Polarimetric Pose Prediction

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

Technical University of Munich, Germany
{d.gao,...,b.busam}@tum.de

Abstract

Light has many properties that can be passively measured by vision sensors. Colour-band separated wavelength and intensity are arguably the most commonly used ones for monocular 6D object pose estimation. This paper explores how complementary polarisation information, i.e. the orientation of light wave oscillations, can influence the accuracy of pose predictions. A hybrid model that leverages physical priors jointly with a data-driven learning strategy is designed and carefully tested on objects with different amount of photometric complexity. Our design not only significantly improves the pose accuracy in relation to photometric state-of-the-art approaches, but also enables object pose estimation for highly reflective and transparent objects.

1. Introduction

"Fiat lux".[1] Light has always fascinated mankind. It is not only the inherent centre of attention for many of the greatest scientific discoveries in the last century, but also plays a crucial role for society and even sets the basis for religions. Typical light sensors used in computer vision either send or receive pulses and waves for which the wavelength and energy are measured to retrieve colour and intensity within a specified spectrum. However, intensity and wavelength are not the only properties of an electromagnetic (EM) wave. The oscillation direction of the EM-field relative to the light ray defines its polarisation. Most natural light sources such as the sun, a lamp or a candle emit unpolarised light, which means that the light wave oscillates in a multitude of directions. When such a wave is reflected off an object, light becomes either perfectly or partially polarised. Polarisation therefore carries information on surface structure, material and reflection angle which can complement passively retrieved texture information from a scene [30]. These additional measurements can be particularly interesting for photometrically challenging objects with metallic, reflective or transparent materials which all pose challenges to vision pipelines effectively hampering their use for automation.

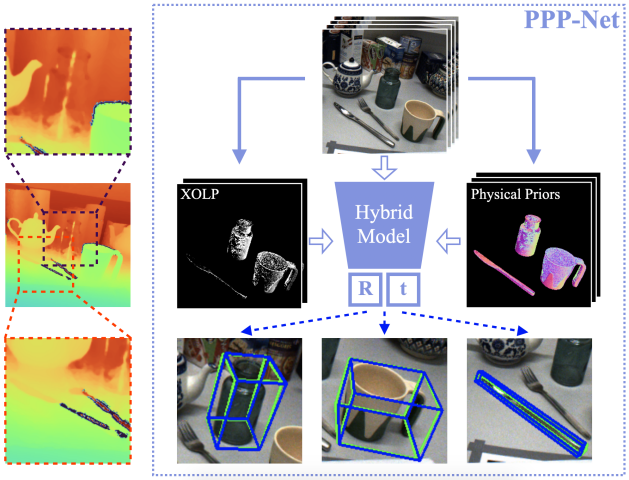

Figure 1. **PPP-Net.** Orthogonal to colour and depth images (left), polarisation data provides cues to surface normals especially for highly reflective (cutlery) and translucent (glass bottle) objects. Our **P**olarimetric **P**ose **P**rediction Pipeline (right) leverages the input of an RGBP camera and uniquely combines physical surface cues from polarisation properties with a data-driven approach to estimate accurate poses even for challenging objects which cannot be accurately predicted by current state-of-the-art approaches based on RGB and RGB-D.

While robust pipelines [23, 41, 10, 13] have been designed for the task of 6D pose estimation and textureless [25, 14] objects have been successfully predicted, photometrically challenging objects with reflectance and partial transparency have become the focus of research only very recently [39]. These objects pose challenges to RGB-D sensing and the field still lacks methods to cope with these problems. We move beyond previous methods based on

* *Equal contribution; Alphabetical order*
[1] Latin for "let there be light".

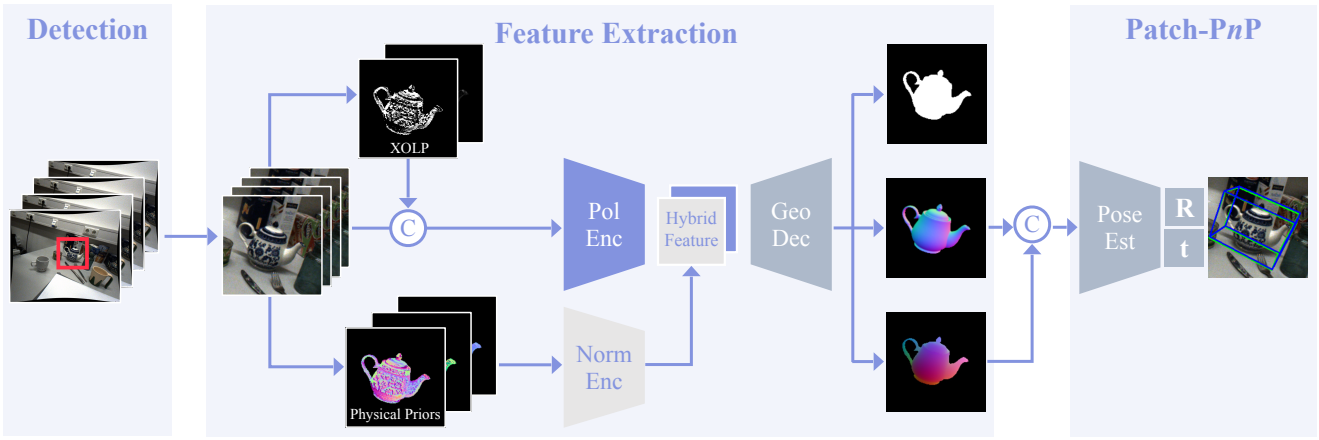

Figure 2. **PPP-Net Pipeline Overview.** After the initial detection of the object of interest, the RGBP image - a quadruple of four differently polarised RGB images - is utilised to compute AOLP/DOLP and polarised normal maps through our physical model. The polarised information and the physical cues are individually encoded and fused in our hybrid model. The decoder predicts an object mask, normal map and NOCS, and finally the 6D object pose is predicted by Patch-PnP [53].

light intensity and exploit the polarisation property of light as an additional prior for surface normals. This allows us to build a hybrid method combining a physical model with a data-driven learning approach to facilitate 6D pose estimation. We show that this not only facilitates pose estimation for photometrically challenging objects, but also improves the pose accuracy for classical objects. To this end, our core contributions are:

1. We propose **polarisation** as a new modality **for object pose estimation** and explore its advantages over previous modalities
2. We design a **hybrid pipeline** for pose estimation that leverages polarisation cues through a **combination of physical model cues with learning**.
3. As a result, we propose the first solution to estimate **6D poses for photometrically challenging objects with high reflectance and translucency** using polarisation.

2. Related Work

2.1. Polarimetric Imaging

Polarisation for 2D. Polarisation cues provide complementary information useful for various tasks in 2D computer vision that involve photometrically challenging objects. This has inspired a series of works on semantic [58] and instance [30] segmentation for reflective and transparent objects. The absence of strong glare behind specific polarisation filters further helps to remove reflections from images [36]. While one polarisation camera can already provide significant improvements compared to photometric acquisition setups, the use of multispectral polarimetric light fields [28] boosts the performance even more.

Polarisation for 3D. Due to the inherent connection of polarisation with surface shape and texture, the natural field of application seems to be 3D computer vision. Indeed, previous works on shape from polarisation (SfP) investigate the estimation of surface normals and depth from polarimetric data. However, intrinsic model ambiguities constraint setups in early works. Classical methods leverage an orthographic camera model and restrict the investigations to lab scenarios with very controlled environment conditions [18, 4, 56, 48]. Yu et al. [56] mathematically connect polarisation intensity with surface height and optimise for depth in a controlled scenario, while Atkinson et al. [4] recover surface orientation for fully diffuse surfaces. Others [48] add shape from shading principles or investigate the normal estimation using circular polarised light [18]. While these methods rely on monocular polarisation, more than one view can be combined with physical models for SfP [3, 11]. Some works also explore the use of complementary photometric stereo [2] and hybrid RGB+P approaches [61] which complement each other and allow for metrically accurate depth estimates if the light direction is known. If an initial depth map (e.g. from RGB-D) exists, polarimetric cues can further refine the measurements [29]. Furthermore, the polarimetric sensing model help estimate the relative transformation of a moving polarisation sensor [12] assuming the scene is fully diffuse. Data-driven approaches can mitigate any assumptions on surface properties, light direction and object shapes. Ba et al. [5] estimate surface normals by presenting a set of plausible cues to a neural network which can use these ambiguous cues for SfP. We take inspiration from this approach to complement our pose estimation pipeline with physical priors. In contrast to these works, we are interested in the object poses in an unconstrained setup without further assumption on the

reflection properties or lighting. The insights of previous works enable, for the first time, the design of a pipeline to address pose prediction for photometrically challenging objects made of transparent and highly reflective materials.

2.2. 6D Pose Prediction

Monocular RGB. Methods that predict 6D pose from a single image can be separated into three main categories: the ones that directly optimise for the pose, learn a pose embedding or establish correspondences between the 3D model and the 2D image. Works that leverage pose parameterisation either directly regress the 6D pose [55, 37, 41, 35] or discretise the regression task and solve for classification [32, 10]. Networks trained this way directly predict pose parameters in the form of $SE(3)$ elements given the parameterisation used for training. Pose parameterisation can also be implicitly learned [60]. The second branch of methods [54, 51, 50] utilises this to learn an implicit space to encode the pose from which the predictions can be decoded. Latest and also the currently best-performing methods follow a two-stage approach. A network is used to predict 2D-3D correspondences between image and 3D model which are used by a consecutive RANSAC/PnP pipeline that optimises the displacement robustly. Some methods in this field use sparse correspondences [45, 43, 49, 27] while others establish dense 2D-3D pairs [57, 42, 38, 24]. While these methods typically learn the correspondences alone, some works managed to learn the task end-to-end [26, 53, 13]. Inspired by the success of this, we also structurally follow the design of GDR-Net [53].

RGB-D and Refinement. Since the task of monocular pose estimation from RGB is an inherently ill-posed problem, depth maps serve as a geometrical rescue. The spatial cue given by the depth map can be leveraged to establish point pairs for pose estimation [16] which can be further improved with RGB [7]. In general, pose can be recovered from depth or combined RGB-D and most RGB-only methods (e.g. [51, 38, 42, 35]) benefit from a depth-driven refinement using ICP [6] or from indirect multi-view cues [35]. The complementary information of RGB and depth has also inspired the seminal work DenseFusion [52] in which deeply encoded features from both modalities are fused. FFB6D [20] further improves this through a tight coupling strategy with cross-modal information exchanges in multiple feature layers combined with a keypoint extraction [21] that leverages geometry and texture cues. These works however, crucially depend on input quality and depth sensing suffers in photometrically challenging regions, where polarisation cues for depth could expedite the pose prediction. However, to the best of our knowledge, this has not been proposed, yet.

Photometric Challenges. The field of 6D pose estimation usually tests on a set of well established dataset with RGB-D input [23, 8, 55, 31]. Photometrically challenging objects such as texture-less and reflective industrial parts are also part of publically available dataset [25, 15]. While most of these datasets are carefully annotated for the pose, polarisation input is not available. Transparency is a further challenge which has been addressed already in the pioneering work of Saxena et al. [47] where the robotic grasp point of objects is determined from RGB stereo without a 3D model. Philipps et al. [44] demonstrate how transparent object with rotation symmetry can be reconstructed from two views using an edge detector and contour fitting and more recently, KeyPose [40] investigates instance and category level pose prediction from RGB stereo. Since their depth sensor fails on transparent objects, they leverage an opaque-transparent object pair to establish ground truth depth. ClearGrasp [46] constitutes an RGB-D method that can be used on transparent objects. More recently, Liu et al. [39] presented the extensive StereOBJ-1M dataset. It includes transparent, reflective and translucent objects with variations in illumination and symmetry using a binocular stereo RGB camera for pose estimation. However, none of these dataset comprised RGBP data.

To this end, the next natural step connects the shape cues from polarisation to recover object geometry in challenging environments. We further ask the question how to do so by starting with a look into polarimetric image formation.

3. Polarimetric Pose Prediction

In contrast to RGBP sensors (see Fig. 3), RGB-D sensors enjoy a wide use in the pose estimation field. Their cost-efficiency and tight integration in many devices present a lot of possibilities in the vision field, but their design also comes with a few drawbacks.

3.1. Photometric Challenges for RGB-D

Commercial depth sensors typically use active illumination either by projecting a pattern (e.g. intel RealSense

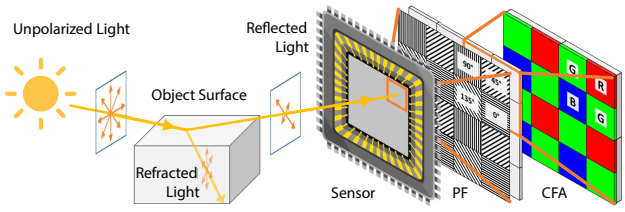

Figure 3. **Polarisation Camera.** Light from an unpolarised light source reflects on an object surface. The refracted and reflected part are partially polarised. A polarisation sensor captures the light. In front of every pixel there are four polarisation filters (PF) arranged at different angles ($0°$, $45°$, $90°$, $135°$). The colour filter array (CFA) separates lights into different wavebands.

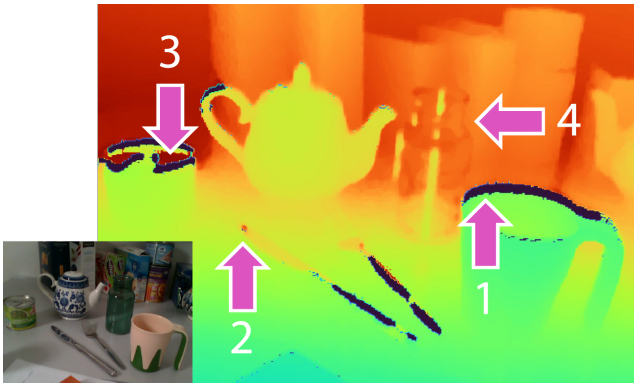

Figure 4. **Depth Artifacts.** A depth sensor (RealSense L515) miscalculates depth values for typical household objects. Reflective boundaries (1,3) invalidate pixels while strong reflections (2,3) lead to incorrect values too far away. Semi-transparent objects such as the vase (4) becomes partly invisible for the depth sensor which measures the distance to the objects behind.

D series) or using time-of-flight (ToF) measurements (e.g. Kinect v2 / Azure Kinect, intel RealSense L series). While the former triangulate depth using stereo vision principles on projected or scene textures, the latter measures the roundtrip time of a light pulse that reflects from the scene. Since the measurement principle is photometric, both suffer on photometrically challenging surfaces where reflections artificially extend the roundtrip time of photons and translucent objects deteriorate the projected pattern to an extent that makes depth estimation infeasible. Fig 4 illustrates such an example for a set of common household objects. The semi-transparent vase becomes almost invisible for the used ToF sensor (RealSense L515) which measures the distance to the objects behind. The reflections on both cutlery and can lead to incorrect depth estimates significantly further than the correct value while strong reflections at boundaries invalidate pixel distances.

3.2. Surface Normals from Polarisation

Before working with RGBP data, we introduce some of the physics behind polarimetric imaging. Natural light and most artificially emitted light is unpolarised, meaning that the electromagnetic wave oscillates along all planes perpendicular to the direction of propagation of light [17]. When unpolarised light passes through a linear polariser or is reflected at Brewster's angle from a surface, it becomes perfectly polarised. How fast light travels through the material, how much of it is reflected is determined by the *refractive index*. It also determines the Brewster's angle of that medium. When light is reflected at the same angle to the surface normal as the incident ray, we speak of *specular reflection*. The remaining part penetrates the object as refracted light. As the light wave traverses through the medium, it becomes partially polarised. Following this, it

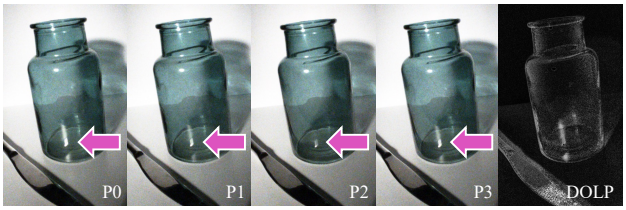

Figure 5. **DOLP.** Polarisation changes for reflection of diffuse light on a translucent surface. Note the indicated differences in the polarimetric image quadruplet that directly relate to the surface normal. The degree of linear polarisation (DOLP) for the translucent and reflective surfaces are considerably higher than for the rest of the image.

escapes from the object and creates *diffuse reflection*. For all real physical objects, the resulting reflection is a combination of specular and diffuse reflection, where the ratio largely depends on the refractive index and the angle of incident light as exemplified in Fig. 5

Light reaches the sensor with a specific intensity I and wavelength λ. The colour filter array (CFA) of the sensor then separates the incoming light into RGB wavebands as illustrated in Fig. 3. The incoming light also has a degree of linear polarisation (DOLP) ρ and a direction (angle) of polarisation (AOLP) ϕ. The measured intensity behind a polariser with an angle $\varphi_{pol} \in \{0°, 45°, 90°, 135°\}$ depends on these parameters and the unpolarised intensity I_{un} [30]:

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

We find φ and ρ from the over-determined system of linear equations in 1 using linear least squares. Depending on the surface properties, AOLP is calculated as

$$\begin{cases} \phi_d[\pi] &= \alpha & \text{for diffuse reflection} \\ \phi_s[\pi] &= \alpha - \frac{\pi}{2} & \text{for specular reflection} \end{cases} \tag{2}$$

where $[\pi]$ indicates the π-ambiguity and α is the azimuth angle of the surface normal $\mathbf{n}$. We can further relate the viewing angle $\theta \in [0, \pi/2]$ to the degree of polarisation by considering Fresnel coefficients, thus DOLP is similarly given by [4]

$$\begin{cases} \rho_d = \frac{(\eta - 1/\eta)^2 \sin^2(\theta)}{2 + 2\eta^2 - (\eta + 1/\eta)^2 \sin^2(\theta) + 4\cos(\theta)\sqrt{\eta^2 - \sin^2(\theta)}} \\ \rho_s = \frac{2\sin^2(\theta)\cos(\theta)\sqrt{\eta^2 - \sin^2(\theta)}}{\eta^2 - \sin^2(\theta) - \eta^2 \sin^2(\theta) + 2\sin^4(\theta)} \end{cases} \tag{3}$$

with the refractive index of the observed object material η. Solving equation 3 for θ, we retrieve three solutions $\theta_d, \theta_{s1}, \theta_{s2}$, one for the diffuse case and two for the specular case. For each of the cases, we can now find the 3D orientation of the surface by calculating the surface normals:

$$\mathbf{n} = (\cos\alpha \sin\theta, \sin\alpha \sin\theta, \cos\theta)^{\mathrm{T}} \tag{4}$$

We use these plausible normals $\mathbf{n}_d, \mathbf{n}_{s1}, \mathbf{n}_{s2}$ as physical priors per pixel to guide our neural network to estimate the 6D object pose.

3.3. Hybrid Polarimetric Pose Prediction Model

In this section, we present our Polarimetric Pose Prediction Network, short **PPP-Net**. Given polarimetric images at four different angles $I_0, I_{45}, I_{90}, I_{135}$, together with the calculated AOLP ϕ, DOLP ρ, and normal maps N_d, N_{s1}, N_{s2} as physical priors, we aim to utilise a network to learn the pose $\mathbf{P} = [\mathbf{R}|\mathbf{t}]$ that can transform the target object from the object frame to the camera frame given the 3D CAD model of the object.

Network Architecture. Our network architecture is depicted in Fig. 2. The network has two encoders, which take joint polarisation information from the native polarimetric images and the calculated AOLP/DOLP maps as well as the physical normals as priors with zoomed-in ROI of size 256×256 as inputs separately. The decoder takes the combined encoded information from both encoders, together with skip connections from different hierarchical levels of the encoders, to decode the object mask, normal map, and a 3-channel dense correspondence map (NOCS) which maps each pixel to its corresponding normalised 3D coordinate. The predicted normal map together with the dense correspondence map are consecutively fed into a pose estimator as used in GDR-Net [53]. The pose estimator is composed of convolution layers and fully connected layers, to output the final estimated 3D rotation and translation.

Pose Parametrisation. Inspired by recent works [60, 38, 53], we parameterise our rotation as allocentric continuous 6-dimensional representation, and translation as scale-invariant representation [38, 53, 13].
The continuous 6-dimensional representation $\mathbf{R}_{6d}$ for rotation comes from the first two columns of the original rotation matrix $\mathbf{R}$ [60], and we further turn it into allocentric representation [53, 13], since our network only perceives the ROI of the target object, which favors the viewpoint-independant representation.

The zoomed-in ROI can help the network focus on more relevant information in the image, i.e. our target object. To overcome the limitations of direct translation vector regression, we estimate the scale-invariant translation composed of relative differences between projected object centroids and the detected bounding box center location with respect to the bounding box size. The latter is given by δ_x, δ_y and the relative zoomed-in depth, δ_z, where

$$\begin{cases} \delta_x = (o_x - b_x)/b_w \\ \delta_y = (o_y - b_y)/b_h \\ \delta_z = t_z/r \end{cases} \tag{5}$$

with (o_x, o_y) and (b_x, b_y) being the projected object centroids and bounding box center coordinates. The size of the bounding box (b_w, b_h) is also used for calculating the zoomed-in ratio $r = s_{out}/s_{in}$ where $s_{in} = \max(b_w, b_h)$ and s_{out} is the size of the output. Note that we can recover both the rotaion matrix and translation vector with known camera intrinsics K [34, 38].

Object Normal Map. The surface normal map contains the surface orientation at each discrete pixel coordinate and thus ecodes the shape of the object. Inspired by the previous works in SfP, we also aim to retrieve the surface normal map in a data-driven manner [5]. To better encode the geometric cue from the input physical priors apart from the polarisation cue, we do not concatenate the physical normals with the polarised images as suggested by Ba et al. [5], but encode them separately into two ResNet encoders. The decoder then learns to produce object shape encoded by surface normal map. The estimated normals are L2-normalised to unit length. As shown in Tab. 1, with the given physical normals as shape prior, we can achieve high quality normal map prediction.

Dense Correspondence Map. The dense correspondence map stores the normalised 3D object coordinates given associated poses. This explicitly models correspondences between object 3D coordinates and projected 2D pixel locations. As shown by Wang et al. [53], this representation helps the consecutive differentiable pose estimator to achieve high accuracy in comparison with RANSAC/PnP.

3.4. Learning Objectives

The overall objective is composed of both geometrical features learning and the pose optimisation [53] as:

$$\mathcal{L} = \mathcal{L}_{pose} + \mathcal{L}_{geo}, \tag{6}$$

with

$$\mathcal{L}_{pose} = \mathcal{L}_R + \mathcal{L}_{center} + \mathcal{L}_z \tag{7}$$

$$\mathcal{L}_{geo} = \mathcal{L}_{mask} + \mathcal{L}_{normals} + \mathcal{L}_{xyz}. \tag{8}$$

Specifically, we employ separate loss terms for given ground truth rotation $\mathbf{R}$, (δ_x, δ_y) and δ_z as

$$\begin{cases} \mathcal{L}_R & = \underset{\mathbf{x} \in \mathcal{M}}{\text{avg}} \|\mathbf{R}\mathbf{x} - \hat{\mathbf{R}}\mathbf{x}\|_1 \\ \mathcal{L}_{center} & = \|(\delta_x - \hat{\delta}_x, \delta_y - \hat{\delta}_y)\|_1 \\ \mathcal{L}_z & = \|\delta_z - \hat{\delta}_z\|_1 \end{cases} \tag{9}$$

where $\hat{\bullet}$ denotes prediction. For symmetrical objects, the rotation loss will be calculated based on the smallest loss from all possible ground-truth rotations under symmetry.

To learn the intermediate geometrical features, we employ $L1$ losses for mask and dense correspondences map

learning, and cosine similarity loss for normal estimation:

$$\begin{cases} \mathcal{L}_{mask} & = \|\mathbf{M} - \hat{\mathbf{M}}\|_1 \\ \mathcal{L}_{xyz} & = \mathbf{M} \odot \|\mathbf{M}_{xyz} - \hat{\mathbf{M}}_{xyz}\|_1 \\ \mathcal{L}_{normal} & = 1 - \langle \mathbf{n}, \hat{\mathbf{n}} \rangle \end{cases} \tag{10}$$

where $\odot$ indicates the Hadamard product of element-wise multiplication, and $\langle \bullet, \bullet \rangle$ denotes the dot product.

4. Experimental Results

The motivation of our proposed pipeline is to show the advantage of leveraging pixelwise physical priors from polarised light (a.k.a. RGBP) for accurate 6D pose estimation of photometrically challenging objects - for which RGB-only and RGB-D methods often fail. For this purpose, we train and test **PPP-Net** with different modalities first on two exemplary objects with very different level of photometric complexity, i.e. a plastic *cup*, and a photometrically very challenging, reflective and textureless stainless steel cutlery *knife*. As detailed later, we find that polarimetric information yields significant performance gain for photometrically challenging objects.

4.1. Polarimetric Data Acquisition

To evaluate our pipeline we leverage 6 models from the PhoCal [1] category-level pose estimation dataset which comprises 60 household objects with high-quality 3D models scanned by a structured light 3D stereo scanner (EinScan-SP 3D Scanner, SHINING 3D Tech. Co., Ltd., Hangzhou, China). The scanning accuracy of the device is ≤ 0.05 mm which allows for highly accurate models. We select the models *cup*, *teapot*, *can*, *fork*, *knife*, *bottle* with increasing photometric complexity which we illustrate in Fig. 6. The last three models do not include texture due to their surface structure. The 3D scanning has been done with a vanishing 3D scanning spray that made the surface temporarily opaque. To acquire RGB-D images, we use a direct Time-of-Flight (dToF) camera, intel RealSense LiDAR Camera L515 (intel, Santa Clara, California, USA), which captures RGB and Depth data at 640x480 pixel resolution.

RGBP is acquired using the polarisation camera Phoenix 5.0 MP PHX050S1-QC comprising a Sony IMX264MYR CMOS (Color) Polarsens sensor (LUCID Vision Labs, Inc., Richmond B.C, Canada) through a Universe Compact C-Mount 5MP 2/3" 6mm f/2.0 lens (Universe, New York, USA) at 612x512 pixel resolution. Both cameras are mounted jointly to a KUKA iiwa (KUKA Roboter GmbH, Augsburg, Germany) 7 DoF robotic arm that guarantees a positional reproducibility of ± 0.1 mm. Intrinsic and extrinsic calibration is performed following the standard pinhole camera model [59] with five distortion coefficients [22]. For pose annotation, we leverage the mechanical pose annotation method proposed in PhoCal [1] where the robotic manipulator is used to tip the object of interest and extract a point cloud. This point cloud is consecutively aligned to the 3D model using ICP [6] to allow for highly accurate pose labels even for photometrically challenging objects. We plan a robot trajectory and use this setup to acquire four scenes with four different trajectories each and utilise a total of 8740 image sets for the dataset.

4.2. Experiments Setup

Implementation Details. We initially refine an off-the-shelf detector Mask RCNN [19] directly on the polarised images I_0 to provide useful object crops on our data (as is needed for the RGB-only benchmark and ours). We follow similar training/testing split strategy as commonly used for the public datasets [9], and employ $\approx 10\%$ of the RGBP images for training and 90% for testing. We train our network end-to-end with Adam optimiser [33] for 200 epochs. The initial learning rate is set to 1e-4, which is halved every 50 epochs. As the depth sensor has a different field of view and is placed beneath the polarisation camera on a customised camera rig, the RGB-D benchmark split differs from the RGB training/testing split.

Evaluation Metrics. To establish our proposed novel 6D pose estimation approach, we report the pose estimation accuracy per object as the commonly used average distance (ADD) and its equivalent for symmetrical objects (ADD-S) [23] for different benchmarks. For the surface normal estimation, we calculate the mean and median errors (in degrees) and the percentage of pixels where the estimated normals vary less than $11.25°$, $22.5°$ and $30°$ from the ground truth. We additionally give valuable insights into our proposed pipeline by performing detailed ablations on the input modalities, the fusion of complementary modalities, and the effect of explicit learning of physically plausible geometric information and its effect on pose prediction accuracy (see Tab. 1), and discuss limitations of our proposed approach.

4.3. PPP-Net

Here, we perform a series of experiments to study the influence of the input modality on the pose estimation accuracy (compare Tab. 1), where we specifically analyse the influence of polarimetric image information for the task of 6D object pose estimation. We demonstrate that our network with RGBP input performs at the state-of-the-art level

Figure 6. **3D Models.** Test objects with increasing photometric complexity (left to right). Three objects have no texture in as they are reflective (cutlery) or transparent (bottle).

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
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**
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**

Table 1. **PPP-Net Modalities Evaluation.** Different combinations of input and output modalities are used for training to study their influence on pose estimation accuracy ADD for objects with different photometric complexity. Where applicable, metrics for estimated normals are reported as well. Results for other objects in Supplementary Material.

for non-reflective, textured objects, which we define as less photometrically challenging, e.g. plastic *cup*, and outperforms current models for photometrically complex objects, e.g. stainless steel cutlery.

To identify the direct influence of polarisation imaging for the task of accurate object pose estimation, we first establish an RGB-only baseline by neglecting our contributions of **PPP-Net**. To compute the unpolarised RGB image, we average over polarimetric images at complementary angles and use this as input for RGB-only. As shown in the first two rows in Tab. 1 for each object (RGB against Polar RGB), the polarisation modality yields larger accuracy gains for the photometrically challenging object *knife* as compared to *cup*. Auxiliary network predictions for normals and NOCS marginally enhance the performance as the network is encouraged to explicitly encode this information from the input modalities. The physically-induced normals from polarisation images provide orthogonal information that significantly boosts the pose prediction quality and thus achieves best ADD performance across all experiments. This behaviour is most promiment for the photometrically challenging *knife*.

4.4. Comparison with established Benchmarks

The input modality experiments already demonstrate the strong capabilities of polarimetric imaging inputs for **PPP-Net** to successfully learn reliable 6D pose prediction with high accuracy for photometrically challenging objects. The depth map of an RGB-D sensor can also provide geometric information that can be utilised for the task of 6D object pose estimation. FFB6D [20] is currently the best-performing state-of-the-art learning pipeline which combines RGB and geometric information from depth maps. Hence, the design of FFB6D is motivated by similar principles as our proposed method, since it leverages geometric information for the task of 6D pose estimation, and is therefore chosen as a strong geometric benchmark for comparison. The unique Full-Flow-Bidirectional fusion network [20] of FFB6D learns to combine appearance and depth information as well as local and global information from the two individual modalities.

We train FFB6D on our data for each object individually and report the best ADD(-S) metric for all objects in Tab. 2. The photometric challenge that each object constitutes is summarised in the Tab. 2 and detailed by its properties (compare with Fig. 6). The objects are categorised into three classes based on the depth map quality for the depth sensor (compare also Fig. 4). We can observe that objects with good depth maps and minor photometric challenges achieve high ADD values for FFB6D. For challenging objects, the increase in photometric complexity (and lower depth map quality) correlates with a decrease in ADD. The transparent *Bottle* object is an exception to this pattern. The depth map is completely invalid (compare Fig. 4), but FFB6D still achieves high ADD. Our hypothesis is that the network successfully learns to ignore the depth map input from early training onward (see Sec. 5 for details). **PPP-Net** achieves comparable results for easy objects and outperforms the strong benchmark for photometrically complex objects. Our method does not suffer from reduced ADD due to noisy or inaccurate depth maps but rather leverages the orthogonal surface information from RGBP data.

As **PPP-Net** profits vastly from physical priors from polarisation, we thoroughly investigate to which extent this additional information impacts the improvement of estimated poses, especially for photometrically challenging objects, by comparing the results also against the monocular RGB-only method GDR-Net [53]. We observe that while using polarimetric information slightly improves pose estimation accuracy for non-challenging objects, we can achieve superior performance for items with inconsistent photometric information due to reflection or transparency. In Tab. 2 the accuracy gain of **PPP-Net** against GDR-Net increases proportionally to the photometric complexity, since our physical priors provide additional information about the geometry of an object.

5. Discussion

Limitations of current geometric methods. As mentioned earlier, we postulate that the RGB-D method ignores

Object	Photo. Chall.	Properties					Depth Quality	RGB-D Split		RGB Split	
		Reflective	Metallic	Textureless	Transparent	Symmetric		FFB6D	**Ours**	GDR	**Ours**
Cup							(+)	**99.4**	98.1	96.7	**97.2**
Teapot	†	(*)					++	86.8	**94.2**	99.0	**99.9**
Can	†	*	*				–	80.4	**99.7**	96.5	**98.4**
Fork	††	*	*	*			––	37.0	**72.4**	86.6	**95.9**
Knife	††	*	*	*			–––	36.7	**87.2**	92.6	**96.4**
Bottle	† † †	*		*	*	*	None	61.5	**93.6**	94.4	**97.5**
Mean								67.0	**90.9**	94.3	**97.6**

Table 2. **Benchmark comparisons.** We compare our method against recent RGB-D (FFB6D [20]) and RGB-only (GDR-Net [53]) methods on a variety of objects with different level photometric challenges (†), and depth map quality (good: + to low:−) which serves as input for FFB6D. RGB-D and RGB-only comparisons are trained and tested on different splits due to different field of view of depth camera (see Sec. 4 for details). We report the Average Recall of ADD(-S).

invalid depth data already in early stages of training (e.g. for the transparent *bottle)* and eventually learns to also ignore noisy or corrupted depth information. To prove this assumption, we perform adversarial attacks on the input depth map for the FFB6D [20] encoder to analyse which parts of input modalities the network relies on when making a prediction. For this purpose we add small Gaussian noise on the depth-related feature embedding in the bottleneck of the network and compare the ADD under this attack. We purposely "overfit" the model on objects of different photometric complexity and compute the relative decrease in ADD under the attack. We observe that the relative decrease is smaller for photometrically challenging objects as compared to objects with accurate depth maps (27% drop in ADD for *knife* and 63% for *cup*). These findings suggest that the network indeed relies on the RGB information only.

Benefits of Polarisation. We have shown that physical priors from polarised light can significantly improve 6D pose estimation results for photometrically challenging objects. RGB-only methods do not incorporate any geometric information and therefore show worse results in scenarios with reflective surfaces or objects of little texture. Methods which try to leverage geometric priors from RGB-D [20], often cannot reliably recover the 6D pose of such objects as the depth map is usually degenerated and corrupt. Our **PPP-Net**, as the first RGBP 6D object pose estimation method, successfully achieves to learn accurate poses even for very challenging objects by extracting geometric information from physical priors. Qualitative results are shown in Figs. 1, 2 and 7, and additionally in the supplementary material. Another benefit of using RGBP lies in the sensor itself: as the polarisation filter is directly integrated on the same sensor as the Bayer filter, both modalities are intrinsically calibrated and the image can be acquired passively paving the way to sensor-integration on low-energy and mobile devices. RGB-D cameras, on the contrary, often require energy-costly active illumination and extrinsic calibration, which prevents simple integration and introduces additional uncertainty to the final RGB-D image.

Limitations. Our physical model requires the refractive index of the respective object to reliably compute the physical priors. To explore the potential of the physical model, distinct to prior works [48, 5] which fix the refractive index to $\eta = 1.5$ for all experiments, we use physically plausible values according to the materials.[2] This means one would need to manually choose such parameter, which limits the performance of the physical model when encountering objects with unknown composite materials. Moreover, strong changes in texture also affect the reflection of light and thus DOLP calculation which, in turn, influences our physical priors.

6. Conclusion

We have presented **PPP-Net**, the first learning-based 6D object pose estimation pipeline which leverages geometric information from polarisation images through physical cues. Our method outperforms current state-of-the-art RGB-D and RGB methods for photometrically challenging objects and demonstrates at par performance for ordinary objects. Extensive ablations show the importance of the complementary polarisation information for accurate pose estimation - specifically for objects without texture, reflective surfaces or transparency.

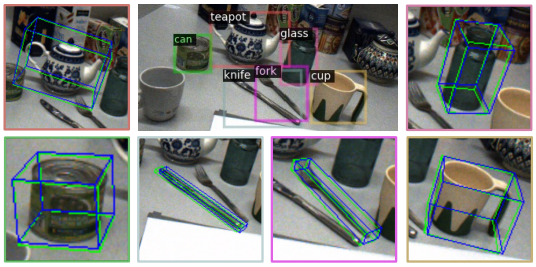

Figure 7. **Qualitative Results.** Input image with 2D detections are shown. Predicted and GT 6D poses are illustrated by *blue* and *green* bounding boxes, respectively.

[2] We approximate the refractive index by the look-up table provided by https://refractiveindex.info/

A. 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 serves 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:

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

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.

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

Table A1. **Refractive Indices.**

B. Additional Results

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. 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 object under consideration in the dataset.

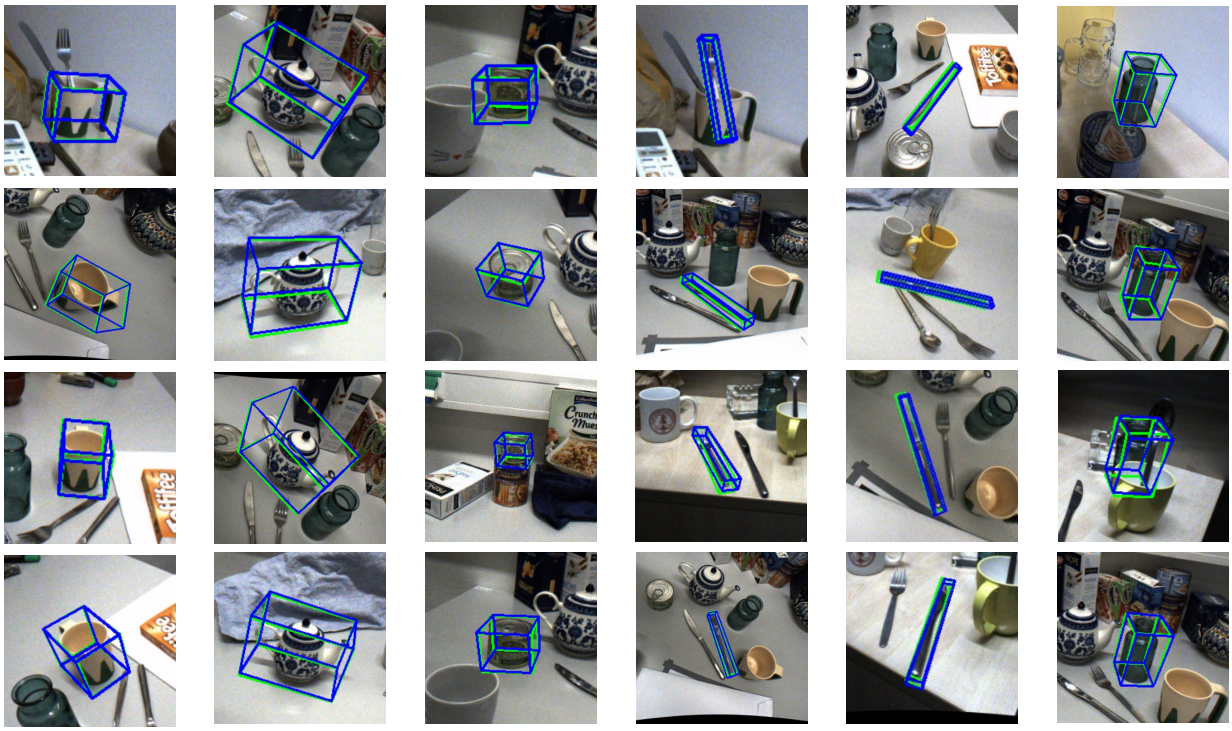

Figure A1. **Qualitative Results.** Predicted and GT 6D poses are illustrated by *blue* and *green* bounding boxes, respectively.

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)
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**
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 A2. **PPP-Net 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.

References

[1] Anonymous. Phocal: A multimodal dataset for category-level object pose estimation with photometrically challenging objects. In *Under Submission*, 2021. 6

[2] Gary A Atkinson. Polarisation photometric stereo. *Computer Vision and Image Understanding*, 160:158–167, 2017. 2

[3] Gary A Atkinson and Edwin R Hancock. Multi-view surface reconstruction using polarization. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 309–316. IEEE, 2005. 2

[4] Gary A Atkinson and Edwin R Hancock. Recovery of surface orientation from diffuse polarization. *IEEE transactions on image processing*, 15(6):1653–1664, 2006. 2, 4

[5] Yunhao Ba, Alex Gilbert, Franklin Wang, Jinfa Yang, Rui Chen, Yiqin Wang, Lei Yan, Boxin Shi, and Achuta Kadambi. Deep shape from polarization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 554–571. Springer, 2020. 2, 5, 8

[6] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992. 3, 6

[7] Tolga Birdal and Slobodan Ilic. Point pair features based object detection and pose estimation revisited. In *2015 International Conference on 3D Vision*, pages 527–535. IEEE, 2015. 3

[8] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *European conference on computer vision*, pages 536–551. Springer, 2014. 3

[9] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, et al. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3364–3372, 2016. 6

[10] 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. 1, 3

[11] Zhaopeng Cui, Jinwei Gu, Boxin Shi, Ping Tan, and Jan Kautz. Polarimetric multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1558–1567, 2017. 2

[12] Zhaopeng Cui, Viktor Larsson, and Marc Pollefeys. Polarimetric relative pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2671–2680, 2019. 2

[13] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, and Federico Tombari. So-pose: Exploiting self-occlusion for direct 6d pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12396–12405, 2021. 1, 3, 5

[14] Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Hartinger, and Carsten Steger. Introducing mvtec itodd - a dataset for 3d object recognition in industry. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017. 1

[15] Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Hartinger, and Carsten Steger. Introducing mvtec itodd-a dataset for 3d object recognition in industry. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2200–2208, 2017. 3

[16] Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 998–1005. Ieee, 2010. 3

[17] Torsten Fließbach. *Elektrodynamik: Lehrbuch zur Theoretischen Physik II*, volume 2. Springer-Verlag, 2012. 4

[18] N Missael Garcia, Ignacio De Erausquin, Christopher Edmiston, and Viktor Gruev. Surface normal reconstruction using circularly polarized light. *Optics express*, 23(11):14391–14406, 2015. 2

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[20] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 3, 7, 8

[21] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[22] Janne Heikkila and Olli Silvén. A four-step camera calibration procedure with implicit image correction. In *Proceedings of IEEE computer society conference on computer vision and pattern recognition*, pages 1106–1112. IEEE, 1997. 6

[23] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision*, pages 548–562. Springer, 2012. 1, 3, 6

[24] Tomas Hodan, Daniel Barath, and Jiri Matas. Epos: Estimating 6d pose of objects with symmetries. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11703–11712, 2020. 3

[25] Tomáš Hodan, Pavel Haluza, Štepán Obdržálek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 880–888. IEEE, 2017. 1, 3

[26] Yinlin Hu, Pascal Fua, Wei Wang, and Mathieu Salzmann. Single-stage 6d object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2930–2939, 2020. 3

[27] Yinlin Hu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. Segmentation-driven 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3385–3394, 2019. 3

[28] Md Nazrul Islam, Murat Tahtali, and Mark Pickering. Specular reflection detection and inpainting in transparent object through msplfi. *Remote Sensing*, 13(3):455, 2021. 2

[29] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. Depth sensing using geometrically constrained polarization normals. *International Journal of Computer Vision*, 125(1-3):34–51, 2017. 2

[30] 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. 1, 2, 4

[31] Roman Kaskman, Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects. *International Conference on Computer Vision (ICCV) Workshops*, 2019. 3

[32] 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 international conference on computer vision*, pages 1521–1529, 2017. 3

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[34] Abhijit Kundu, Yin Li, and James M Rehg. 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3559–3568, 2018. 5

[35] 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. 3

[36] Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, and Qifeng Chen. Polarized reflection removal with perfect alignment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1750–1758, 2020. 2

[37] 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. 3

[38] Zhigang Li, Gu Wang, and Xiangyang Ji. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7678–7687, 2019. 3, 5

[39] Xingyu Liu, Shun Iwase, and Kris M Kitani. Stereobj-1m: Large-scale stereo image dataset for 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10870–10879, 2021. 1, 3

[40] Xingyu Liu, Rico Jonschkowski, Anelia Angelova, and Kurt Konolige. Keypose: Multi-view 3d labeling and keypoint estimation for transparent objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11602–11610, 2020. 3

[41] Fabian Manhardt, Diego Martin Arroyo, Christian Rupprecht, Benjamin Busam, Tolga Birdal, Nassir Navab, and Federico Tombari. Explaining the ambiguity of object detection and 6d pose from visual data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6841–6850, 2019. 1, 3

[42] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7668–7677, 2019. 3

[43] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4561–4570, 2019. 3

[44] Cody J Phillips, Matthieu Lecce, and Kostas Daniilidis. Seeing glassware: from edge detection to pose estimation and shape recovery. In *Robotics: Science and Systems*, volume 3, 2016. 3

[45] Mahdi Rad and Vincent Lepetit. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3828–3836, 2017. 3

[46] Shreeyak Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. Clear grasp: 3d shape estimation of transparent objects for manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3634–3642. IEEE, 2020. 3

[47] Ashutosh Saxena, Justin Driemeyer, and Andrew Y Ng. Robotic grasping of novel objects using vision. *The International Journal of Robotics Research*, 27(2):157–173, 2008. 3

[48] William AP Smith, Ravi Ramamoorthi, and Silvia Tozza. Height-from-polarisation with unknown lighting or albedo. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2875–2888, 2018. 2, 8

[49] Chen Song, Jiaru Song, and Qixing Huang. Hybridpose: 6d object pose estimation under hybrid representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 431–440, 2020. 3

[50] Martin Sundermeyer, Maximilian Durner, En Yen Puang, Zoltan-Csaba Marton, Narunas Vaskevicius, Kai O Arras, and Rudolph Triebel. Multi-path learning for object pose estimation across domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13916–13925, 2020. 3

[51] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 699–715, 2018. 3

[52] 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. 3

[53] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16611–16621, 2021. 2, 3, 5, 7, 8

[54] Paul Wohlhart and Vincent Lepetit. Learning descriptors for object recognition and 3d pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3109–3118, 2015. 3

[55] 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. 3

[56] 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. 2

[57] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1941–1950, 2019. 3

[58] Yifei Zhang, Olivier Morel, Marc Blanchon, Ralph Seulin, Mojdeh Rastgoo, and Désiré Sidibé. Exploration of deep learning-based multimodal fusion for semantic road scene segmentation. In *VISIGRAPP (5: VISAPP)*, pages 336–343, 2019. 2

[59] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000. 6

[60] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 3, 5

[61] 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. 2