depth alignments to refine the predicted object poses with a differentiable renderer. [32] extends the self-supervision approach to occluded objects. For category-level objects, [16] uses a student-teacher network and bidirectional point filtering to align the predicted point cloud with the depth image for self-supervision of the network. [23] predicts the attention maps in the region of interest, and aligns the predicted object point cloud with the depth maps in the attention maps with Chamfer distances. Instead of predicting the object shape as point cloud, [24] leverages implicit representations of the category objects with depth images for self-supervision.

2.4 Object Pose Detection and Shape Recovery with Polarization Image

Depth cameras fail to measure the surface distances of photometrically challenging objects, while polarization camera can reveal the material properties by capturing the change in the polarization state. [13] utilizes the polarization information for the segmentation of transparent objects and the results improved greatly in comparision with rgb inputs. Methods such as [40], [39] estimates the depth and normal of objects from a stereo pair or a sequence of images with physical derivations. [38] estimates the missing values of the depth camera with the help of polarization images for metallic car components. [1] further directly estimates the object surface normal from polarization images with neural networks. [10] uses the polarization images for instance-level 6d pose by predicting the normal, NOCS map of the objects. However, no work has been done on polarization images for category-level object pose estimations.

3 Dataset Preparation

Synthetic dataset generation To generate the synthetic dataset, object models of cutlery with various shapes and sizes are collected. Afterwards, BlenderProc [7] is leveraged to generate the synthetic dataset. The elevation angle is set between 30 and 60 degrees and the in-plane rotations are set between -90 to 90 degrees. Physical positioning of the objects are used and the cutlery are rendered with glossy material in blender. In total 10k images are generated for the training, we will opensource the rgb and depth images, along with the annotated groundtruth in the dataset.

Real dataset generation A polarization camera and rgb-d camera are used for recording the real dataset. The cutlery objects are pre-scanned and placed in a household environments. After calibration the image sequences are recorded for both cameras as in [35]. Since the two cameras are calibrated with the same calibration board, the depth camera transformation with respect to the polarization image is calculated and the depth images are aligned to the polarization images by rendering the transformed point cloud.

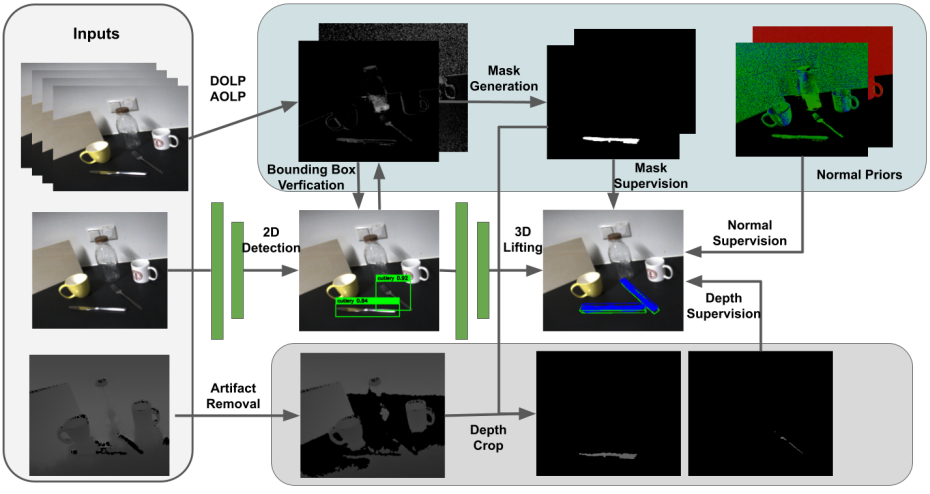


Figure 1: The illustration of our proposed self-supervision approach. The inputs are polarization images, which are averaged as a single RGB image, and the depth image. From the polarization images input, the degree of linear polarization (DOLP), the angle of linear polarization (AOLP) and normal priors are derived. Afterwards the object mask is extracted from DOLP and utilized to crop the depth map, which is processed with artifact removal, as the supervision signal. After the 2D self-supervision and lifting to 3D space, the predicted object shapes, poses and sizes are rendered into the object mask, depth, normal images, which are leveraged in combination with extracted masks, normal priors, cropped depth images for the self-supervision.

4 Methodology

Although the former works [23], [16] focus on the domain adaptation for the 6D pose of object, 2D object detections trained from synthetic images are assumed to be accurate in the real environment. However, since the cutlery objects have high reflectivity and it is appearance depends highly on the light sources and surrounding illuminations, which is hard to fully simulate during rendering, there are false positives and missing detections in the trained network. The false positives in the detections convey wrong information in the self-supervision stage of 6D poses and reduce the overall performance. Therefore, we design a novel approach to distinguish between good and bad detections in the challenging environment.

Self-Supervision for 2D Detections Though the light signal sent from the depth sensor can not be reflected back and measured by the receiver on metallic surfaces, the change of the polarization state of the environment light is captured by the polarization camera. The environment light is unpolarized and becomes partially polarized by the specular reflection on the object surface, which results in higher value of degree of linear polarization than the surrounding environments. Therefore we use a threshold to get the potential object mask inside the bounding box. It is observed that the cutlery contains thin structures, which are neglected when searching for the largest connected component. Therefore morphological operations, which includes dilation followed by erosion with a kernel of 5x5, are performed. The connected components are then extracted within the bounding box and the one with

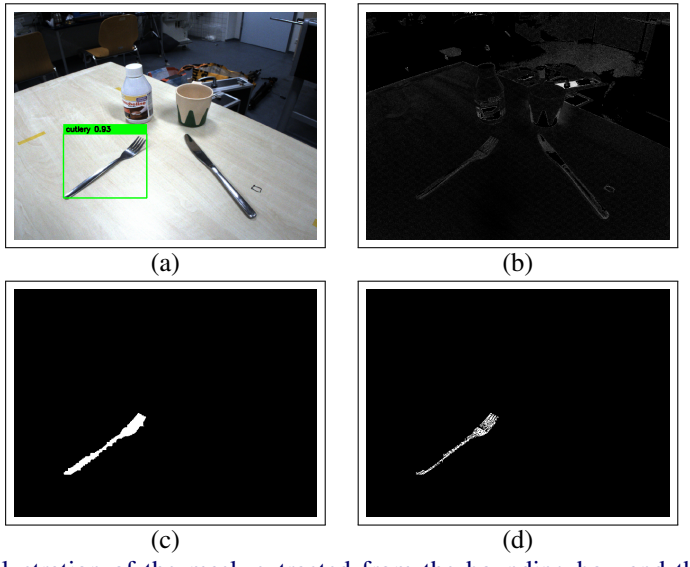


Figure 2: Illustration of the mask extracted from the bounding box and the polarization image, (a) input image, (b) degree of linear polarization, (c) extracted object mask with morphological operation, (d) extracted object mask without morphological operation.

largest area is assumed to be the object mask. The procedure of mask extraction from polarization images is illustrated in Fig. 2. Based on the assumption that the object should exist and be inside the bounding box, several criteria are set to determine whether the predictions are good or not. Firstly the proposals with mask area smaller than a threshold, which we set as 50, are considered as false positives and discarded. The areas look bright but actually contain objects with diffuse reflections, which is visualized in Fig. 3 (a). Secondly, the minimum and maximum of the x,y coordinates are calculated from the extracted mask, which should be inside the bounding box and not reaching the boundary of the bound box. With this step, bounding boxes such as objects are partially inside are avoided. After the processing of unfitted bound boxes, the remaining object bounding boxes are used as 2D self-supervision signal for domain adaptation.

Polarization Image Processing The polarization image consists of rgb image of four different angles I_0,I_1,I_2,I_3 at 0, 45, 90, 135 degrees. The polarization rgb is calculated as

$$I_{rgb} = \frac{1}{4} \cdot (I_0 + I_1 + I_2 + I_3) \tag{1}$$

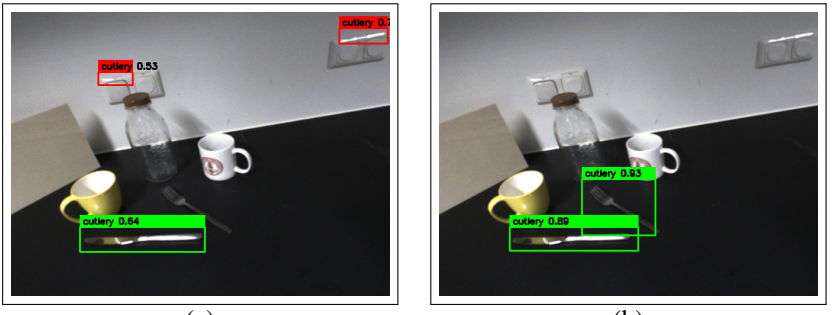
The intensity of the polarization image is defined as

$$I(\Phi) = \frac{I_{max} + I_{min}}{2} + \frac{I_{max} - I_{min}}{2} \cdot \cos(2\Phi - 2\phi) \tag{2}$$

The degree of polarization is calculated as

$$\rho = \frac{I_{max} - I_{min}}{I_{max} + I_{min}} \tag{3}$$

I_{max} and I_{min} are the maximum and minimum values of the four observations in Equ. 2 and 3. The degree of polarization ρ can be derived from zenith angle θ for diffuse surfaces



(a) (b)

Figure 3: (a) 2D detections without self-supervision, where red boxes indicate false positives while the green boxes represent verified good predictions, (b) 2D detections after self-supervision, where objects are correctly predicted.

in Equ. 4 and specular surfaces in Equ. 5, where n is the refractive index of the material, which is normally set as 1.5.

$$\rho = \frac{(n-\frac{1}{n})^2 \sin^2\theta}{2+2n^2-(n+\frac{1}{n})^2\sin^2\theta+4\cos\theta\sqrt{n^2-\sin^2\theta}} \tag{4}$$

$$\rho = \frac{2\cos\theta\sin^2\theta\sqrt{(n^2-\sin^2\theta)}}{\cos^2\theta(n^2-\sin^2\theta)+\sin^4\theta} \tag{5}$$

The zenith angle θ is θ_1 and θ_2 in Equ. 5 and there are also two possible azimuth angles $\phi+\frac{\pi}{2}$,$\phi-\frac{\pi}{2}$, because of π-ambiguity. Although two zenith angles can be recovered from DOLP, only the zenith angle with a lower value is considered, because the range of the higher value is close to 90 degrees for specular surfaces and less probable in practice, similar to [38]. Afterwards, the two possible surface normals are recovered by two possible azimuth angles and the zenith angle with lower value in Equ. 6.

$$n = \begin{bmatrix} cos\psi \cdot \sin\theta \\ sin\psi \cdot \sin\theta \\ cos\theta \end{bmatrix} \tag{6}$$

Self-Supervision for Category-Level 6D Poses and Sizes The YoloX [11] Nano model is leveraged for the 2D detections, where the FPN features of sizes 64x60x80, 128x30x40, 256x15x20 are upsampled and concatenated as features of size 448x60x80. The ROI features inside the 2D bounding boxes are extracted by ROI alignment and used for 3D lifting. As shown in Fig. 1, the 6D pose and shape of objects are decomposed into allocentric rotations in quaternion, translation, object scales, and the normalized point cloud, similar to [23]. The translation is represented by the projection residual of the object center to the 2D bounding box center, and the depth of the object center. The normalized point cloud is learned by a point cloud autoencoder pretrained on synthetic object models with a bottleneck size of 32. The quaternion, projection residual, center depth, scales and the shape encodings are generated by individual encoders from the 2D features. Then the objects are rendered with a differentiable renderer module to the object masks and depth image, where the depth image is transformed into normal map with Korinia [25].

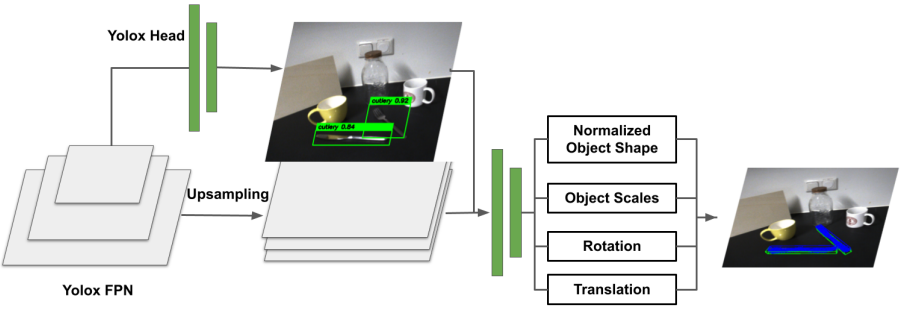


Figure 4: The 3D lifting module to derive the object 6D poses and sizes. The YoloX [□] FPN features are upsampled to the same scale and fed into the network to estimate the normalized object shape, scales, translation and rotation of the objects.

For the self-supervision of the object 6D poses and shapes, novel losses are proposed consisting of three parts, the polarization mask loss, the surface normal loss, the geometric loss (Equ. 7). The pipeline is visualized in Fig. 1.

$$\mathcal{L}_{all} = \mathcal{L}_{pol}^{mask} + \mathcal{L}_{pol}^{normal} + \mathcal{L}_{pol}^{geo} \tag{7}$$

Mask Loss As described in above sections, the object mask extracted from image bounding boxes and degree of linear polarization (DOLP) is marked as M_{init}. After the morphological operations and extraction of the largest connected component, the mask is recorded as M_{pol}. The loss term is formulated by comparing M_{render} and M_{pol} with focal loss, to deal with the imbalance of the positive and negative samples, as shown in Equ. 8. Using polarization image for mask loss leverages the unique material property of metallic objects and segments the object out even under noisy color information, which is challenging for monocular images.

$$\mathcal{L}_{pol}^{mask} = -\frac{1}{|N_+|}\sum_{j\in N_+} M_j^{pol}\log M_j^{render} - \frac{1}{|N_-|}\sum_{j\in N_-}(1-M_j^{pol})\log M_j^{render} \tag{8}$$

Polarization Normal Loss The differentiable rendering module provides rendered depth images. To convert the depth image to normal map, the depth image are reprojected to 3d space and the spatial gradients are calculated. The normal map $\hat{N}_r$ is calculated by cross product of the two gradients. Given the two possible normal directions for specular surfaces $\hat{N}_{\theta_1}, \hat{N}_{\theta_2}$, D is the intersection of the predicted mask and polarization mask, the normal loss is defined in Equ. 9. The log function is used to reduce the influence of the possible normal outliers.

$$\mathcal{L}_{pol}^{normal} = \frac{1}{N_j}\sum_{j\in D}\min(\log(1+\arccos(\hat{N}_{j,r}\cdot\hat{N}_{j,\theta_1})) + \log(1+\arccos(\hat{N}_{j,r}\cdot\hat{N}_{j,\theta_2}))) \tag{9}$$

Geometric Loss In [23],[16], the depth images are assumed to be accurate and used directly for the self-supervision of 6D poses. However, it is observed that for metallic object, the depth map from ToF camera have missing pixels and artifacts. Part of the depth image is missing because the receiver of the depth sensor fails to measure the light signal after specular reflection on object surfaces. Parts of the depth images have larger depth values than the ground truth, because the light signal from the depth camera is reflected to other nearby objects and reflected back to the receiver, which measures a longer flight time of the light and results in larger depth values. Therefore, we adopt the plane assumptions for the environment and remove the depth values which are physically impossible as artifacts in the depth image. The L1 loss is applied to the rendered depth map and filtered real depth map for self-supervision.

5 Experiment

In this section, we introduce the experiment settings and results of the proposed pipeline. The model is pretrained on the synthetic dataset and trained with our self-supervision method on real dataset.

5.1 Training Settings

The training is in two steps. Firstly, the Yolo-X model is trained on the synthetic dataset and fine-tuned on the real scenes. Secondly, the 3D lifting module is trained to estimate the 9D bounding box with ground truth synthetic data, and then refined with the proposed self-supervision losses in the real scene. The mask loss, the normal loss and the geometric loss are multiplied with factors of 50, 50, 1000. The self-supervision network is trained for 15000 iterations with a SGD optimizer and a base learning rate of 1e-5.

5.2 Analysis on 2D Detections

The 2D detection results are evaluated with and without self-supervision. The results of the average precision and recall, along with F1 scores, are recorded in Tab. 1. The result shows that through the self-supervision with polarization images, both the average accuracy and recall become higher. Especially for AP_{50} and $Recall_{50}$, the results are improved greatly, which makes the self-supervision signal in the next stage more accurate. Qualitative results are visualized in Fig. 3.

5.3 Ablation Study

To evaluate the effectiveness of the proposed losses, an ablation study is conducted and the results are listed in Tab. 2. The result shows that with improved 2D detections, the 3D IoU at a threshold of 0.25 can reach 85% and 3D IoU at a threshold of 0.5 can reach 30%. The valid depth pixels without artifacts play an important role in estimating the scales of objects in monocular images and the normal priors from polarization images improve the 3D IoU results.

	AP_{50}	Recall $_{50}$	AP	Recall	F1
w/o self-supervision	60.47	63.75	36.28	38.75	37.47
with self-supervision	**100**	**100**	**45.89**	**45.89**	**45.89**

Table 1: Evaluation results of 2D detections

$3D_{25}$ / $3D_{50}$	$3D_{25}$	$3D_{50}$
mask+normal	0	0
mask+depth	81.25	25
mask+normal+depth	**85**	**30**

Table 2: Ablation. We report 3D IoU results on the real dataset.

$3D_{25}$ / $3D_{50}$	$3D_{25}$	$3D_{50}$
CPS++	42.5	0
Ours	**88.75**	**30**

Table 3: Comparison with state of the art on 3D IoU results

5.4 Comparison with State-of-the-art

Since no prior work has been done focusing on the self-supervision of photometrically challenging categories, we compare our method with CPS++, which performs domain adaptation for categories with diffuse reflections. The result shows that our method outperforms CPS++ by 42.5% for 3D IoU at a threshold of 0.25 and by 30% for 3D IoU at a threshold of 0.5. The overall performance improvements come from cross-model learning from polarization and depth inputs, where the accurate object mask and normal priors from polarization images help to reconstruct missing depth pixels.

6 Conclusions

In this paper, we introduce a multimodal dataset including synthetic and real images, and address the problem of self-supervision of photometrically challenging categories. To this end, we propose a cross-modal learning pipeline, which combines the strengths of polarization and depth modality, for estimating the 6D poses and sizes of object categories with specular surfaces. The evaluation result shows the effectiveness of our pipeline. In addition to specular objects, refractive objects such as glass are also quite common in household environments, which is our target in the future work.

6.1 Limitations

The polarization segmentations could be affected by other reflective materials in the background. However this can be avoided in many occasions.

7 Acknowledgements

The authors would like to express their thanks to Dr Seena Rejal of Shapes AI (www.shapes.ai) for supplying the dataset of high-quality 3D models used in this work.

References

[1] Yunhao Ba, Alex Gilbert, Franklin Wang, Jinfa Yang, Rui Chen, Yiqin Wang, Lei Yan, Boxin Shi, and Achuta Kadambi. Deep shape from polarization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 554–571. Springer, 2020.

[2] Yannick Bukschat and Marcus Vetter. Efficientpose: An efficient, accurate and scalable end-to-end 6d multi object pose estimation approach. *arXiv preprint arXiv:2011.04307*, 2020.

[3] Benjamin Busam, Hyun Jun Jung, and Nassir Navab. I like to move it: 6d pose estimation as an action decision process. *arXiv preprint arXiv:2009.12678*, 2020.

[4] Kai Chen and Qi Dou. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2773–2782, 2021.

[5] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Ales Leonardis. Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1581–1590, 2021.

[6] Xu Chen, Zijian Dong, Jie Song, Andreas Geiger, and Otmar Hilliges. Category level object pose estimation via neural analysis-by-synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–156. Springer, 2020.

[7] Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Youssef Zidan, Dmitry Olefir, Mohamad Elbadrawy, Ahsan Lodhi, and Harinandan Katam. Blenderproc. *arXiv preprint arXiv:1911.01911*, 2019.

[8] Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6781–6791, 2022.

[9] Zhaoxin Fan, Zhenbo Song, Jian Xu, Zhicheng Wang, Kejian Wu, Hongyan Liu, and Jun He. Object level depth reconstruction for category level 6d object pose estimation from monocular rgb image. *arXiv preprint arXiv:2204.01586*, 2022.

[10] Daoyi Gao, Yitong Li, Patrick Ruhkamp, Iuliia Skobleva, Magdalena Wysocki, HyunJun Jung, Pengyuan Wang, Arturo Guridi, and Benjamin Busam. Polarimetric pose prediction. In *European Conference on Computer Vision (ECCV)*, October 2022.

[11] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[12] HyunJun Jung, Patrick Ruhkamp, Guangyao Zhai, Nikolas Brasch, Yitong Li, Yannick Verdie, Jifei Song, Yiren Zhou, Anil Armagan, Slobodan Ilic, et al. Is my depth ground-truth good enough? hammer–highly accurate multi-modal dataset for dense 3d scene regression. *arXiv preprint arXiv:2205.04565*, 2022.

[13] Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. Deep polarization cues for transparent object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8602–8611, 2020.

[14] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1521–1529, 2017.

[15] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *European Conference on Computer Vision*, pages 574–591. Springer, 2020.

[16] Taeyeop Lee, Byeong-Uk Lee, Inkyu Shin, Jaesung Choe, Ukcheol Shin, In So Kweon, and Kuk-Jin Yoon. Uda-cope: Unsupervised domain adaptation for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14891–14900, 2022.

[17] Fu Li, Ivan Shugurov, Benjamin Busam, Minglong Li, Shaowu Yang, and Slobodan Ilic. Polarmesh: A star-convex 3d shape approximation for object pose estimation. *IEEE Robotics and Automation Letters*, 7(2):4416–4423, 2022.

[18] Fu Li, Ivan Shugurov, Benjamin Busam, Shaowu Yang, and Slobodan Ilic. Ws-ope: Weakly supervised 6-d object pose regression using relative multi-camera pose constraints. *IEEE Robotics and Automation Letters*, 7(2):3703–3710, 2022.

[19] Fu Li, Hao Yu, Ivan Shugurov, Benjamin Busam, Shaowu Yang, and Slobodan Ilic. Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation. *arXiv preprint arXiv:2203.04802*, 2022.

[20] Xiaolong Li, He Wang, Li Yi, Leonidas J Guibas, A Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3706–3715, 2020.

[21] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 683–698, 2018.

[22] Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. 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*, pages 3560–3569, 2021.

[23] Fabian Manhardt, Gu Wang, Benjamin Busam, Manuel Nickel, Sven Meier, Luca Minciullo, Xiangyang Ji, and Nassir Navab. Cps++: Improving class-level 6d pose and shape estimation from monocular images with self-supervised learning. *arXiv preprint arXiv:2003.05848*, 2020.

[24] Wanli Peng, Jianhang Yan, Hongtao Wen, and Yi Sun. Self-supervised category-level 6d object pose estimation with deep implicit shape representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2082–2090, 2022.

[25] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3674–3683, 2020.

[26] Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic. Osop: A multi-stage one shot object pose estimation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6835–6844, 2022.

[27] Yongzhi Su, Mahdi Saleh, Torben Fetzer, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari. Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6738–6748, 2022.

[28] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[29] Yannick Verdié, Jifei Song, Barnabé Mas, Benjamin Busam, Ales Leonardis, and Steven McDonagh. Cromo: Cross-modal learning for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3937–3947, June 2022.

[30] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3343–3352, 2019.

[31] Gu Wang, Fabian Manhardt, Jianzhun Shao, Xiangyang Ji, Nassir Navab, and Federico Tombari. Self6d: Self-supervised monocular 6d object pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 108–125. Springer, 2020.

[32] Gu Wang, Fabian Manhardt, Xingyu Liu, Xiangyang Ji, and Federico Tombari. Occlusion-aware self-supervised monocular 6d object pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[33] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019.

[34] Pengyuan Wang, Fabian Manhardt, Luca Minciullo, Lorenzo Garattoni, Sven Meier, Nassir Navab, and Benjamin Busam. Demograsp: Few-shot learning for robotic grasping with human demonstration. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5733–5740. IEEE, 2021.

[35] Pengyuan Wang, HyunJun Jung, Yitong Li, Siyuan Shen, Rahul Parthasarathy Srikanth, Lorenzo Garattoni, Sven Meier, Nassir Navab, and Benjamin Busam. Phocal: A multimodal dataset for category-level object pose estimation with photometrically challenging objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21222–21231, 2022.

[36] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

[37] Yan Xu, Kwan-Yee Lin, Guofeng Zhang, Xiaogang Wang, and Hongsheng Li. Rnnpose: Recurrent 6-dof object pose refinement with robust correspondence field estimation and pose optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14880–14890, 2022.

[38] Tomonari Yoshida, Vladislav Golyanik, Oliver Wasenmüller, and Didier Stricker. Improving time-of-flight sensor for specular surfaces with shape from polarization. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1558–1562. IEEE, 2018.

[39] Ye Yu, Dizhong Zhu, and William AP Smith. Shape-from-polarisation: a nonlinear least squares approach. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2969–2976, 2017.

[40] Dizhong Zhu and William AP Smith. Depth from a polarisation+ rgb stereo pair. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7586–7595, 2019.