International Journal of Computer Vision
https://doi.org/10.1007/s11263-023-01965-w

S^{2}P^3: Self-Supervised Polarimetric Pose Prediction

Patrick Ruhkamp[1] · **Daoyi Gao**[1] · **Nassir Navab**[1] · **Benjamin Busam**[1]

Received: 3 April 2023 / Accepted: 28 November 2023

Abstract
This paper proposes the first self-supervised 6D object pose prediction from multimodal RGB + polarimetric images. The novel training paradigm comprises (1) a physical model to extract geometric information of polarized light, (2) a teacher–student knowledge distillation scheme and (3) a self-supervised loss formulation through differentiable rendering and an invertible physical constraint. Both networks leverage the physical properties of polarized light to learn robust geometric representations by encoding shape priors and polarization characteristics derived from our physical model. Geometric pseudo-labels from the teacher support the student network without the need for annotated real data. Dense appearance and geometric information of objects are obtained through a differentiable renderer with the predicted pose for self-supervised direct coupling. The student network additionally features our proposed invertible formulation of the physical shape priors that enables end-to-end self-supervised training through physical constraints of derived polarization characteristics compared against polarimetric input images. We specifically focus on photometrically challenging objects with texture-less or reflective surfaces and transparent materials for which the most prominent performance gain is reported.

Keywords Self-supervision · Multi-modalities · Pose estimation · Differentiable rendering

1 Introduction

"Fiat lux, et facta est lux".[1] Light has been the foundation of many significant scientific findings in history. Early horological devices utilized changing shadows cast from the sun to measure time throughout centuries across different civilizations all over the globe. Based on the constant speed of the electromagnetic wave (EM) with which light travels, it is possible to determine the distance of an object after emitting a light pulse by measuring its return time after reflection: a principle used in many active depth sensors. However, measurements are affected by artifacts such as multi-path interference (MPI) (Cui et al., 2010) due to reflective materials, ambient light (Jung et al., 2021), or inherently incorrect estimates when the light passes through transparent objects such as glass. This leads to inaccurate depth estimates, most noticeable for photometrically challenging objects (Jung et al., 2022). Still, many methods that learn geometric tasks from images use such geometry information from depth data.

6D object pose estimation is one of those geometric tasks and essential in many computer vision and AR applications, ranging from robotics (Wang et al., 2021c) to safety-critical autonomous driving (Ost et al., 2021) and medical applications (Busam et al., 2018). Recent methods integrate geometric information either directly as input (He et al., 2021) or leverage it for self-supervision (Wang et al., 2021a). Reliable geometric cues can improve pose estimation performance, while unreliable and noisy depth information would interfere with what information a neural network has learned to extract.

Recent approaches integrate the geometric information of polarized light by learning features from both the estimated normal from polarization, and their polarization characteristics, for the task of 6D object pose estimation in a supervised

[1] Latin for "let there be light, and there was light".

Patrick Ruhkamp and Daoyi Gao have contributed equally to this work.

✉ Patrick Ruhkamp
p.ruhkamp@tum.de

✉ Benjamin Busam
b.busam@tum.de

Daoyi Gao
daoyi.gao@tum.de

Nassir Navab
n.navab@tum.de

[1] TUM School of Computation, Information and Technology, Technical University of Munich, Munich, Germany

Published online: 12 January 2024

way (Gao et al., 2022). In the case of photometrically complex objects, it is shown that the deterioration of measured depth is even inferior to the use of this modality, ultimately making the direct geometric measurement obsolete. The authors report impressive results for texture-less, reflective and translucent objects, outperforming state-of-the-art RGB-only (Wang et al., 2021b) and RGB-D (He et al., 2021) methods. However, an extensive training dataset with ground-truth annotations is required, which may be challenging to obtain in practice, especially with high accuracy (Wang et al., 2022).

In $\mathbf{S}^2\mathbf{P}^3$, we study how a neural network can encode the geometric shape priors from polarized light captured with a multi-modal polarization camera for the task of 6D object pose estimation without the need for annotated real data. We leverage the aforementioned supervised polarimetric 6D object pose estimation method (Gao et al., 2022) as a teacher network and pre-train it on synthetically rendered polarimetric image data only. We then utilize its noisy predictions on real data, to support a student network with weak labels for guidance. A differentiable renderer is employed to enable self-supervision with dense geometric cues. Additionally, we propose an invertible formulation of the physical polarization model to analytically compute pixel-wise image characteristics from the geometric normal representation after the differentiable rendering of the student with the predicted 6D pose. This analytic inversion closes the self-supervision loop and allows for direct comparison with the input polarization as illustrated in Fig. 1.

While we adopt the architecture of PPP-Net for a teacher network with an additional differentiable renderer, different from Gao et al. (2022), we use this network to only train on synthetic data. This pre-trained model then produces predictions on the 6D pose of objects on real data, which are leveraged in our proposed teacher–student scheme as weak labels. The teacher network, based on PPP-Net, is thus merely one element of the overall method $\mathbf{S}^2\mathbf{P}^3$ as introduced here.

Inspired by the advancements in self-supervised learning and the use of differentiable renderers in end-to-end learning pipelines, as e.g. in Self6D++ (Wang et al., 2021a), we transfer such knowledge to the multi-modal imaging domain of polarization. Unlike Self6D++, where a renderer produces geometric information in terms of a depth map, which is then compared against a presumably noisy depth map from an active depth sensor, i.e., as explained in later sections here, we carefully study the physical properties of light and integrate encoded shape priors into a self-supervised scheme. This is possible through the differentiable analytical derivation of the physical properties from surface normal information.

The full pipeline of $\mathbf{S}^2\mathbf{P}^3$ thus includes (a) novel architectural designs for the encoding of physical shape priors that extend the findings from PPP-Net to a student–teacher scheme; (b) integrates RGB-agnostic shape information as surface normal maps from a differentiable renderer, offering a more resilient alternative to the issues posed by active depth sensors for photometrically complex objects; (c) entails weak pseudo-labels in the form of geometric and pose information for self-supervision from the teacher network; and most notably, (d) proposes an inverted physical model to leverage shape priors. The lightweight student network predicts and encodes these into a surface normal representation through a differentiable renderer. This encoded representation is

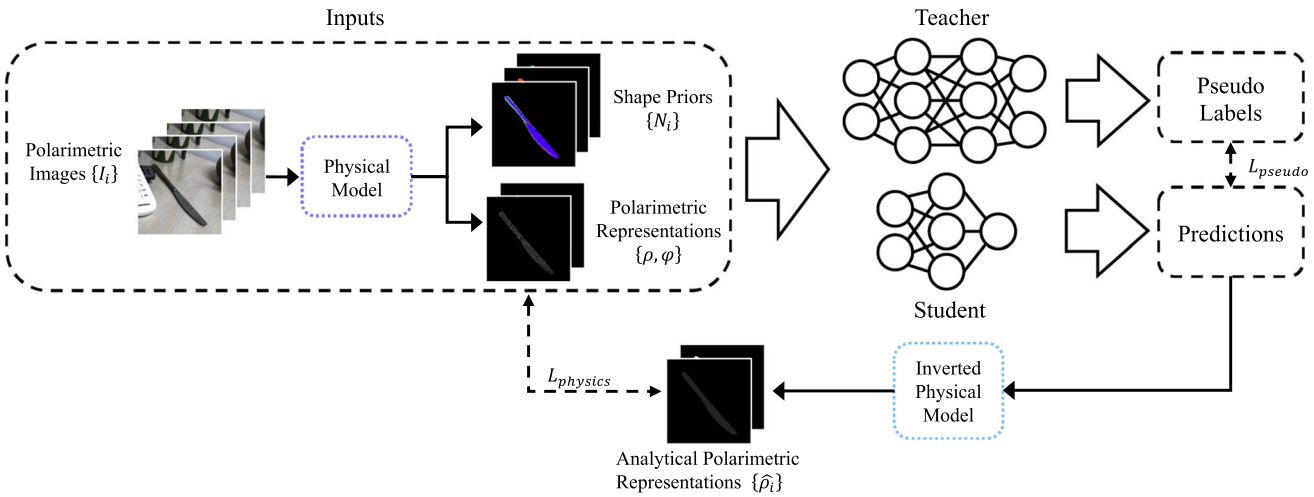

Fig. 1 $\mathbf{S}^2\mathbf{P}^3$ pipeline overview. Our proposed teacher–student training scheme takes four polarization images taken under different polarization filter angles as well as polarimetric and geometrical representations derived from the analytical physical model as multi-modal inputs to both the teacher and student networks, individually. The student network is optimized not only towards the pseudo labels generated from the teacher denoted as L_{pseudo}, but also by $L_{physics}$ which minimizes the discrepancy between the polarimetric representations ρ from the input images after the analytical physical model (cf. Inputs) and $\hat{\rho}$ derived through the inverted physical model from the predicted surface normal of the student network

then utilized to derive the object's analytical polarimetric representation. By integrating this representation into a new physical loss, we achieve complete end-to-end self-supervision using raw polarimetric images.

To this end we contribute in summary:

1. $\mathbf{S}^2\mathbf{P}^3$ as a *hybrid neural-physics* approach to learn *6D object pose* prediction with photometric challenges through *self-supervision* with *neural encodings of geometric shape priors* from mutli-modal data.
2. Insights on the interplay of *differentiable rendering* with the *invertible physical model* through extensive experiments on objects of *varying photometric complexity*.
3. An *instance-level synthetic polarimetric image dataset* for 6D pose estimation that comprises objects present in PPP-Net (Gao et al., 2022) and PhoCal (Wang et al., 2022).

2 Related Work

We revise related work in the realm of polarimetric imaging and 6D object pose estimation, including relevant datasets and recent self-supervised approaches, to provide a solid overview in the research field.

2.1 Polarimetric Imaging

Early works on shape from polarization (SfP) investigate how the relation between polarization and the object's surface can be used to estimate surface normals and depth information, but focus on lab scenarios with controlled conditions of the environment (Atkinson & Hancock, 2006; Garcia et al., 2015; Smith et al., 2018; Yu et al., 2017). These methods only rely on monocular polarization images, but multiple views can also be used for SfP (Atkinson & Hancock, 2005; Cui et al., 2017), also extending to depth estimation (Verdie et al., 2022) from a freely moving camera. In Verdie et al. (2022) the goal is to predict dense depth for outdoor scenes with photometrically easy objects in a (partly) supervised manner with depth measurements from an active structured light sensor while leveraging multi-modal input to account for other artefacts that affect depth predictions. Polarimetric images are also combined with photometric information from either stereo (Atkinson, 2017) or monocular RGB (Zhu & Smith, 2019) to complement each other for depth predictions. Polarized light can also improve initial noisy depth maps from other sensors (Kadambi et al., 2017). Ba et al. (2020) compute a set of plausible cues from polarimetric images to predict surface normals with a neural network which can disambiguate such cues for SfP. Lei et al. (2022) present a novel method for scene-level surface normal estimation from a single polarization image. By introducing a unique real-world dataset and employing advanced neural architecture with a multi-head self-attention module and viewing encoding, the study achieves superior performance in complex scenes. Our approach is inspired from these findings to complement the pose estimation with shape priors from physical properties extracted from the polarized light.

2.2 6D Object Pose Estimation

Dense correspondence-based methods (Hodan et al., 2020; Li et al., 2019; Park et al., 2019; Shugurov et al., 2021; Zakharov et al., 2019) gained popularity in recent years for 6D object pose estimation. The key idea is to train a neural network to predict 2D–3D correspondences between each object pixel in the image and the 3D location of the corresponding point on the object's surface. Those correspondences are consecutively used either with PnP + RANSAC (Fischler & Bolles, 1981; Lepetit et al., 2009), the Umeyama algorithm (Umeyama, 1991), or direct regression to compute the 6D object pose. Hierarchical feature representations are proposed in ZebraPose (Su et al., 2022), and also zero-shot methods are being investigated for the task of 6D pose estimation (Shugurov et al., 2022). Many works on correspondence-based methods (Hodan et al., 2020; Li et al., 2019; Park et al., 2019; Shugurov et al., 2021; Zakharov et al., 2019) are limited by the computationally expensive post-processing for the RANSAC-based pose solver. GDR-Net (Wang et al., 2021b) and its follower SO-Pose (Di et al., 2021) use learning-based MLP networks to directly predict the target pose from the predicted dense correspondences to improve the computing efficiency. In $\mathbf{S}^2\mathbf{P}^3$ we build upon these findings to directly regress the object pose.

2.3 Geometric Depth Information

FFB6D (He et al., 2021) introduces a tight coupling strategy from cross-modal information exchanges with a keypoint extraction (He et al., 2020) that leverages geometry from depth. Also other methods like Uni6D (Jiang et al., n.d.), ESA6D (Mo et al., 2022), FS6D (Yisheng et al., 2022) and DGECN (Cao et al., 2022) include depth information into their prediction pipelines. These approaches however, all critically depend on depth quality which suffers for photometrically challenging objects (Gao et al., 2022). Geometric cues from polarization could mitigate such issues.

2.4 Self-Supervision

Self-supervised learning avoids the problem of lacking properly labeled data. In the realm of 6D pose estimation, differentiable rendering is being used to render synthetic images with a predicted pose to compare against input images (Sock et al., 2020). Self6D (Wang et al., 2020) pro-

poses such approach, where a network is first trained on synthetic RGB data and then fine-tuned on real RGB-D data without pose annotations in a self-supervised manner. They use depth data to align the visual and geometric cues which is the core part in the self-supervision stage. Building on top of Self6D, Self6D++ (Wang et al., 2021a) replaces the one-stage pose regression backbone to two-stage GDR-net (Wang et al., 2021b) backbone, and additionally introduces a pose refiner on top of the teacher network to improve the accuracy and the robustness towards occlusions.

2.5 Polarimetric 6D Pose Prediction

With recently published annotated datasets for real-world polarimetric category-level (Wang et al., 2022) and instance-level (Gao et al., 2022) 6D pose estimation, it is now possible to study methods with this mostly unexplored imaging modality (Jung et al., n.d.). PPP-Net (Gao et al., 2022) investigates the advantages of using polarization for supervised object pose estimation, and designs a hybrid pipeline leveraging polarization through a combination of physical model cues with learning, yielding impressive performance for photometrically challenging objects when compared against RGB and RGB-D baselines. However, acquiring real training data with accurate annotations is still difficult and not easily reproducible for other scholars without complex and expensive hardware (Gao et al., 2022; Wang et al., 2022). Inspired by the strengths of polarimetric information in the supervised learning, we investigate the logical, yet non-trivial, next step towards exploring how this interesting modality can be integrated into a self-supervised scheme to reduce the need for annotated data. Different from Self6D (Wang et al., 2020) and Self6D++ (Wang et al., 2021a), we leverage polarimetric images, and extend the differentiable renderer to yield—besides appearance information—geometric representations in terms of normal maps of the object of interest. We further utilize this representation to compute polarimetric properties used for additional self-supervision through our proposed invertible physical model. To the best of our knowledge, we present the first method to utilize the geometric information from polarization in a self-supervised learning scheme.

3 Polarimetric Physical Model

Commonly used sensors in computer vision send or receive light to measure the wavelength and energy within some specific spectrum. Additionally to this information, the relative oscillation of the electromagnetic wave defines its polarization. Emitted unpolarized natural light becomes polarized after being reflected from a surface, hence it carries information about the object's surface characteristics. The utilization of RGB-D sensors in pose estimation has gained popularity owing to their cost-effectiveness and easy integration into various devices. These sensors utilize active illumination for depth measurement, either through projection of a pattern or time-of-flight measurements. However, they are prone to photometric challenges such as translucency and reflections that can result in erroneous depth estimates. This paper presents a solution to these challenges through the use of surface normals derived from polarization of an RGB-P sensor (refer to Fig. 2). After discussing some issus of RGB-D sensors, this section will introduce how aforementioned information can be measured with a passive sensor with integrated polarization filters. Then we will introduce how the physical model computes geometric shape priors from the information encoded in the polarimetric images and how our invertible formulation is integrated into our network architecture to enable direct self-supervision.

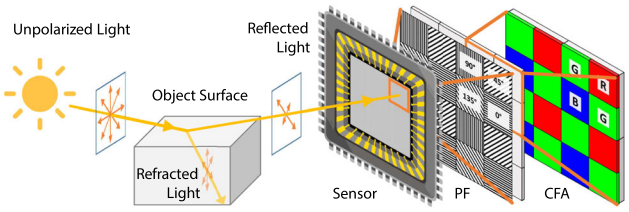

Fig. 2 Polarization camera. When an unpolarized light source reflects on an object surface, the resulting reflection comprises both a refracted and a reflected part, both of which are partially polarized. A polarization sensor captures this reflected light. In front of each pixel of the sensor, there are four polarization filters (PF) arranged at different angles: 0°, 45°, 90°, and 135°. Additionally, a colour filter array (CFA) is used to separate the reflected light into different wavebands

3.1 Photometric Challenges for RGB-D

Commercial depth sensors rely on photometric measurements to estimate depth, by using active illumination either by projecting a pattern (e.g. intel RealSense D series) or using time-of-flight (ToF) measurements (e.g. Kinect v2/Azure Kinect, intel RealSense L series). This makes them susceptible to challenges such as reflections and translucency, which can artificially extend the roundtrip time of photons or deteriorate the projected pattern. As a result, accurate depth estimation becomes infeasible in such scenarios, as illustrated in Fig. 3 for a set of common household objects. The ToF sensor (RealSense L515) used in the experiment struggles to detect the semi-transparent vase, which appears almost invisible to the sensor. Additionally, reflections on the *cutlery* and *can*, cause the sensor to generate depth estimates that are significantly further from the true value, while strong reflections at boundaries result in pixel distances that are invalidated.

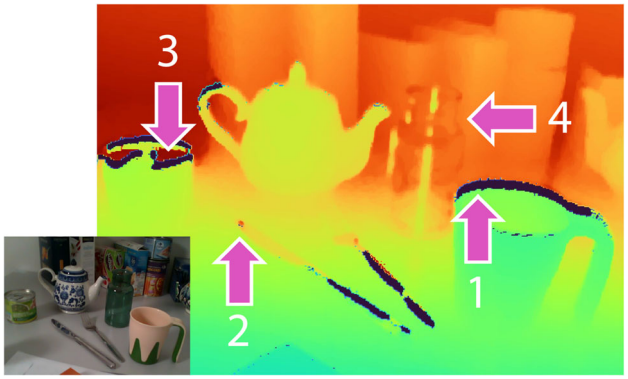

Fig. 3 Depth artifacts. The RealSense L515 depth sensor exhibits miscalculations in depth values for common household objects. Specifically, boundaries (1,3) invalidate pixels, and strong reflections (2,3) lead to incorrect depth estimates that are too far from the true value. In the case of semi-transparent objects like the vase (4), the depth sensor has difficulty detecting them, resulting in partially invisible objects and inaccurate measurements of the distance to objects behind them

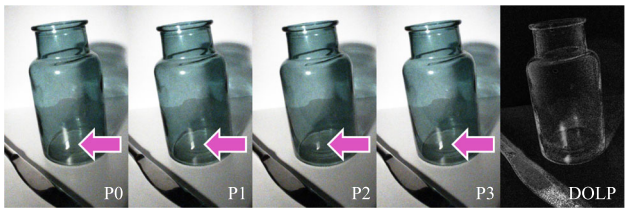

Fig. 4 Degree of polarization. The polarization of light changes when it reflects off a translucent surface, resulting in differences in the polarimetric image quadruplet, with different polarization angles (P0–P3), that are directly related to the surface normal. In particular, the degree of polarization (DoP) for both the translucent and reflective surfaces is considerably higher than for the rest of the image, as shown in the indicated areas in the image

3.2 Surface Normals from Polarization

Most artificial and natural light is unpolarized, meaning the electromagnetic wave oscillates along all planes perpendicular to the direction of propagation of the light (Fließbach, 2012). When unpolarized light passes through a linear polarizer or is reflected at Brewster's angle from a surface, it becomes perfectly polarized. The refractive index of a material determines how fast light travels through it, how much of it is reflected, and the Brewster's angle of that medium. When light is reflected at the same angle to the surface normal as the incident ray, we call it specular reflection. The remaining part penetrates the object as refracted light, which becomes partially polarized as it traverses through the medium. This light wave escapes from the object and creates diffuse reflection (Fliesßbach, 2012). We use Fig. 4 to provide an example that illustrates these concepts.

For 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. We propose to use surface normals obtained from polarization to overcome the photometric challenges faced by RGB-D sensors. Our method can be applied to various applications, including pose estimation, where accurate 3D information is crucial.

3.3 Image Formation Model

We present the fundamental polarization image formation model and our invertible physical model that links the polarimetric and geometrical representations. When light with a specific intensity I and wavelength λ reaches the sensor, it passes through the color filter array (CFA), which separates the light into RGB wavebands, as shown in Fig. 2. The incoming light also has a degree of polarization (DoP) ρ and a direction (angle) of polarization (AoP) ϕ. As light passes through a polarizer array on top of a pixel unit with four different polarization angles $\varphi_{pol} \in \{0°, 45°, 90°, 135°\}$, the oscillation state of light is recorded alongside its wavelength and energy (Kalra et al., 2020). The polarization image formation model in Eq. 1 defines the underlying parameters that contribute to the captured polarized intensities as:

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

where the unpolarized intensity I_{un} can be computed via averaging over polarized intensities $I_{\varphi_{pol}}$ under different polarization filter angles $\varphi_{pol} \in \{0°, 45°, 90°, 135°\}$. The degree of polarization (DoP) ρ and angle of polarization (AoP) ϕ can be solved from a linear least squares system (Huynh et al., 2010) from a set of polarization images captured under different polarization filter angles as:

$$\begin{bmatrix} I_{\varphi_{pol,1}} \\ \vdots \\ I_{\varphi_{pol,4}} \end{bmatrix} = \begin{bmatrix} 1 & \cos 2\varphi_{pol,1} & \sin 2\varphi_{pol,1} \\ & \vdots & \\ 1 & \cos 2\varphi_{pol,4} & \sin 2\varphi_{pol,4} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \tag{2}$$

where the unknowns x_i in the linear system represent $x_1 = I_{un}$, $x_2 = I_{un}\rho \cos 2\phi$, and $x_3 = I_{un}\rho \sin 2\phi$.

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

$$\begin{cases} \phi_d[\pi] = \alpha & \text{for diffuse reflection} \\ \phi_s[\pi] = \alpha - \frac{\pi}{2} & \text{for specular reflection} \end{cases}, \tag{3}$$

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 polarization by considering Fresnel coefficients, thus DoP is similarly given

by (Atkinson & Hancock, 2006):

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

with the refractive index of the observed object material η. Solving Eq. 4 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:

$$\mathbf{n} = (\cos\alpha \sin\theta, \sin\alpha \sin\theta, \cos\theta)^{\mathrm{T}}. \tag{5}$$

We use these plausible normals $\mathbf{n}_d, \mathbf{n}_{s1}, \mathbf{n}_{s2}$ as physical priors per pixel as input to the neural network.

With the help of the physical model defined by Eqs. 1 and 2, we can now derive physical polarimetric characteristics which encode shape information as geometric normals. More formally, when light gets reflected by the object's surface, the shape information is encoded in the captured polarization intensities accordingly. The physical model in our pipeline reveals the implicitly encoded shape information to provide object-centric priors orthogonal to intensity information. We derive a set of explicit object shape priors N_i based on polarimetric intensities $I_{\varphi_{pol}}$ and properties ρ, ϕ as Ba et al. (2020) and Zou et al. (2020). The ambiguities within this process lead to non-unique solutions as in Ba et al. (2020), yet we encode them in a pixel-exclusive manner to guide the network to distinguish between different priors and extract meaningful geometrical features.

3.4 Invertible Physical Model

Inverting the model and assuming a given normal map of an object, e.g., from a differentiable renderer with an estimated 6D pose as in our training scheme, we define an invertible solution to solve for the polarimetric representation analytically. This serves to close the loop from the network's prediction by transferring the information of the object's pose parameterized as 6D transformation through a differentiable renderer into a geometric form and further into encoded physical properties of light reflections that can be compared against the original input information in a self-supervised scheme.

The inverted physical model aims to bring a loop closure from the other end by taking the rendered object surface normal map to analytical polarimetric parameters considering different reflection properties. We obtain the viewing angle θ_v from $\cos\theta_v = \mathbf{n} \cdot \mathbf{v}$ where $\mathbf{n}$ is the rendered object surface normal map, and the viewing vector $\mathbf{v}$ is defined as $\mathbf{v} = -\pi^{-1}(u, v, K)$ with π^{-1} which serves as backprojection operation for pixel (u, v) with camera intrinsics K. The analytical DoP $\hat{\rho}$ is then derived via formulations for diffuse and specular reflection cases:

$$\begin{cases} \hat{\rho_d} = \frac{(\eta - 1/\eta)^2 \sin^2(\theta_v)}{2+2\eta^2-(\eta+1/\eta)^2 \sin^2(\theta_v)+4\cos(\theta_v)\sqrt{\eta^2-\sin^2(\theta_v)}} \\ \hat{\rho_s} = \frac{2\sin^2(\theta_v)\cos(\theta_v)\sqrt{\eta^2-\sin^2(\theta_v)}}{\eta^2-\sin^2(\theta_v)-\eta^2\sin^2(\theta_v)+2\sin^4(\theta_v)} \end{cases}, \tag{6}$$

where η is a constant defined by the refractive index of object materials. The inverted physical model offers the possibility to optimize the model via object shape cues, which is more robust in photometrically challenging scenarios compared to active depth sensors.

4 Methodology

The objective of $\mathbf{S^2P^3}$ is to achieve 6D object pose prediction without relying on annotated real data. To accomplish this, a teacher–student training approach is suggested, which utilizes pre-training on synthetic data and pseudo-labels from the teacher during self-supervision as depicted in Fig. 1. By additionally incorporating the proposed invertible physical model for self-supervision, $\mathbf{S^2P^3}$ makes full use of the geometric data encoded in the polarimetric images. This section outlines the hybrid polarization-based pipeline for learning object pose and explains the physics-induced self-supervision approach in detail.

4.1 $\mathbf{S^2P^3}$ Network Architecture

$\mathbf{S^2P^3}$, consisting of a teacher network (cf. Fig. 5) with a larger capacity and a light student network (cf. Fig. 6), is illustrated in Fig. 7 as a schematic overview. Both networks are pre-trained on synthetic data, whereas the teacher later provides pseudo labels on real data to guide the student network in a self-supervised manner. The detailed architecture illustrates essential extensions, modifications, and important design choices of $\mathbf{S^2P^3}$ compared against established student–teacher training schemes in the community of 6D object pose estimation (Wang et al., 2021a). These are explained in detail in the following and justified with ablations in our experiments section.

4.1.1 Teacher Network

Inspired by the architecture of PPP-Net (Gao et al., 2022), we propose our polarimetric network with an extended differentiable renderer, as the teacher of $\mathbf{S^2P^3}$ (cf. Fig. 5). Here, the inputs of polarimetric intensities and geometrical shape priors are encoded through separate input heads, followed by an explicit decoder to predict an object mask $\tilde{\mathbf{M}}_\mathbf{t}$, an object normal map $\tilde{\mathbf{N}}_\mathbf{t}$, and the dense correspondences as

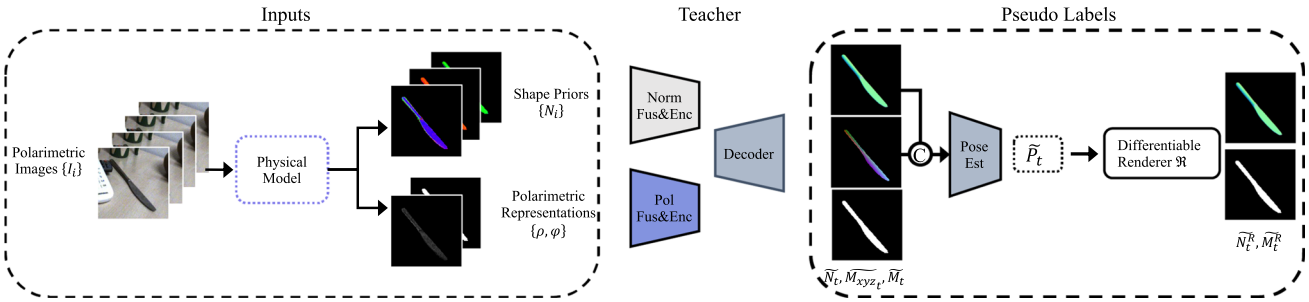

Fig. 5 $\mathbf{S^2P^3}$ teacher network. The network takes the shape priors and polarimetric representations, both derived from the analytical physical model from four polarized images, as input. Before retrieving the 6D object pose, intermediate geometrical representations are predicted. A differentiable renderer utilizes the predicted pose to provide a rendered normal map and object mask

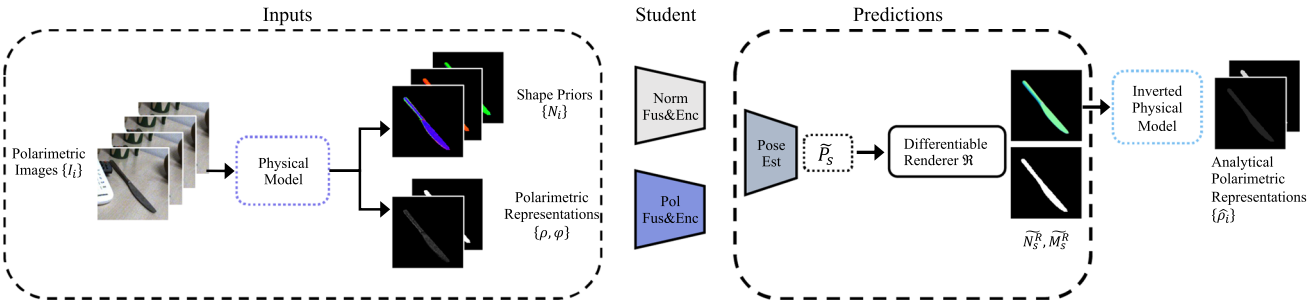

Fig. 6 $\mathbf{S^2P^3}$ student network. Different from the teacher network in Fig. 5, the student is more light-weight by neglecting the explicit decoding of predicted geometric representations

normalized object coordinate map $\tilde{\mathbf{M}}_{\mathbf{xyz_t}}$. The spatial and shape correlation of $\tilde{\mathbf{M}}_{\mathbf{xyz_t}}$ and $\tilde{\mathbf{N}}_\mathbf{t}$ serve as inputs to an object pose estimation module (Wang et al., 2021b), in which the predicted rotation vector is parameterized in the form of allocentric continuous 6D representation (Zhou et al., 2019) and the predicted translation as scale-invariant vector (Li et al., 2019). We further convert them into a standard rotation matrix $\tilde{\mathbf{R}}_\mathbf{t} \in \mathbb{R}^{3\times3}$ and a translation vector $\tilde{\mathbf{t}}_\mathbf{t} \in \mathbb{R}^3$ and denote the final pose as $\tilde{\mathbf{P}}_\mathbf{t} = [\tilde{\mathbf{R}}_\mathbf{t} \mid \tilde{\mathbf{t}}_\mathbf{t}]$. Here, we extend the neural network of PPP-Net. To compute pixel-wise geometrical pseudo labels from the predicted pose, a differentiable renderer takes the object's CAD model and $\tilde{\mathbf{P}}_\mathbf{t}$ as inputs to render an object mask $\tilde{\mathbf{M}}_\mathbf{t}^\mathbf{R}$ and an object normal map $\tilde{\mathbf{N}}_\mathbf{t}^\mathbf{R}$. All the predicted and rendered quantities serve as weak pseudo labels for the student network.

4.1.2 Student Network

We propose a lightweight student network without explicit geometric decoder, different to Self6D++ (Wang et al., 2021a), where the network directly regresses the predicted pose for the student $\hat{\mathbf{P}}_\mathbf{s}$ (cf. Fig. 6). This also favors fast inference while maintaining high accuracy. Our ablations, discussed later in Table 4, indicate the superiority of our student network design. The teacher network consists of about 5.5 million weights, whereas our lightweight teacher does not need the explicit decoder, thus reducing the network to about 5 million weights. While the number of parameters is not significantly reduced, the inference time and also pose prediction accuracy is greatly improved by not predicting the intermediate geometric representations, as discussed later in the results section. We test this against the design choice of Self6D++ (Wang et al., 2021a) of having the student network identical to the teacher but without a subsequent pose refiner. Our student network converges towards better predictions without the redundant explicit prediction of intermediate geometric representations with our proposed self-supervision. The final output of our student in $\mathbf{S^2P^3}$, thus only consists of the predicted pose $\hat{\mathbf{P}}_\mathbf{s}$. To link the predictions with geometric and polarimetric properties, we render an object normal map $\hat{\mathbf{N}}_\mathbf{s}$ and an object mask $\hat{\mathbf{M}}_\mathbf{s}$ given $\hat{\mathbf{P}}_\mathbf{s}$ via the differentiable renderer—analogous to the teacher network. We will detail how this polarimetric representation of the geometric information is utilized in a self-supervised loss term in the following.

4.2 Physics-Induced Self-Supervised Training Scheme

As detailed before, the polarimetric images contain rich information that we provide as explicit representations to the network to learn neural geometric encodings. This sec-

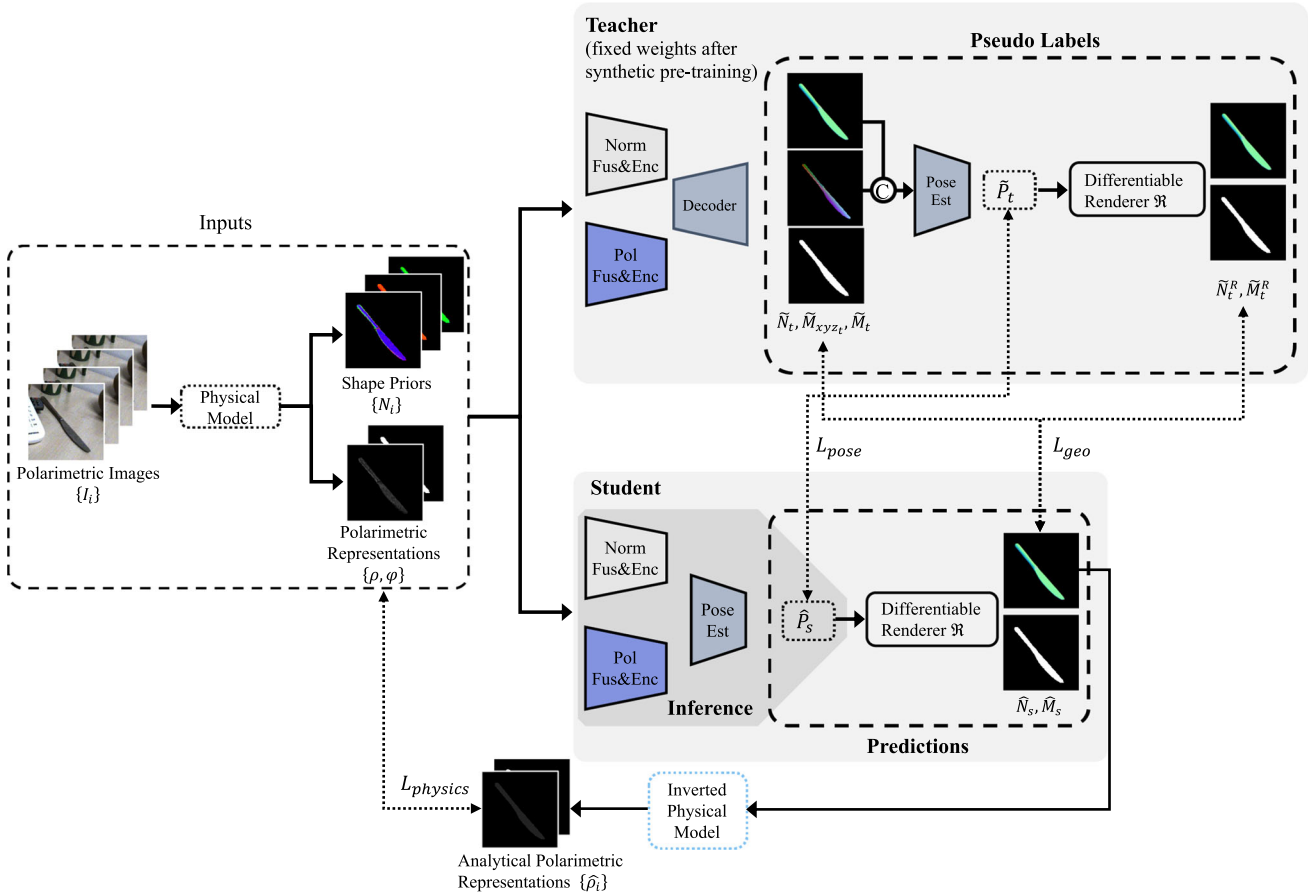

Fig. 7 $\mathbf{S^2P^3}$ pipeline overview. Our proposed teacher–student training scheme takes four polarization images taken under different polarization filter angles as well as polarimetric and geometrical representations derived from the physical model as inputs to both the teacher and student networks. The student network is optimized not only towards the pseudo labels generated from the teacher denoted as L_{pseudo}, but also by $L_{physics}$ which minimizes the discrepancy between ρ from the physical model and $\hat{\rho}$ from the inverted physical model. During inference, the lightweight student network only predicts direct pose estimates as indicated by the gray background color

tion defines how these representations are further leveraged and integrated into our physically induced self-supervised scheme, firstly through implicit and explicit weak pseudo-labels of the teacher network, and second as direct coupling by closing the loop towards the input information of the pipeline.

4.3 Loss Formulations

Our proposed optimization scheme comprises two complementary paradigms. The first passes knowledge of the pre-trained teacher to the student in the form of weak labels of the pose $\tilde{\mathbf{P}}_\mathbf{t}$ and related object shape knowledge $\{\tilde{\mathbf{M}}_\mathbf{t}, \tilde{\mathbf{N}}_\mathbf{t}, \tilde{\mathbf{M}}_\mathbf{t}^\mathbf{R}, \tilde{\mathbf{N}}_\mathbf{t}^\mathbf{R}\}$, which we define as pseudo label loss $\mathcal{L}_{pseudo}$. The second is to utilize the inverted physical model to optimize the student prediction $\hat{\mathbf{P}}_\mathbf{s}$ via raw polarization data in our physical loss term $\mathcal{L}_{physics}$ detailed below.

To account for potential misalignment between the decoded shape knowledge $\{\tilde{\mathbf{M}}_\mathbf{t}, \tilde{\mathbf{N}}_\mathbf{t}\}$ and pose knowledge $\tilde{\mathbf{P}}_\mathbf{t}$, we compare the predicted mask $\tilde{\mathbf{M}}_\mathbf{t}$ and the rendered mask $\tilde{\mathbf{M}}_\mathbf{t}^\mathbf{R}$ and normalize the discrepancy to a scalar value of δ, which serves as the criteria of choosing pseudo ground truth for the geometrical regularization term $\mathcal{L}_{geo}$ and a dynamic weighting term in the overall learning objective. The final formulation is then:

$$\mathcal{L}_{pseudo} = \lambda_1 \mathcal{L}_{pose} + \mathcal{L}_{geo}, \tag{7}$$

with:

$$\mathcal{L}_{pose} = \underset{\mathbf{x}\in\mathcal{M}}{\text{avg}} \, \|(\tilde{\mathbf{R}}_\mathbf{t}\mathbf{x} + \tilde{\mathbf{t}}_\mathbf{t}) - (\hat{\mathbf{R}}_\mathbf{s}\mathbf{x} + \hat{\mathbf{t}}_\mathbf{s})\|_1, \tag{8}$$

$$\mathcal{L}_{geo} = \mathcal{L}_{mask} + \mathcal{L}_{normals}, \tag{9}$$

in which we define the $\mathcal{L}_{mask}$ as mean squared error and $\mathcal{L}_{normals}$ as cosine similarity loss. The rendered represen-

tations $\{\tilde{\mathbf{M}}_{\mathbf{t}}^{\mathbf{R}}, \tilde{\mathbf{N}}_{\mathbf{t}}^{\mathbf{R}}\}$ are chosen as geometrical pseudo ground truth if δ is within a predefined threshold r, otherwise the predicted representations are selected, also leading to a reduced weighting factor $\lambda_1 = (1 - \delta)$ on direct pseudo pose loss $\mathcal{L}_{pose}$.

4.3.1 Physical Constraints

To enable self-supervision via the invertible physical model, the rendered geometric normal map $\hat{\mathbf{N}}_{\mathbf{s}}$ serves as input to solve for analytical diffuse and specular DoP $\{\hat{\rho_d}, \hat{\rho_s}\}$ according to Eq. 6. To benefit from the underlying physical process of polarimetric imaging, $\mathcal{L}_{physics}$ deploys a pixel-wise minimum selection mechanism inspired by (Verdie et al., 2022):

$$\mathcal{L}_{physics} = \min_{\mathbf{x} \in \{\hat{\rho_d}, \hat{\rho_s}\}} \|\rho - \mathbf{x}\|_1. \tag{10}$$

To avoid the domain gap between the analytically solved intensity map and the real polarimetric images as in Verdie et al. (2022), we directly formulate the loss function based on polarimetric properties instead of polarimetric intensities. Hence, the student's output is optimized to align with raw DoP ρ from real polarization images.

The overall loss combines the knowledge from the teacher and the raw data as:

$$\mathcal{L} = \mathcal{L}_{pseudo} + \mathcal{L}_{physics}. \tag{11}$$

5 Experimental Results

We perform extensive evaluations and ablations on the instance-level polarimetric 6D pose dataset on which PPP-Net (Gao et al., 2022) provides a strong baseline against RGB-only (Wang et al., 2021b) and RGB-D (He et al., 2021) state-of-the-art supervised methods (Wang et al., 2022). This section first states implementation parameters for training, outlines the synthetic dataset generation, and describes the real polarimetric dataset. Detailed quantitative results on real data are discussed, and extensive ablations on different loss terms and modalities are analyzed. Our experiments specifically study the influence of polarimetric physical cues in a self-supervised scheme on objects of varying photometric complexity for instance-level 6D object pose prediction. Polarimetric images and self-supervised schemes are both mostly unexplored tracks in 6D pose estimation. As such, we take the supervised PPP-Net (Gao et al., 2022), and the self-supervised Self6D++ (Wang et al., 2021a) trained on RGB and RGB-D data, as strong baselines for comparison. Self6D++ (Wang et al., 2021a) is the SOTA method in self-supervised 6D object pose estimation with RGB-D information, outperforming other baselines by a large margin (Sock et al., 2020; Wang et al., 2020). As such, it represents a valid comparison and justifies the improvements of our method. Likewise, PPP-Net (Gao et al., 2022) outperforms state-of-the-art RGB-only methods on photometrically challenging objects as under consideration here. Hence, it constitutes a legitimate representative of RGB-only methods as strong baseline for the experiments under consideration here.

5.1 Synthetic Data Generation

Given a CAD model of an object, we randomly sample camera locations on its upper hemisphere for rendering. To further enforce realistic renderings and to reduce the domain gap, we set up backgrounds with different textures and lighting positions in Mitsuba2 renderer (Nimier-David et al., 2019) to acquire 200–800 sets of polarization images for each object.

We present illustrations of our synthetic dataset for different viewpoints in Fig. 8 to illustrate the variety of sampled poses, objects of different photometric complexity, and their appearance in the image. The synthetic dataset is used to pre-train the teacher and student networks. We render a set of four polarimetric images with different angles of the polarization filter according to the camera used in the real setup.

As rendering is very time-consuming, we provide the dataset.[2] We also train a customized object detector on synthetic data to later provide predicted masks and bounding boxes on the real domain.

We present samples of our real polarimetric dataset with annotated object poses in Fig. 9. The objects rendered using ground truth pose labels indicate the high quality of data annotation, and the object models with white color rendering indicate their textureless nature, which supports our design of removing the need for color texture supervision.

5.2 S^{2}P^3 Training

We detail the two phases of the training: "Synthetic Pre-training" on rendered data and "Self-supervised Training on Real Data". The former uses synthetic data with 6D pose annotations for supervised pre-training of the teacher and the student networks individually. In the latter phase, we use real data to train the student network in a self-supervised fashion by leveraging our proposed novel training scheme and loss function.

5.2.1 Synthetic Pre-training

Both the teacher and student models go through a pre-training phase in which they receive supervision exclusively based on

[2] https://daoyig.github.io/.

Fig. 8 Synthetic dataset. Samples of objects with varying photometric complexity are illustrated from different viewpoints

Fig. 9 Real dataset. Samples of objects are illustrated from different viewpoints. The rendered objects using GT pose illustrate the white-color rendered texture

the 6D pose information, derived from ground truth annotations from synthetic data. During this phase, the loss function has a two-part structure: an L1 loss is utilized for translation, while a point matching loss is applied for rotation. Notably, the differentiable renderer is not integrated into this pre-training stage. In terms of computational time, the pre-training process takes several hours, typically ranging from 4 to 5 h for each object. Subsequently, the self-supervised phase is more time-intensive, demanding approximately 10 h per object.

5.2.2 Self-Supervised Training on Real Data

We evaluate our method on a specific data split of the instance-level 6D pose estimation dataset introduced in Gao et al. (2022) containing objects with varying photometric complexity with highly accurate annotations from robotic forward-kinematics. The RGB-P data is acquired with the polarization camera Phoenix 5.0 MP PHX050S1-QC comprising a Sony IMX264MYR CMOS (Color) Polarsens sensor (LUCID Vision Labs, Inc., Richmond B.C, Canada) and a Universe Compact C-Mount 5MP 2/3″ 6 mm f/2.0 lens (Universe, New York, USA).

As the amount of real data differs between objects, we follow common practice in instance-level object pose estimation literature by sampling around 15–20% of total data for training, and the rest for testing (Gao et al., 2022), which results in 200–300 sets of real polarization images as training data for each object, and 1000–2000 sets of images as testing.

We ensure that the poses for the data split of the rendered synthetic domain are similar to the poses of the real domain

in terms of overall distribution, to ensure comparability when analyzing the domain shift and the influence of our proposed self-supervision scheme later. The predicted bounding box crops out the region containing the object of interest and is resized to 256×256 as inputs to the networks. The predicted object mask serves as input to the physical model to produce only object-related polarimetric parameters as well as shape priors.

5.3 Implementation Details

We implement our model using Pytorch (Paszke et al., 2019) and train on an NVIDIA 2080 GPU and using ADAM optimizer (Kingma & Ba, 2014) on a commodity desktop PC with an Intel i7 CPU processor and 32 GB RAM. The teacher and student networks are trained for 100 epochs for each object individually, both for synthetic and real data. The initial learning rate is set to 1×10^{-4}, and halved every 25 epochs. The weights for all encoders are initialized with ImageNet weights. For synthetic pre-training, we use a batch size of 8, and for the self-supervised training on real data a batch size of 4.

5.4 Refractive Indices

As a material-related coefficient, the refractive index η for each object is listed in Table 1. The index serves as input to both the forward and inverted physical model. The refractive index is assumed to be known, but it has only a minor influence on object pose predictions [cf. also PPP-Net (Gao et al., 2022)]. In terms of objects with different composite materials, we can observe the plastic cup in our experiments, where the plastic material is slightly different for the beige and green parts and the texture changes as well. Given the results for the different refractive indices [cf. Table A6. Refractive Index Ablation. in the Supp.Mat. of PPP-Net (Gao et al., 2022)], we expect that objects with different composite materials can still be handled. An extensive study of composite objects is out of scope for this work, as the PhoCal dataset (Wang et al., 2022) does not include other such objects.

Table 1 Refractive indices

Object	Material	Refractive index
Fork	Stainless steel	2.75
Knife	Stainless steel	2.75
Bottle	Glass	1.52
Cup	Plastics	1.50

5.5 Evaluation Metrics

The results are evaluated using the common Average Distance of Distinguishable Model Points (ADD) metric (Hinterstoisser et al., 2013) for non-symmetrical objects, in which 10% of the object's diameter is set as the threshold to judge the average deviation of the transformed model points. For symmetric objects, the average deviation to the closest model points is measured as in the Distance of Indistinguishable Model Points (ADD-S) metric (Hodaň et al., 2016).

5.6 Quantitative Results: Baseline Comparisons

$\mathbf{S^2P^3}$ proposes to leverage polarimetric information for self-supervised 6D object pose estimation and focuses on photometrically challenging objects, where self-supervised RGB-D methods may fail due to inherent sensor data artifacts, and supervised approaches, either RGB-only or RGB-P methods as e.g. PPP-Net (Gao et al., 2022), would require a large amount of annotated real data. Therefore, the experiments are deliberately chosen to analyse the multi-modal self-supervision through the physical constraints, its loss functions, as well as the architecture and design choices for the student–teacher scheme in the ablation studies to yield best scientific insights into self-supervised polarimetric 6D pose estimation. As such, we compare $\mathbf{S^2P^3}$ against PPP-Net (Gao et al., 2022) on our data split, as a very strong supervised baseline, in order to analyze the effect of self-supervision. PPP-Net already outperforms other strong state-of-the-art RGB-only methods as reported in Gao et al. (2022), and is thus a valid upper threshold for comparison. Self6D++ (Wang et al., 2021a) is the state-of-the-art self-supervised RGB-D method on many standard benchmark datasets, and is thus chosen for establishing polarimetric self-supervision in $\mathbf{S^2P^3}$ as a strong baseline. See also the qualitative results in Figs. 10 and 11, also including occlusions of the object as in Figs. 12 and 13, for visual results which are discussed later in more detail.

We prove the effectiveness of the self-supervision pipeline by quantitative results in Table 2. Please note, that PPP-Net (trained on annotated real data) is the identical network as we use in our teacher model but without the differentiable renderer. In our full model $\mathbf{S^2P^3}$ however, we do not train the teacher in a supervised manner on real data, but only pre-train it on the synthetic data. Then, the weights of the teacher are frozen and it only provides weak pseudo-labels on real data for the teacher–student scheme. Our model $\mathbf{S^2P^3}$, consistently outperforms the self-supervised learning-based state-of-the-art RGB-D method Self6D++ by Wang et al. (2021a),[3] and even reaches comparable performance against

[3] Self6D++ is trained and tested on our dataset, with RGB-D information from Realsense L515 sensor.

Fig. 10 $\mathbf{S^2P^3}$ qualitative results before and after self-supervision. The projected bounding boxes in blue, red and green represent the ground-truth 6D object poses, the results before and after applying self-supervision, respectively

Fig. 11 $\mathbf{S^2P^3}$ qualitative results before and after self-supervision (zoomed-in from Fig. 10). The projected bounding boxes in blue, red and green represent the ground-truth 6D object poses, the results before and after applying self-supervision, respectively

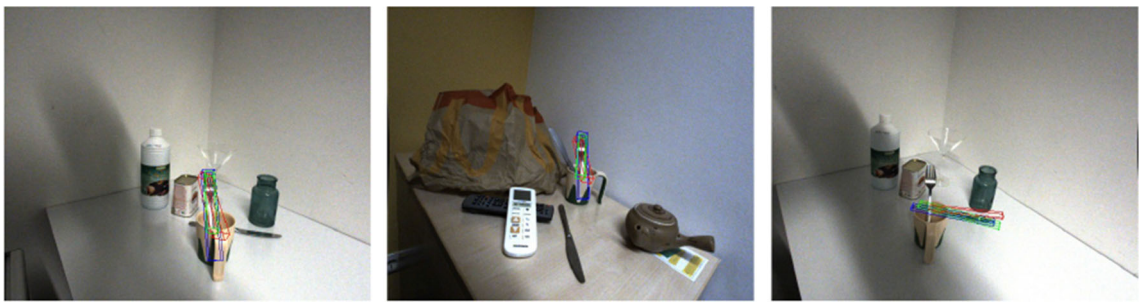

Fig. 12 $\mathbf{S^2P^3}$ qualitative results before and after self-supervision with occlusions. The projected bounding boxes in blue, red and green represent the ground-truth 6D object poses, the results before and after applying self-supervision, respectively

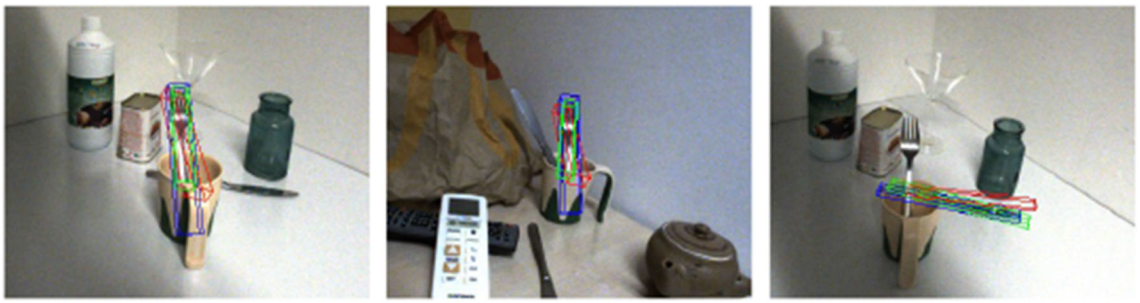

Fig. 13 $\mathbf{S^2P^3}$ qualitative results before and after self-supervision with occlusions (zoomed-in from Fig. 12). The projected bounding boxes in blue, red and green represent the ground-truth 6D object poses, the results before and after applying self-supervision, respectively

Table 2 $\mathbf{S^2P^3}$ quantitative results

Methods	Training	Cup	Fork	Knife	Bottle	Mean
PPP-Net	Supervised	91.4	91.7	90.0	89.4	90.6
Self6D++	Self-supervised (RGB-D)	68.4	14.3	17.8	33.5	34.0
$\mathbf{S^2P^3}$ (Ours)	Self-supervised (**RGB-P**)	**93.8**	**72.4**	**78.4**	**78.2**	**80.7**

Average recall of ADD(-S) metric is reported for different objects with increasing photometric complexity. Self6D++ from Wang et al. (2021a). PPP-Net from Gao et al. (2022)
Bold values indicate best results

the fully supervised upper bound baseline (Gao et al., 2022) for photometrically complex objects.

5.7 Ablation Studies

Our evaluation comprises several ablation studies to analyze the nuances of our model's components. We assess performance variations between synthetic and real data domains, particularly in the absence of self-supervision, to answer the question: how well can the student and the teacher network perform on real data, when trained in a supervised fashion on synthetic or real data, respectively, and how much performance gain does $\mathbf{S^2P^3}$ achieve when supervising on synthetic data only and performing self-supervision with real data. We further explore the impact of the student's architecture within the student–teacher paradigm, focusing on whether a lightweight student could match or outperform the teacher when refined on real data. Or to put it simple: Do we need a large student model, identical to the teacher network with a decoder and dedicated geometrical predictions? Or is the design choice of $\mathbf{S^2P^3}$ to directly regress the 6D pose for the student beneficial? Additionally, we dissect the influence of individual loss components, emphasizing the significance of our physically-induced self-supervised loss. And finally investigate the role of depth versus polarimetric information, gauging their relative contributions to the model's efficacy.

5.7.1 Ablation on Domain Shift: $\mathbf{S^2P^3}$'s Self-supervision

Table 3 summarizes the results when training the individual student and teacher network separately (not within the $\mathbf{S^2P^3}$ training scheme), without the differentiable renderer, with supervision on the pose estimation as in the synthetic pre-training. We differentiate whether training is performed on annotated real or synthetic data, and test on real data. As expected, the student and teacher networks, perform worse on real data when trained on synthetic data only, due to the domain shift, compared to training on real data (compare top and lower rows of Table 3 for student and teacher, respectively). The larger teacher network with a dedicated decoder and explicit intermediate geometrical representations, which is identical to PPP-Net (Gao et al., 2022) and marked with † in the table, outperforms the smaller student network when trained in a supervised fashion in both scenarios. Our full pipeline of $\mathbf{S^2P^3}$(where the student is trained self-supervised on real data and the teacher weights are fixed), with our proposed small student network and a teacher, which are both only pre-trained on synthetic data (i.e., the synthetically pre-trained networks correspond to the numbers of the lower part of Table 3), achieves impressive results without being trained on annotations from real images due to our proposed self-supervision paradigm. $\mathbf{S^2P^3}$ even partly outperforms the fully supervised training on real data (cf. top rows against $\mathbf{S^2P^3}$) and achieves comparable results to PPP-Net as fully supervised upper boundary (indicated by †). Notably, the self-supervision of $\mathbf{S^2P^3}$ improves the results against the synthetically pre-trained student network (cf. Table 3 "Student ⋆" against $\mathbf{S^2P^3}$). While this trend holds true for all objects, the observation from before is not as significant for the fork, which may result from large occlusions for this object in the majority of the data (cf. Figs. 12 and 13 where the fork is inside the cup).

Table 3 Domain shift and $\mathbf{S^2P^3}$'s self-supervision

Configuration	Supervised	Self-supervised	Tested on	Cup	Fork	Knife	Bottle	Mean
Student	Real	–	Real	86.4	88.0	91.1	80.4	86.5
Teacher †	Real	–	Real	**91.4**	**91.7**	**90.0**	**89.4**	**90.6**
Student ⋆	Synthetic	–	Real	53.7	64.4	46.1	47.5	52.9
Teacher	Synthetic	–	Real	72.3	**75.0**	67.3	76.2	72.7
$\mathbf{S^2P^3}$ (Ours)	Syn. (Pre-trained)	Real	Real	**93.8**	72.4	**78.4**	**78.2**	**80.7**

Average recall of ADD(-S) metric is reported for different objects with increasing photometric complexity for the student and teacher network individually, when trained in a supervised setting on either real or synthetic data and tested on real data. The full $\mathbf{S^2P^3}$ pipeline, with synthetic pre-training, and self-supervised training of the student on non-annotated real data, is also reported for comparison. "Teacher †' as upper bound is identical to PPP-Net (Gao et al., 2022). "Student ⋆" corresponds to the setting of $\mathbf{S^2P^3}$ **before** applying our proposed self-supervision scheme
Bold values indicate best results

Table 4 Ablation on network architecture

Config	Self-sup.	Cup	Fork	Knife	Bottle	Mean
Our student	None	53.7	64.4	46.1	47.5	52.9
Large student	None	72.3	75.0	67.3	76.2	72.7
Our student	✓ ($\mathbf{S^2P^3}$)	**93.8**	**72.4**	**78.4**	**78.2**	**80.7**
Large Student	✓	88.6	55.9	69.4	77.8	73.0

We compare different architecture designs for the student network, i.e., our small student and a larger student which would be equivalent to the teacher architecture. Our proposed self-supervised student network (Ours) achieves best results across all objects. We report average recall of ADD(-S) metric
Bold values indicate best results

5.7.2 Ablation on Network Architecture: Exchanging the Student

We follow the motivation to utilize a lightweight student for faster inference. We exchange the network architecture for the student network in $\mathbf{S^2P^3}$, with the one that is normally used as teacher, i.e., instead of the network in Fig. 6 we use the one of Fig. 5 as the student, to analyze the influence of the larger network with its dedicated decoder and intermediate geometrical representations. While intermediate geometrical outputs of the large student are beneficial during pre-training in a supervised fashion (cf. first two rows in Table 4), those outputs introduce more optimization objectives when used as the student network to learn the 6D pose of the object. The ablation in Table 4 demonstrates that the lightweight student (Our Student) can achieve better performance than a larger student (Large Student) network after fine-tuning the student on real data with our student–teacher training scheme and self-supervision through physical constraints $\mathcal{L}_{physics}$. The additional parameters and the intermediate geometrical representations of the large student make convergence more difficult. Still, the physical constraints improve the large student significantly after self-supervision (cf. Large student with None and with Self-Supervision). The ablations demonstrate that the lightweight student can achieve better performance than the larger student network after fine-tuning on real data with our self-supervision scheme through physical constraints $\mathcal{L}_{physics}$, as employed in $\mathbf{S^2P^3}$.

5.7.3 Ablation on Loss Terms

We first verify the influence of various loss terms by training the network without each specific loss term for the self-supervision stage as summarized in Table 5. We find that the direct geometrical point matching loss of $\mathcal{L}_{pose}$ is crucial to self-supervision. Without enforcing $\mathcal{L}_{pose}$ for the student against the weak pseudo-labels of the teacher, the training would easily diverge. The physically-induced self-supervised loss $\mathcal{L}_{physics}$, that is derived from our invertible physical derivations, indicates a larger impact on training results compared to geometrical supervision signals from the teacher network, e.g., $\mathcal{L}_{normal}$ and $\mathcal{L}_{mask}$. The captured real polarimetric images contain more robust underlying object shape information compared to the output of the differentiable renderer. The overall performance of the model reaches best accuracy metrics for all objects with varying photometric complexity when all loss ingredients are present, as indicated in the last row of Table 5.

These results indicate, that the convergence of the student can only be guaranteed when weak labels of the teacher network roughly guide the pose predictions. One reason to explain such behavior, is that the differentiable renderer would be completely unconstrained without $\mathcal{L}_{pose}$, thus potentially rendering outputs with pose predictions that are out of the field of view. Dense supervision of the appearance and geometric representations after differentiable rendering further improve the networks performance, while the boost in pose accuracy is most noticeable with our proposed self-supervised physically-induced loss formulation. The contribution of the self-supervision is also apparent in the qualitative results in Figs. 10 and 11. The projected bounding boxes in *green* show better alignment with ground truth (*blue*) after self-supervision, compared against predictions of

Table 5 Ablation on loss terms

Methods	Cup	Fork	Knife	Bottle	Mean
w/o $\mathcal{L}_{pose}$	6.8	0.2	2.3	0.6	2.5
w/o $\mathcal{L}_{physics}$	71.8	72.1	70.8	74.4	72.3
w/o $\mathcal{L}_{normal}$	87.5	61.0	67.3	74.9	72.7
w/o $\mathcal{L}_{mask}$	89.9	64.9	70.1	72.7	74.4
$\mathbf{S^2P^3}$ (Ours)	**93.8**	**72.4**	**78.4**	**78.2**	**80.7**

Average recall of ADD(-S) metric is reported
Bold values indicate best results

Table 6 $\mathbf{S^2P^3}$ ablations on depth modality

Ours with:	Cup	Fork	Knife	Bottle	Mean
RGB-D chamfer	**100.0**	11.6	59.1	40.7	52.9
RGB-D pixel-wise	86.8	32.3	62.5	50.3	58.0
RGB-P ($\mathbf{S^2P^3}$)	93.8	**72.4**	**78.4**	**78.2**	**80.7**

Average recall of ADD(-S) metric is reported
Bold values indicate best results

the pre-trained teacher (*red*). Figures 12 and 13 show additional results for cases where part of the object, here *fork* and *knife*, is occluded.

5.7.4 Ablation on Modalities

RGB-Texture Supervision For textureless and transparent objects, the rendered object texture will only be white, since it does not have any color [cf. also Figure 7 in PhoCal (Wang et al., 2022) and Figure 5 in PPP-Net (Gao et al., 2022)]. This would reduce the RGB-texture loss essentially to the mask loss in our pipeline. Hence, we eliminate the need for texture rendering and instead rely on the physical properties of polarized light.

Depth Supervision To analyze the importance of accurate and reliable geometric representations for the task of 6D object pose estimation, we train our pipeline with depth maps from a direct time of flight (D-ToF) sensor and compare it against the polarimetric $\mathbf{S^2P^3}$ method with our physically-induced self-supervised loss. For this purpose, we adapt our network to have an additional loss term utilizing depth information aside from having almost all other components unchanged. We let the differentiable renderer of the student network additionally render depth maps $\mathbf{D}^R$ given the predicted pose $\hat{\mathbf{P}}_\mathbf{s}$, and employ a chamfer distance loss $\mathcal{L}_{chamfer}$ between the point cloud $\mathbf{P}^R$ back-projected from the rendered depth $\mathbf{D}^R$ and the point cloud $\mathbf{P}$ back-projected from the depth map in the polarization camera coordinate system, to optimize for alignments without explicit 3D–3D correspondence registrations as:

$$\mathcal{L}_{chamfer} = \underset{\mathbf{p}\in\mathbf{P}}{\text{avg}} \min_{\mathbf{p}^r\in\mathbf{P}^R} \|\mathbf{p}-\mathbf{p}^r\|_2 + \underset{\mathbf{p}^r\in\mathbf{P}^R}{\text{avg}} \min_{\mathbf{p}\in\mathbf{P}} \|\mathbf{p}-\mathbf{p}^r\|_2. \quad (12)$$

Besides adding $\mathcal{L}_{chamfer}$ to the pipeline, we remove the $\mathcal{L}_{physics}$ to have a fair comparison of the effectiveness of direct spatial cues from depth and object shape cues from polarimetric physical properties. The results listed in Table 6 indicate the depth cues can be beneficial when the quality is reliable, i.e., the performance on the *cup* peeks when $\mathcal{L}_{chamfer}$ is introduced to the pipeline.

We conduct additional ablations using a pixel-wise depth loss instead of the chamfer distance loss, as reported in Table 6. The experiment illustrates that also with the pixel-wise depth loss, inaccurate depth information would inject incorrect geometric guidance into the pipeline, leading to degraded performance on photometrically challenging objects.

The inherent limitations of the depth sensor cause severe degradation of the depth quality (Jung et al., 2022). The reflective and semi-transparent objects are measured incorrectly due to reflective and translucent object materials. This is also illustrated in detailed large figures of real examples in Fig. 14. In such cases, the strong signal coming from depth alignment loss introduces incorrect spatial awareness, leading to low pose prediction performance.

On the contrary, the shape of the object that is encoded in the polarimetric image modality can provide stable geometric information for objects of all material characteristics presented here, across a variety of photometric complexity, e.g., from a matte plastic cup, to reflective stainless steel cutlery, and translucent and transparent colored glass objects. The analytically retrieved diffuse and specular solutions after the differentiable renderer are stable across all discussed objects. These polarization properties are computed through our invertible model and then utilized in the physics-induced self-supervision scheme against the raw DoP illustrated on the top left in Fig. 14. Please note that $\mathcal{L}_{physics}$ is a pixelwise minimum loss of the diffuse and specular reflection.

5.7.5 Runtime Analysis

On a desktop PC with an Intel i7 4.20 GHz CPU and an NVIDIA 2080 GPU, given a 512×612 image, our student network takes ≈ 7.3 ms for inferring the 6D pose for a single object, which is around 30% faster than the teacher model. Additionally, the preprocessing for the physical prior calculation takes 13.0 ms, and the object detection takes 15.4 ms.

6 Conclusion

6.1 Limitations

The performed experiments highlight the importance of reliable geometric priors for the task of 6D object pose estimation. When the quality of the depth map is reliable and

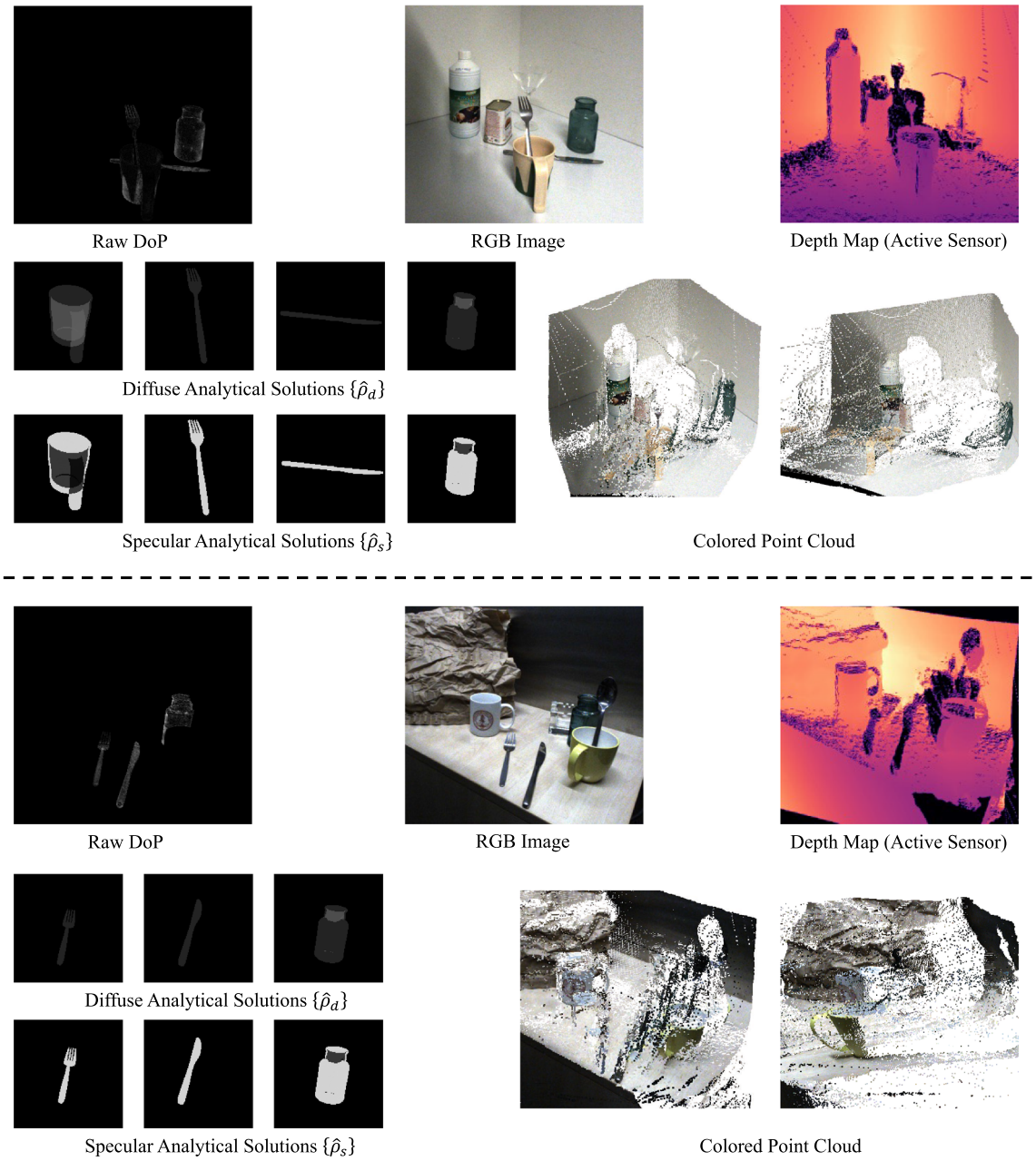

Fig. 14 Examples of polarimetric and depth quality

accurate, the spatial loss term introduced by the source depth map may lead to better performance than pure object-shape-based optimization through polarization. The current model focuses on instance-level pose estimation and does not generalize to unseen objects during training. An interesting future direction is to include the idea in a category-level pipeline.

6.2 Self-Supervised Polarimetric Pose Prediction

This paper bridges two worlds and combines a hybrid model for polarimetric pose estimation that fuses an invertible physical model with neural shape extraction from data within a self-supervised framework. $\mathbf{S}^2\mathbf{P}^3$ solves instance-level object pose estimation from polarimetric images without annotated real data. In our proposed pipeline, a teacher pre-trained on a small set of synthetic renderings ensures convergence of a lightweight student network through weak pseudo-labels. Our employed differentiable renderer additionally provides the appearance and geometric outputs and enables self-supervision. $\mathbf{S}^2\mathbf{P}^3$ outperforms methods that use depth measurements from active sensors for photometrically challenging objects. We achieve this by carefully

integrating distinct design choices in the student–teacher architecture and proposing our invertible physical model for self-supervision by leveraging XoP properties, instead of raw polarimetric data as in Verdie et al. (2022), to reduce the domain gap. Our contributions are validated through extensive ablation studies.

Our experimental results show the importance of self-supervision through geometric and physical cues for the task of 6D pose estimation and yield scientific insights into the robustness of polarimetric images. Such observations are most noticeable for photometrically challenging, textureless, reflective, or translucent objects.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Atkinson, G. A. (2017). Polarisation photometric stereo. *Computer Vision and Image Understanding, 160*, 158–167.

Atkinson, G. A., & Hancock, E. R. (2005). Multi-view surface reconstruction using polarization. In *Tenth IEEE international conference on computer vision (ICCV'05)* (Vol. 1, pp. 309–316).

Atkinson, G. A., & Hancock, E. R. (2006). Recovery of surface orientation from diffuse polarization. *IEEE Transactions on Image Processing, 15*(6), 1653–1664.

Ba, Y., Gilbert, A., Wang, F., Yang, J., Chen, R., Wang, Y., Yan, L., Shi, B., & Kadambi, A. (2020). Deep shape from polarization. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, Part XXIV 16* (pp. 554–571).

Busam, B., Ruhkamp, P., Virga, S., Lentes, B., Rackerseder, J., Navab, N., Hennersperger, C. (2018). Markerless inside-out tracking for 3D ultrasound compounding. In *Simulation, image processing, and ultrasound systems for assisted diagnosis and navigation* (pp. 56–64). Springer.

Cao, T., Luo, F., Fu, Y., Zhang, W., Zheng, S., Xiao, C. (2022). *DGECN: A depth-guided edge convolutional network for end-to- end 6D pose estimation*.

Cui, Y., Schuon, S., Chan, D., Thrun, S., Theobalt, C. (2010). 3D shape scanning with a time-of-flight camera. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 1173–1180).

Cui, Z., Gu, J., Shi, B., Tan, P., Kautz, J. (2017). Polarimetric multi-view stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1558–1567).

Di, Y., Manhardt, F., Wang, G., Ji, X., Navab, N., Tombari, F. (2021). SO-pose: Exploiting self-occlusion for direct 6D pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 12396–12405).

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM, 24*(6), 381–395.

Fließbach, T. (2012). *Elektrodynamik: Lehrbuch zur theoretischen physik II* (Vol. 2). Springer.

Gao, D., Li, Y., Ruhkamp, P., Skobleva, I., Wysock, M., Jung, H., Wang, P., Guridi, A., & Busam, B. (2022). Polarimetric pose prediction. In *Proceedings of the European conference on computer vision (ECCV)*.

Garcia, N. M., De Erausquin, I., Edmiston, C., & Gruev, V. (2015). Surface normal reconstruction using circularly polarized light. *Optics Express, 23*(11), 14391–14406.

He, Y., Huang, H., Fan, H., Chen, Q., Sun, J. (2021). FFB6D: A full flow bidirectional fusion network for 6D pose estimation. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

He, Y., Sun, W., Huang, H., Liu, J., Fan, H., Sun, J. (2020). PVN3D: A deep point-wise 3D keypoints voting network for 6DoF pose estimation. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., & Navab, N. (2013). Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes. In *Computer vision—ACCV 2012: 11th Asian conference on computer vision, Daejeon, Korea, November 5–9, 2012, revised selected papers, part I 11* (pp. 548–562).

Hodan, T., Barath, D., & Matas, J. (2020). EPOS: Estimating 6D pose of objects with symmetries. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11703–11712).

Hodaň, T., Matas, J., & Obdržálek, Š. (2016). On evaluation of 6D object pose estimation. In *European conference on computer vision* (pp. 606–619).

Huynh, C. P., Robles-Kelly, A., & Hancock, E. (2010). Shape and refractive index recovery from single-view polarisation images. In *2010 IEEE conference on computer vision and pattern recognition* (pp. 1229–1236).

Jiang, X., Li, D., Chen, H., Zheng, Y., Zhao, R., & Wu, L. (n.d.). *Uni6D: A unified CNN framework without projection breakdown for 6D pose estimation*.

Jung, H., Brasch, N., Leonardis, A., Navab, N., & Busam, B. (2021). Wild ToFu: Improving range and quality of indirect timeof-flight depth with RGB fusion in challenging environments. In *2021 International conference on 3D vision (3DV)* (pp. 239–248).

Jung, H., Ruhkamp, P., Zhai, G., Brasch, N., Li, Y., Verdie, Y., Song, J., Zhou, Y., Armagan, A., Ilic, S., & Busam, B. (n.d.). *On the importance of accurate geometry data for dense 3D vision tasks*.

Jung, H., Ruhkamp, P., Zhai, G., Brasch, N., Li, Y., Verdie, Y., Song, J., Zhou, Y., Armagan, A., Ilic, S., Leonardis, A., & Busam, B. (2022). Is my depth ground-truth good enough? HAMMER–Highly accurate multi-modal dataset for DEnse 3D scene regression. arXiv preprint arXiv:2205.04565

Kadambi, A., Taamazyan, V., Shi, B., & Raskar, R. (2017). Depth sensing using geometrically constrained polarization normals. *International Journal of Computer Vision, 125*(1–3), 34–51.

Kalra, A., Taamazyan, V., Rao, S. K., Venkataraman, K., Raskar, R., & Kadambi, A. (2020). Deep polarization cues for transparent object segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8602–8611).

Kingma, D. P., & Ba, J. (2014). ADAM: A method for stochastic optimization. arXiv preprint arXiv:1412.6980

Lei, C., Qi, C., Xie, J., Fan, N., Koltun, V., & Chen, Q. (2022). Shape from polarization for complex scenes in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 12632–12641).

Lepetit, V., Moreno-Noguer, F., & Fua, P. (2009). EP n P: An accurate O (n) aolution to the P n P problem. *International Journal of Computer Vision, 81*, 155–166.

Li, Z., Wang, G., & Ji, X. (2019). 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* (pp. 7678–7687).

Mo, N., Gan, W., Yokoya, N., & Chen, S. (2022). *ES6D: A computation efficient and symmetry-aware 6D pose regression framework*.

Nimier-David, M., Vicini, D., Zeltner, T., & Jakob, W. (2019). *Mitsuba 2: A retargetable forward and inverse renderer* (Vol. 38, pp. 1–17). New York: ACM.

Ost, J., Mannan, F., Thuerey, N., Knodt, J., & Heide, F. (2021). Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 2856–2865).

Park, K., Patten, T., & Vincze, M. (2019). Pix2Pose: Pixel-wise coordinate regression of objects for 6D pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7668–7677).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., & Antiga, L. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems* (vol. 32).

Shugurov, I., Li, F., Busam, B., & Ilic, S. (2022). *OSOP: A multi-stage one shot object pose estimation framework*.

Shugurov, I., Zakharov, S., & Ilic, S. (2021). DPODv2: Dense correspondence-based 6 DoF pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(11), 7417–7435.

Smith, W. A., Ramamoorthi, R., & Tozza, S. (2018). Height-from-polarisation with unknown lighting or albedo. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(12), 2875–2888.

Sock, J., Garcia-Hernando, G., Armagan, A., & Kim, T.-K. (2020). Introducing pose consistency and warp-alignment for self-supervised 6D object pose estimation in color images. In *2020 International conference on 3D vision (3DV)* (pp. 291–300).

Su, Y., Saleh, M., Fetzer, T., Rambach, J., Navab, N., Busam, B., Stricker, D., & Tombari, F. (2022). *ZebraPose: Coarse to fine surface encoding for 6DoF object pose estimation*.

Umeyama, S. (1991). Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 13*(04), 376–380.

Verdie, Y., Song, J., Mas, B., Busamm, B., Leonardis, A., & McDonagh, S. (2022). CroMo: Cross-modal learning for monocular depth estimation. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Wang, G., Manhardt, F., Liu, X., Ji, X., & Tombari, F. (2021a). Occlusion-aware self-supervised monocular 6D object pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, G., Manhardt, F., Shao, J., Ji, X., Navab, N., & Tombari, F. (2020). Self6D: Self-supervised monocular 6D object pose estimation. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, Part I 16* (pp. 108–125).

Wang, G., Manhardt, F., Tombari, F., & Ji, X. (2021b). 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* (pp. 16611–16621).

Wang, P., Jung, H., Li, Y., Shen, S., Srikanth, R.P., Garattoni, L., Meier, S., Navab, N., & Busam, B. (2022). 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)*.

Wang, P., Manhardt, F., Minciullo, L., Garattoni, L., Meier, S., Navab, N., & Busam, B. (2021c). DemoGrasp: Few-shot learning for robotic grasping with human demonstration. In *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 5733–5740).

Yisheng, H., Yao, W., Haoqiang, F., Qifeng, C., & Jian, S. (2022). *Fs6d: Few-shot 6d pose estimation of novel objects*.

Yu, Y., Zhu, D., & Smith, W. A. (2017). Shape-from-polarisation: A nonlinear least squares approach. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 2969–2976).

Zakharov, S., Shugurov, I., & Ilic, S. (2019). DPOD: 6D pose pbject detector and refiner. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1941–1950).

Zhou, Y., Barnes, C., Lu, J., Yang, J., & Li, H. (2019). On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5745–5753).

Zhu, D., & Smith, W.A. (2019). Depth from a polarisation + RGB stereo pair. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7586–7595).

Zou, S., Zuo, X., Qian, Y.,Wang, S., Xu, C., Gong, M., & Cheng, L. (2020). 3D human shape reconstruction from a polarization image. In *European conference on computer vision* (pp. 351–368).