Polarimetric Pose Prediction
—
Supplementary Material

Daoyi Gao*, Yitong Li*, Patrick Ruhkamp*, Iuliia Skobleva*, Magdalena Wysocki*, HyunJun Jung, Pengyuan Wang, Arturo Guridi, and Benjamin Busam

* *Equal contribution; Alphabetical order*
Technical University of Munich, Germany
{d.gao,...,b.busam}@tum.de

A1 Physical Priors

We use physical priors as inputs in our network to improve the estimated 6D pose of an object. These priors form relations between polarisation properties and azimuth and zenith angle of the surface normal, which serve as geometric cues orthogonal to color information. We calculate the physical priors under the assumption of either specular or diffuse reflection. To recover the azimuth and zenith angle of the surface normal, we present the calculation for solving the unknowns of Eq. A1.

A polarimetric camera registers intensity behind four linear polarisers with angles $0°, 45°, 90°, 135°$, which depends on unpolarised intensity I_{un}, degree of polarisation ρ, and angle of polarisation ϕ:

$$I_{\varphi_{pol}} = I_{un} \cdot \ (1 + \rho \ \cos(2(\phi - \varphi_{pol}))). \tag{A1}$$

Eq. A1 can be re-written as:

$$I_{\varphi_{pol}} = \underbrace{\begin{pmatrix} 1 \\ \cos 2\varphi_{pol} \\ \sin 2\varphi_{pol} \end{pmatrix}^T}_{\boldsymbol{\beta^T}} \underbrace{\begin{pmatrix} I_{un} \\ \rho \cos 2\phi \\ \rho \sin 2\phi \end{pmatrix}}_{\boldsymbol{x}}. \tag{A2}$$

For all angles $\varphi_{pol} \in \{0°, 45°, 90°, 135°\}$, we get a linear equation system for each pixel location with $\boldsymbol{I_{\varphi_{pol}} \in \mathbb{R}^{4\times1}}$, $\boldsymbol{\beta \in \mathbb{R}^{3\times4}}$ and $\boldsymbol{x \in \mathbb{R}^{3\times1}}$. After solving this over-determined linear equation system using least squares, we find unpolarised intensity, degree of polarisation and angle of polarisation:

$$\begin{aligned} I_{un} &= x_1 \\ \rho &= \sqrt{x_2^2 + x_3^2} \\ \phi &= \frac{1}{2} \arctan \frac{x_3}{x_2} \end{aligned}. \tag{A3}$$

The azimuth angle can be found using Eq.2. Then, we can estimate the azimuth angle θ from Eq.3 by linear interpolation. Both models take in the same value for the refractive index η, since it is an intrinsic property of the material and it does not depend on the reflection model. The values used for our objects can be seen in Tab. A1.

Table A1. Refractive Indices. Refractive indices per object with certain material used for the physical model of **PPP-Net**.

Object	Material	Refractive Index
Teapot	ceramic	1.54
Can	aluminium composite	1.35
Fork	stainless steel	2.75
Knife	stainless steel	2.75
Bottle	glass	1.52
Cup	plastics	1.50

A2 Additional Experiments and Ablation Studies

Runtime Analysis. On a desktop PC with an Intel i7 4.20GHz CPU and an NVIDIA 2080 GPU, given a 512×612 pixel image, our network takes ca. 64 ms for a single object, including 40 ms for detection, and 13 ms to calculate the physical priors with our non-optimized implementation.

A2.1 Ablations on Modalities

Ablations on Input Modalities. Tab. A2 is an extension to Tab.1 in the main paper and summarises the quantitative evaluation for different modalities for **PPP-Net** for all objects under consideration in the dataset.
Ablations on Output Modalities. 6D pose estimation mainly depends on accurate correspondences prediction by NOCS regression as reported in the ablation in Tab. A3. The ADD drops significantly for the model without (w/o) NOCS output before Patch-PnP, i.e. only shape information is utilised for pose prediction. Still, as proven by the ablations in the paper, the auxiliary explicit prediction of object-centric shape information as normals map benefits 6D pose estimation as the network is more strongly guided towards extracting physical shape priors from the input.

A2.2 Ablations on Network Architecture

Tab. A4 indicates naively concatenating geometric priors and RGBP images for direct input to the network (as in [5]) results in inferior normal prediction quality, and also leads to less improvement on pose estimation results (compare `concat`

Table A2. PPP-Net Input Modalities Evaluation. Different combinations of input and output modalities are used for training to study their influence on pose estimation accuracy ADD(-S) for objects with different photometric complexity. Where applicable, metrics for estimated normals are reported as well.

Object	Photo. Chall.	Input Modalities			Output Variants		Normal Metrics					Pose Metric
		RGB	Polar RGB	Physical N	Normals	NOCS	mean↓	med.↓	11.25°↑	22.5°↑	30°↑	ADD(-S)
Cup		✓				✓	-	-	-	-	-	91.1
			✓			✓	-	-	-	-	-	91.3
			✓		✓	✓	7.3	5.5	86.2	96.1	97.9	91.3
			✓	✓	✓	✓	**4.5**	**3.5**	**94.7**	**99.1**	**99.6**	**97.2**
Teapot	†	✓				✓	-	-	-	-	-	97.8
			✓			✓	-	-	-	-	-	99.5
			✓		✓	✓	7.9	5.4	82.5	94.5	97.1	99.2
			✓	✓	✓	✓	**5.3**	**4.0**	**91.6**	**98.7**	**99.5**	**99.9**
Can	†	✓				✓	-	-	-	-	-	91.8
			✓			✓	-	-	-	-	-	93.2
			✓		✓	✓	**5.7**	**3.9**	**90.0**	97.0	98.6	96.7
			✓	✓	✓	✓	6.0	4.5	89.0	**97.3**	**98.9**	**98.4**
Fork	††	✓				✓	-	-	-	-	-	85.4
			✓			✓	-	-	-	-	-	86.1
			✓		✓	✓	11.0	7.3	72.6	90.7	93.9	92.9
			✓	✓	✓	✓	**6.5**	**4.3**	**87.6**	**95.9**	**97.6**	**95.9**
Knife	††	✓				✓	-	-	-	-	-	84.1
			✓			✓	-	-	-	-	-	88.0
			✓		✓	✓	12.2	8.0	68.7	88.5	92.4	89.4
			✓	✓	✓	✓	**6.8**	**5.4**	**88.2**	**97.3**	**98.6**	**96.4**
Bottle	†††	✓				✓	-	-	-	-	-	90.5
			✓			✓	-	-	-	-	-	93.5
			✓		✓	✓	5.6	4.7	**92.9**	**99.0**	**99.6**	94.7
			✓	✓	✓	✓	**5.4**	**4.5**	92.1	**99.0**	**99.6**	**97.5**

Table A3. PPP-Net Output Ablation. With and without NOCS output.

Object	Pose Metric (ADD)	
Teapot	w/ **99.9**	w/o 72.7
Fork	w/ **95.9**	w/o 79.3

against `ours` in Tab. A4). This holds true for all objects, whereas photometrically more challenging objects show a larger relative improvement. These results confirm the importance of our design choices of **PPP-Net** to employ a dedicated encoder for the physics-based derived geometric priors, and its positive effect on 6D object pose estimation results. We thus propose a careful integration design of such physical priors into established principles of 6D object pose estimation within our novel hybrid encoder. We deliberately choose a simple general architecture for PPP-Net for best comparison and evaluation against SOTA, and to show that even such simplistic encoders can achieve significant accuracy for 6D pose prediction with the physical priors from polarisation as inputs.

A2.3 Other Ablations

Ablation on Detector. We train an object detector using Faster R-CNN without additional modification of polarimetric inputs. It is not affected by the

Table A4. Fusion Ablation. Naive concatenation against our proposed fusion strategy of RGB and physical priors in **PPP-Net**.

Object	Fusion	Input Modalities		Output Variants		Normal Metrics					Pose Metric
		Polar RGB	Physical N	Normals	NOCS	mean↓	med.↓	11.25° ↑	22.5°↑	30°↑	ADD
Cup	concat	✓	✓	✓	✓	6.0	4.9	91.1	98.1	99.1	93.6
Cup	ours	✓	✓	✓	✓	**4.5**	**3.5**	**94.7**	**99.1**	**99.6**	**97.2**
Teapot	concat	✓	✓	✓	✓	7.4	5.7	83.4	96.3	98.4	97.3
Teapot	ours	✓	✓	✓	✓	**5.3**	**4.0**	**91.6**	**98.7**	**99.5**	**99.9**
Can	concat	✓	✓	✓	✓	8.5	6.4	81.8	95.1	97.5	92.2
Can	ours	✓	✓	✓	✓	**6.0**	**4.5**	**89.0**	**97.3**	**98.9**	**98.4**
Fork	concat	✓	✓	✓	✓	10.7	7.8	70.0	91.8	95.0	87.6
Fork	ours	✓	✓	✓	✓	**6.5**	**4.3**	**87.6**	**95.9**	**97.6**	**95.9**
Knife	concat	✓	✓	✓	✓	10.8	8.5	67.1	92.8	96.2	86.1
Knife	ours	✓	✓	✓	✓	**6.8**	**5.4**	**88.2**	**97.3**	**98.6**	**96.4**
Bottle	concat	✓	✓	✓	✓	7.6	6.0	86.5	94.8	96.4	93.1
Bottle	ours	✓	✓	✓	✓	**5.4**	**4.5**	**92.1**	**99.0**	**99.6**	**97.5**

photometric challenges of the objects, as indicated by similar results in Tab. A5 when training/testing PPP-Net with the GT bounding box and the predicted ones.

Table A5. BBox Ablations.

Configuration	Cup	Teapot	Can	Fork	Knife	Bottle
Train with GT BBox/Test with pred BBox	97.2	99.9	98.4	95.9	96.4	97.5
Train/Test with GT BBox	99.0	99.9	99.0	96.1	97.6	97.5

Ablation on Refractive Index. As mentioned, the prior knowledge of the refractive index of materials in the scene is one limitation of our model. To analyse the impact of incorrect indices, we report pose accuracy results when trained/tested with minor (1.54 vs. 1.5) and large deviations (2.75 vs. 1.5) of the correct index in Tab. A6. The results in the 2nd row highlight that our model still performs well when providing incorrect refractive indices during inference. This indicates that the model is robust enough to extract relevant features. When training and testing with very different indices, we see a slight decrease in ADD (cf. *fork, knife*).

Table A6. Refractive Index Ablation.

Object	Cup	Teapot	Can	Fork	Knife	Bottle
Refractive Index	1.50	1.54	1.35	2.75	2.75	1.52
Train/Test with correct index	97.2	99.9	98.4	95.9	96.4	97.5
Train with correct index, test with incorrect (1.5)	97.2	99.9	98.3	95.8	96.2	97.5
Train/Test with incorrect index (1.5)	97.2	99.9	98.0	93.5	90.1	97.5

Ablation on Photometric Complexity. Recent RGB-D pipelines (to which we compare) try to overcome photometric challenges, e.g. textureless objects, by incorporating depth information. Correct depth information is essential here, but depth sensors suffer specifically in these areas. On the contrary, the physical properties encoded in the polarimetric images, which are leveraged by **PPP-Net**, preserve object-centric shape information also for very challenging (e.g. reflective, transparent) objects. We train and test CosyPose [2] on our data using single-view mode without ICP refinement or additional 1 million synthetic data as when training on T-LESS [1], which ensures the same settings for all benchmarking experiments. We outperform CosyPose for every object in Tab. A7 with significant improvement for photometrically challenging objects.

Table A7. CosyPose [2] Benchmarking.

Methods	Cup	Teapot	Can	Fork	Knife	Bottle	Mean
CosyPose [33]	88.5	94.3	91.0	83.0	89.5	79.6	87.7
Ours	**97.2**	**99.9**	**98.4**	**95.9**	**96.4**	**97.5**	**97.6**

A3 Qualitative Visualizations

In Fig. A1, we visualise the 6D pose by overlaying the image with the corresponding transformed 3D bounding box. For better visualization we cropped the images and zoomed into the area of interest.

A4 Instance-level Polarimetric Object Pose Dataset

Fig. A2 illustrates our scene settings as well as the pose annotation quality. We cover a wide range of variety in the background, illumination, as well as object settings. And our pose annotations are accurate for all objects, including the challenging reflective and transparent ones. High accuracy of annotations is achieved with the process described in [3], which involves tipping multiple times the surface of objects with a calibrated tool tip attached to a robotic arm and subsequent ICP alignment with the pre-scanned 3D mesh of the object (see Sec.4 for more details). We provide 6D pose annotations for all objects in the scene, but here only consider the objects introduced in Fig. 5 which cover a wide range of photometric complexity. Fig. A2 shows the superimposed 3D meshes of these objects with high accuracy.

Camera Alignment. The extrinsic calibration, which is derived by an hand-eye calibration against the robotic end-effector with high accuracy, is used for aligning different camera modalities. The alignment of cameras is only limited by their form factors. To reduce this effect and to bring the optical centers of all cameras as close to another as possible, we design a custom rig. However, small changes in the viewpoint of each camera cannot be completely avoided.

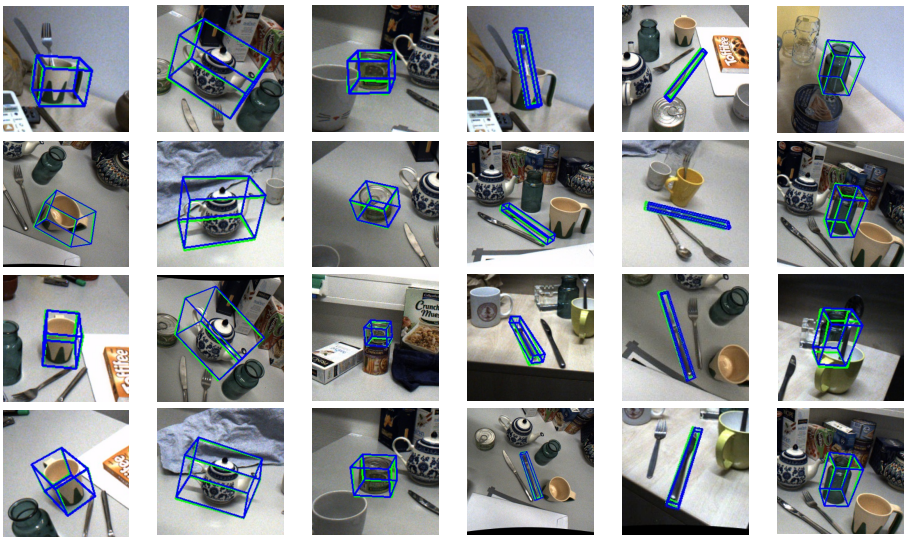

Fig. A1. Qualitative Results. Predicted and GT 6D poses are illustrated by *blue* and *green* bounding boxes, respectively.

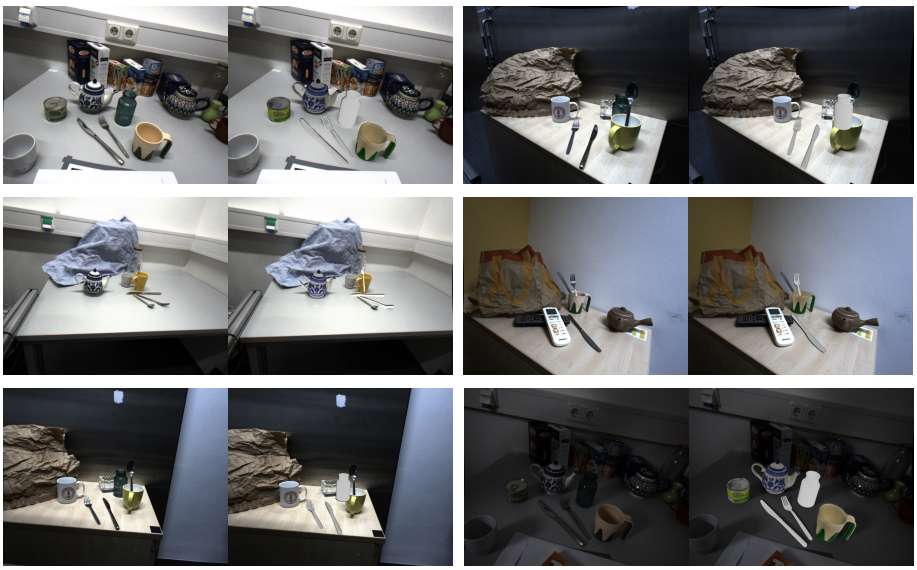

Fig. A2. Dataset and Annotation Examples. The figure shows one polarisation image together with the rendered models.

Dataset Comparison. Tab. A8 gives an overview of different dataset characteristics.

Table A8. Dataset Comparison.

Dataset	RGB	Depth	Polarisation	Robotic GT	Occlusion	Symmetry	Transparent	Reflective	Sequences
YCB-V [55]	✓	✓			✓	✓			92
T-LESS [23]	✓	✓			✓	✓			20
Linemod [21]	✓	✓			✓	✓			15
Ours	✓	✓	✓	✓	✓	✓	✓	✓	20

References

1. Hodan, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., Zabulis, X.: T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In: IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 880–888 (2017)
2. Labbé, Y., Carpentier, J., Aubry, M., Sivic, J.: Cosypose: Consistent multi-view multi-object 6d pose estimation. In: European Conference on Computer Vision (ECCV). pp. 574–591 (2020)
3. Wang, P., Jung, H., Li, Y., Shen, S., Srikanth, R.P., Garattoni, L., Meier, S., Navab, N., Busam, B.: Phocal: A multimodal dataset for category-level object pose estimation with photometrically challenging objects. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)