Accepted for publication at German Conference on Pattern Recognition (GCPR), 2024.
Submitted version. This preprint has not undergone peer review or any post-submission improvements or corrections.
The final version of the paper will be published in Proc. of GCPR by Springer.

Physically Plausible Object Pose Refinement in Cluttered Scenes

Michael Strecke[0000-0002-0322-0653] and Joerg Stueckler[0000-0002-2328-4363]

Embodied Vision Group, Max Planck Institute for Intelligent Systems, Tuebingen, Germany
{firstname.lastname}@tuebingen.mpg.de

Abstract. Estimating the 6-DoF pose of objects from images is a fundamental task in computer vision and a prerequisite for downstream tasks like augmented reality or robotic grasping applications. This task becomes particularly challenging in cluttered scenes, when many objects are present in the image in close proximity and occlude one another. However, the close proximity between objects also provides additional cues about the objects, as objects in physically plausible scenes do not intersect one another and thus occluding objects constrain the ones they occlude. We present a novel approach for utilizing this information in 6-DoF object pose refinement of known objects. Our formulation extends RAFT-based pose refinement to reduce penetrations between objects to a large degree and leads to more plausible object poses with less penetrations. We evaluate our approach quantitatively and qualitatively on two benchmark datasets, demonstrate improvements over baselines, and will make the source code of our approach publicly available to foster future research in this area.

1 Introduction

Estimating the 6 degree of freedom (6-DoF) pose, *i.e.*, position and orientation, of known objects from camera images is a fundamental task in computer vision and a prerequisite for downstream tasks like augmented reality or robotic object grasping applications. Many works have attempted to solve this problem using classical approaches for feature matching (*e.g.*, [15]) or, more recently, by exploiting advances in deep learning [27,6,5,24]. Methods like [11,14,22] target improving initial pose prediction by a neural network with iterative refinement of the pose estimates.

While these methods often yield accurate results for well-visible objects, clutter with severe occlusions often still poses a challenge, if depth and texture features do not provide enough cues for accurately estimating the object pose. We argue that the proximity of objects in cluttered scenes does not only provide a challenge for pose estimation, but also an additional cue as well-visible objects constrain the poses of occluded objects as we illustrate in Fig. 1.

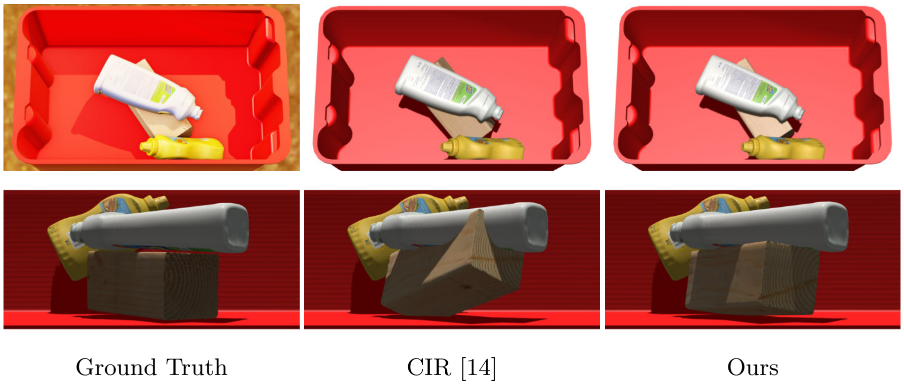


Fig. 1: Physically plausible object pose refinement in cluttered scenes. Given an RGB-D input image (color input top left), in which the wood block is occluded by the bleach cleanser bottle, coupled iterative refinement (CIR) [14] yields an implausible result with penetrations of the wood block with the bleach cleanser bottle and the background tote (center). Our approach (right) resolves these penetrations for more accurate and physically plausible object poses.

Many previous methods separate the object pose estimation task into several stages: 2D object detection, pose estimation/regression and optionally pose refinement. In this work, we focus on the refinement stage of the pose estimation task. We build upon the refinement algorithm of Coupled Iterative Refinement (CIR) [14], which in its basic form considers each object in isolation after the object detection stage. By contrast, we follow a line of physics-informed pose refinement approaches [2,3,23], extend CIR's pose refinement algorithm, and optimize the poses of all detected objects *jointly*, taking into account that no two objects share the same 3D space by physical plausibility. We propose a novel penetration penalty based on a signed distance function representation of the objects.

We evaluate our approach on two multi-object pose estimation benchmarks and demonstrate slight improvements over CIR in pose estimation and strong reductions of inter-object penetrations. Our approach also reduces penetrations better than two baseline approaches (MoreFusion [23] and SporeAgent [3]). Wrt. the latter state-of-the-art reinforcement learning approach (SporeAgent), our approach also demonstrates better generalization capabilities to novel scenes using the same underlying detector.

In summary, we contribute the following:

- We propose a novel intersection penalty for pose refinement with deep learning based correspondences that resolves penetrations between objects and leads to more accurate and physically plausible object pose estimates.
- We evaluate the accuracy of the recovered object poses and their physical plausibility in terms of interpenetration on two benchmark datasets. We demonstrate improved performance in penetration resolution and generaliza-

tion wrt the state-of-the-art pose refinement methods that also use physical plausibility.
- We will make the source code of our method publicly available to foster future research in this area.

2 Related Work

Object pose estimation. Early works such as, e.g., [15], estimate keypoint correspondences between a 2D image and the 3D object model and computed the pose from the 2D-3D keypoint matches using the Perspective-n-Point (PnP) algorithm [10,12]. In recent years, significant improvements in accuracy have been achieved using deep learning based approaches [27,5,24,25]. A complementary line of works attempts to refine the object poses obtained by an initial prediction by render-and-compare, matching a render of the object in the currently estimated poses using learned features [13,11,14] or by differentiable rendering [22].

Object pose refinement with physical plausibility. Most previous works separate the detection and pose estimation stages and consider objects in isolation during pose estimation. Notable exceptions are [11,23,2,3], where [11] jointly estimates a set of camera poses and a collection of objects in an object-level bundle adjustment setting for a globally consistent scene. MoreFusion [23] also considers objects jointly, but like our method tries to estimate poses from a single image. It uses volumetric pose prediction and combines the iterative closest point algorithm with iterative collision checks for refinement using occupancy grids. VeREFINE [2] combines pose verification with physics-guided iterative refinement by embedding objects in simulation. As some object configurations might lead to diverging simulation results, the observation-based verification is needed for robustness. Our approach directly resolves penetrations between objects. Penetrations are typically one of the reasons for diverging behavior in simulation. SporeAgent [3] trains a reinforcement learning agent for object pose refinement. The approach only utilizes object geometry and definitions of plausible poses in which objects do not intersect or float in the air as in [1]. By contrast, we propose to embed a penetration constraint in an optical flow-based refinement optimization, allowing us to exploit both depth-augmented 2D image features and physical plausibility for 6-DoF pose prediction.

3 Method

Our method takes a single RGB-D image as input and outputs a set of object detections with associated pose estimates. We assume the shape and texture of objects that are detected to be known in advance, *i.e.*, we assume available textured 3D meshes for the objects. We build our method upon Coupled Iterative Refinement (CIR) by Lipson *et al.* [14], which targets the same task. Our key contribution is an extension of CIR's refinement algorithm. We add a term for avoiding penetrations between objects in their pose estimates to CIR's optimization objective and thus receive physically more plausible scene configurations.

3.1 Preliminaries

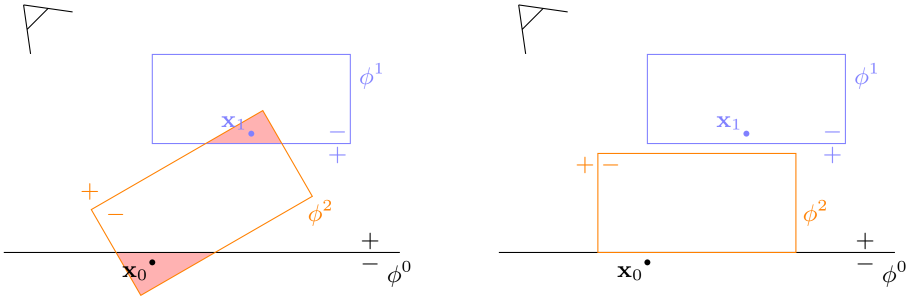
CIR [14] separates the pose estimation task into the 3 stages object detection, pose initialization, and pose refinement. It builds upon Cosypose [11] for detection and pose initialization. Objects are detected by Mask R-CNN [4] and cropped to their bounding boxes. The bounding boxes are used to initialize the object translation $\mathbf{t}_{\text{bbox}}$, which aligns the bounding box of the 3D model with the detected object mask. A rendered image of the object in this estimated pose is fed to a ResNet-based architecture (EfficientNet [20]) together with the image crop. The result is a rotation $\mathbf{R}$ and a translation increment $\mathit{\Delta}\mathbf{t}$, yielding the initial object pose $\mathbf{G}_0^{(0)} \in SE(3)$. After initialization, CIR refines the pose estimates in a render-and-compare approach. Given the textured 3D meshes of the objects and viewpoints parameterized by extrinsic and intrinsic parameters $\mathbf{G}_i \in SE(3)$ and $\mathbf{K}_i$, respectively, images and depth maps of the objects can be rendered using PyTorch3D [18]. Here $\mathbf{G}_i$ denotes the object pose in camera coordinates. Lipson *et al.* [14] then proceed by denoting the pose for the image by $\mathbf{G}_0$ and producing a set of renders with poses $\{\mathbf{G}_1, \ldots, \mathbf{G}_N\}$. They then compute correspondence features between the renders and the input image using RAFT [21] and use these in a bidirectional perspective-n-point (BD-PnP) optimization to refine the pose. We refer to [14] for details and denote the BD-PnP energy for object i (eq. (7) in [14]) as $\mathbf{E}_{\text{BD}-\text{PnP}}(\mathbf{G}_0^i)$ in the following.

3.2 Resolving Penetrations

The formulation from [14] relies on feature matches between rendered objects and the input image crops. In the presence of other objects, occlusions and regions of uniform texture can make the correspondences unreliable and thus lead to bad pose estimates (see Fig. 1). This becomes especially challenging in cluttered scenes where objects are in close proximity, *e.g.*, when multiple objects lie in a box like in the Synpick dataset [17]. However, we can reason about physically plausible scene configurations in this setting: we know that no two objects occupy the same space in 3D. We include this information in an additional data term for retrieving more accurate and more plausible object poses.

Towards this goal, we represent the geometry of the objects by volumetric signed distance functions (SDFs) $\phi^i : \mathbb{R}^3 \to \mathbb{R}$. Evaluating an SDF ϕ at a point $\mathbf{x} \in \mathbb{R}^3$ gives the distance to the closest surface with a sign indicating whether $\mathbf{x}$ lies inside ($\phi^i(\mathbf{x}) < 0$) or outside ($\phi^i(\mathbf{x}) > 0$) the object. The surface is thus represented *implicitly* as the zero level set of the SDF ϕ. We propose the following energy term to resolve penetrations between M objects with poses $\left\{\mathbf{G}_0^1, \ldots, \mathbf{G}_0^M\right\}$:

$$\mathcal{E}_{\text{inter}}\left(\left\{\mathbf{G}_0^1, \ldots, \mathbf{G}_0^M\right\}\right) = \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \mathbf{E}_{\text{inter}}(\mathbf{G}_0^i, \mathbf{G}_0^j), \tag{1}$$



$\ell_{\text{inter}}(\mathbf{x}_0) = \phi^0(\mathbf{x}_0) + \phi^2(\widetilde{\mathbf{x}}_0) < 0$
$\ell_{\text{inter}}(\mathbf{x}_1) = \phi^1(\mathbf{x}_1) + \phi^2(\widetilde{\mathbf{x}}_1) < 0$

$\ell_{\text{inter}}(\mathbf{x}_0) = \phi^0(\mathbf{x}_0) + \phi^2(\widetilde{\mathbf{x}}_0) = 0$
$\ell_{\text{inter}}(\mathbf{x}_1) = \phi^1(\mathbf{x}_1) + \phi^2(\widetilde{\mathbf{x}}_1) > 0$

Fig. 2: The intersection constraint in Eq. (2) penalizes penetrations between objects. In the left illustration, the volume of ϕ^2 overlaps with the volumes ϕ^0 and ϕ^1, yielding negative values for $\ell_{\text{inter}}(\mathbf{x}_i)$ for the points $\mathbf{x}_0$ and $\mathbf{x}_1$ defined in the coordinate systems of objects 0 and 1, respectively ($\widetilde{\mathbf{x}}_i = \mathbf{G}_0^2 \mathbf{G}_0^i \mathbf{x}_i$ is the point transformed to the coordinate system of ϕ^2). Thus $\mathbf{E}_{\text{inter}}$ is positive. In the right illustration, the penetration is resolved by a rotation of the object ϕ^2, yielding $\ell_{\text{inter}}(\mathbf{x}_i) \geq 0$ (where equality is attained for objects in contact) and thus $\mathbf{E}_{\text{inter}} = 0$.

where

$$\mathbf{E}_{\text{inter}}(\mathbf{G}_0^i, \mathbf{G}_0^j) = \int_{\Omega_i \cap \Omega_j} \frac{1}{2} \min\Big\{0, -\underbrace{\left(\phi^i\left(\mathbf{x}\right) + \phi^j\left(\mathbf{G}_0^j {\mathbf{G}_0^i}^{-1} \mathbf{x}\right)\right)}_{=:\ell_{\text{inter}}(\mathbf{x})}\Big\}^2 \mathrm{d}\mathbf{x}, \qquad (2)$$

where $\mathbf{x}$ is defined in the coordinate system of object i and Ω_i and Ω_j are the 3D bounding volumes containing objects i and j, respectively.

Intuitively, we argue that a point $\mathbf{x}$ that is inside one object by a certain distance $|\phi(\mathbf{x})|$ (where *inside* is indicated by $\phi(\mathbf{x}) < 0$) needs to be outside all other objects by at least $|\phi(\mathbf{x})|$. Thus, the sum of any two SDF values, ℓ_{inter} in Eq. (2), needs to be positive. Equation (2) penalizes negative values of ℓ_{inter} in the area in which the bounding volumes of objects i and j overlap. Figure 2 illustrates the values of ℓ_{inter} and $\mathbf{E}_{\text{inter}}$ in Eq. (2) for cases with and without penetration between the objects. A similar formulation has been used in [19] to close object shapes for 3D mapping. In contrast to their work, we assume the object shapes given by CAD models and refine the object poses to resolve penetrations.

We optimize the combined objective

$$\mathcal{E}\left(\left\{\mathbf{G}_0^1, \ldots, \mathbf{G}_0^M\right\}\right) = \mathcal{E}_{\text{BD-PnP}}\left(\left\{\mathbf{G}_0^1, \ldots, \mathbf{G}_0^M\right\}\right) + \mathcal{E}_{\text{inter}}\left(\left\{\mathbf{G}_0^1, \ldots, \mathbf{G}_0^M\right\}\right), \quad (3)$$

where $\mathcal{E}_{\text{BD-PnP}}\left(\left\{\mathbf{G}_0^1, \ldots, \mathbf{G}_0^M\right\}\right) = \sum_{i=1}^M \mathbf{E}_{\text{BD-PnP}}(\mathbf{G}_0^i)$. For minimizing Eq. (3), we follow [14] and linearize Eq. (3) using the current pose estimate and then perform a fixed number of Gauss-Newton updates. Each update step produces a pose update $\boldsymbol{\delta\xi} \in \mathfrak{se}(3)$, which is used to update the pose on the SE(3) manifold $\mathbf{G}_0^{(t+1)} = \exp(\boldsymbol{\delta\xi}) \cdot \mathbf{G}_0^{(t)}$.

3.3 Implementation Details

To implement our intersection penalty in Eq. (2), we compute discrete volumetric signed distance function (SDF) grids at a resolution of 64^3 from the 3D object meshes using mesh2sdf[1] [26]. We then discretize Eq. (2) by turning the integral into a sum over all voxels in object i that when transformed to object j's coordinate frame are inside the bounding volume of object j. For computing the Gauss-Newton updates, gradients of the SDFs are required, which we compute using central finite differences on the discrete grid. For non-integral point queries to SDF or gradient volumes (either by non-integral sampling or after transformation to the other object volume), we perform triliniear interpolation to compute the corresponding value. Following the CIR evalution pipeline, we use 4 outer loops, 40 inner loops, and 10 solver steps in our experiments. In some cases, the EfficientNet initialization is too far off, leading to objects being estimated on the wrong sides of each others. In these cases, our penetration constraint actually prevents CIR from correcting the pose. We thus use the first outer loop as warmup iteration without the penetration penalty in our experiments.

4 Experiments

We evaluate our approach on two publicly available datasets. The main dataset used in our evaluation is SynPick [17]. It consists of sequences in which the 21 objects from the YCB Video Dataset [27] are initially in a box and then manipulated with a suction cup gripper. The data is stored in the format specified in the BOP benchmark suite [8,9]. Additionally, we evaluate on the YCB Video [27] dataset from the BOP benchmark [8,9]. In the following, we evaluate penetration and pose accuracy and show qualiative and quantitative results of our method and baselines. Details on the evaluation measures and further results can be found in the supplementary material, including a video.

4.1 Qualitative Results

We show qualitative results on the Synpick dataset in Fig. 3. In these cluttered scenes our approach manages to reduce penetrations and thus yields results that match the ground truth better than CIR's estimates. Note that in many of these cases, there is only little difference observable in the input view, but alternate views (every second row) highlights how our approach improves object poses. In Fig. 4, we show results on the YCB Video dataset. While for some images (top row and bottom left), our method manages to reduce penetrations and use that information for more accurate pose estimates, in other cases (bottom right), our optimization runs into local optima for the pose due to bad initialization while penetrations are reduced well.

[1] https://github.com/wang-ps/mesh2sdf

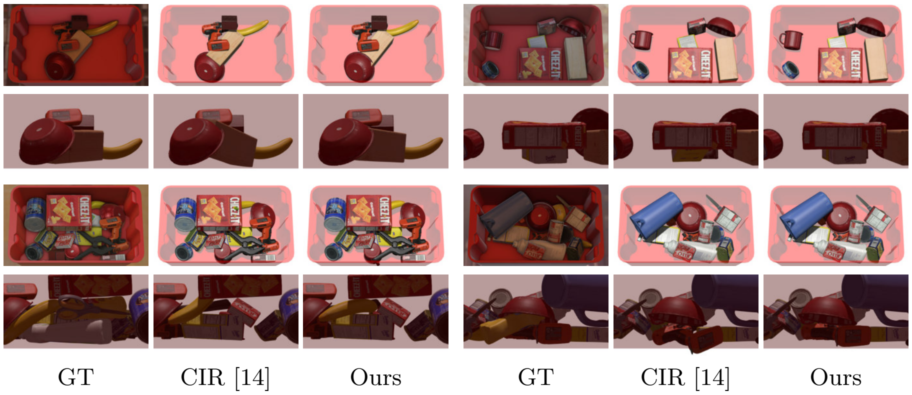


Fig. 3: Qualitative results on the Synpick dataset. Our method manages to reduce penetrations between the objects and the background box, leading to more accurate results for occluded objects (wood block in the top left, sugar box in the top right, banana in the bottom left and power drill in the bottom right). The average AD scores improve from 0.014 to 0.013 (top left), 0.031 to 0.009 (top right) and from 0.031 to 0.023 (bottom right), and remain approximately the same at 0.043 for the bottom left example.

4.2 Evaluation Measures

Penetration. As the key focus of our work lies on improving the physical plausibility of object pose estimates, we evaluate the penetration volume between objects as a measure of physical plausibility. We compute the penetration volume for object i as the volume it shares with other objects:

$$\mathrm{PEN}(i) = \sum_{\mathbf{v}\in\Omega_i} \left[\exists j \in \{1,\dots,M\} : \left(\phi^i(\mathbf{v}) < 0\right) \wedge \left(\phi^j\left(\mathbf{G}^j{\mathbf{G}^i}^{-1}\mathbf{v}\right) < 0\right)\right] \mathrm{vol}(\mathbf{v}), \tag{4}$$

where $[\cdot]$ denotes the Iverson bracket and $\mathrm{vol}(\mathbf{v})$ is the volume of voxel $\mathbf{v}$. In other words, Eq. (4) sums the volume of voxels which lie inside object i that also map to points inside another volume j. We evaluate the absolute penetration in mm^3 as well as the relative penetration $\mathrm{PEN}(i)/\mathrm{vol}(i)$, where $\mathrm{vol}(i) = \sum_{\mathbf{v}\in\Omega_i}[\phi^i(\mathbf{v}) < 0]$ is the volume of object i. The penetration scores are then averages across all objects in an image and across all images in the dataset.

Pose evaluation. For evaluating pose accuracy, we compute the recall scores for the error measure AD [7] at threshold of 2%, 5% and 10% of the object diameter using the BOP toolkit[2]. Additionally we compute the AD, ADI, and ADD area-under-the curve on the YCB Video dataset, following [3].

[2] https://github.com/thodan/bop_toolkit

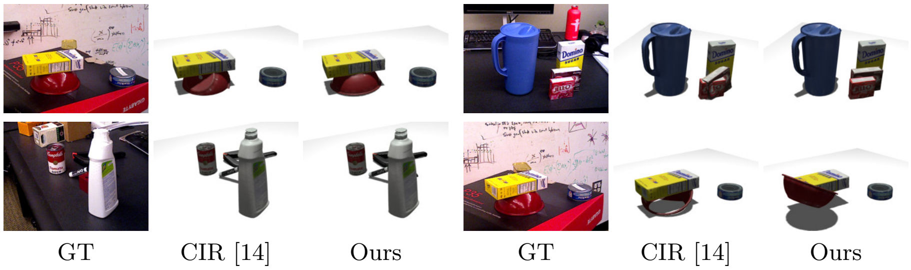


Fig. 4: Qualitative results on YCB Video. Our approach reduces penetrations and improves the object poses of the bowl (top left), the box (top right), and the clamp (bottom left). The average AD scores improve from 0.04 to 0.01 (top left), 0.05 to 0.01 (top right), and 0.03 to 0.02 (bottom left). Bottom right: limit case. A bad initialization can lead to local optima for our optimization, reducing the penetration with the background plane but ending in a wrong pose for the bowl with an increase in the AD scores from 0.05 to 0.12.

4.3 Quantitative Results

We evaluate our method on frames not showing the gripper from the "Test Pick Targeted" and "Test Pick Untargeted" splits of the Synpick dataset and the default test splits of the other BOP datasets as specified by the provided target files. For finding the frames not showing the gripper in Synpick we choose the first frame of every test sequence and every frame in the sequence in which the minimum depth equals the minimum depth of the empty box at the end of the sequence. We compare to the base "vanilla" CIR [14], EfficientNet [20] (EN), MoreFusion [23] (MF), and SporeAgent [3]. MoreFusion includes physical plausibility in object pose estimation by volumetric pose prediction and refinement through iterative closest points (ICP) and iterative collision checks (ICC), and we always evaluate the "full" approach, including ICP and ICC. SporeAgent trains a reinforcement learning agent to refine poses for physical plausibility by avoiding penetrations and floating objects. For SporeAgent, we compare two variants on YCB Video: using PoseCNN [27] for initialization (SP-PC) and using EfficienNet [20] for initialization (SP-EN) to have a fair comparison to our method and CIR, which also use EfficientNet for initialization. Unfortunately, we did not achieve good results for CIR when directly training its full pipeline on Synpick with the same settings as for YCB Video. We thus evaluate EfficientNet, CIR, SporeAgent and MoreFusion models on Synpick which are consistently trained on YCB Video to test their generalization capabilities. For SporeAgent, we only compare to the SA-EN variant as PoseCNN predictions are not available for Synpick.

As the background box is provided for the Synpick dataset, we compute an SDF model for it and use it as additional penetration constraint in our approach and compare to not using this additional cue. Similarly, we use the annotated planes from SporeAgent [3] as background cue for the YCB Video dataset. The version of the YCB Video dataset provided in the BOP benchmark [8,9] contains a shifted coordinate system for YCB objects, which also affects the ground truth. We thus train MoreFusion on the BOP version of the YCB Video dataset and do not use its original pretrained model for this dataset. As MoreFusion evaluated their approach using ground truth masks and we found significantly lower performance when using MoreFusion with Mask R-CNN detections or EfficientNet for initialization, we compare to MoreFusion using ground truth masks. When using the Mask R-CNN detector for our method, CIR, EN, and SA-EN, we set the detection threshold to 0.95 as we discovered a lower threshold would sometimes lead to multiple detections of the same object in the same position, which is an implausible initialization for our method.

As we process all objects in an image in parallel, the number of objects is naturally limited by the available GPU memory. We observe that we can process a maximum of 17 objects on a GPU with 80GB VRAM. Our method requires on average 28.5 s while CIR needs 14.9 s per image for 100 frames from the Synpick "Test Pick Targeted" split on an NVIDIA H100 GPU.

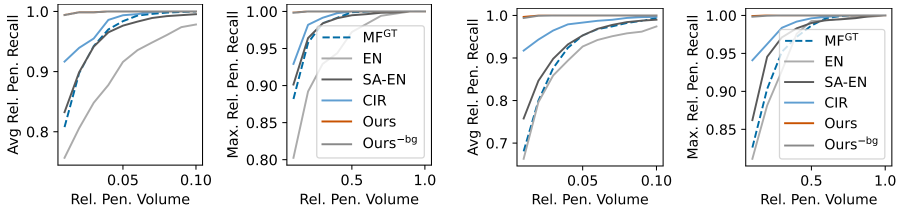
Penetration evaluation. We report penetration evaluation scores on the Synpick dataset in Table 1. Our method reduces the absolute and relative penetration between the objects compared to all baselines. We observe that our approach outperforms EfficientNet (EN) [20], CIR [14], MoreFusion (MF) [23], and SporeAgent [3] with EfficientNet initialization (SA-EN) by approximately one order of magnitude in both absolute and relative penetration. Note that in some cases, not using the background box improves the penetration scores. This is due to the evaluation only considering the detected objects without the background box. Resolving penetrations with the box might thus lead to increased penetration with other objects. In Fig. 5, we plot recall curves for the average and maximum relative penetration. A recall of 1.0 for a threshold θ means that all images in the dataset have an average/maximum relative penetration of at most θ. Our approach achieves higher recall scores than EN, CIR, MF, and SA-EN across the evaluation range. We further evaluate the penetration on the YCB Video dataset from the BOP benchmark in Fig. 6. In addition to the approaches we compared to on Synpick, we compare to SporeAgent with PoseCNN (SA-PC) initialization. Our approach resolves penetrations better than all comparison approaches, while CIR and the two SporeAgent variants perform approximately on-par to each others.

Pose evaluation. While the focus of our method lies on physical plausibility, *i.e.*, resolving penetrations between objects, we also evaluate the pose accuracy on Synpick and YCB Video dataset. We show pose evaluation results in Tables 2 and 3. Our method improves CIR's results slightly on the Synpick dataset (*cf*. Table 2) and the YCB Video dataset (*cf*. Table 3). On Synpick, CIR and our approach achieve better pose accuracy than SporeAgent and MoreFusion, even when we use Mask R-CNN detections for our method and ground truth masks for MoreFusion (*cf*. Table 2), indicating better generalization capabilities for CIR and our method. Note that SporeAgent does not use texture for refinement,

Table 1: Penetration evaluation on the Synpick dataset. The absolute penetration error is given in mm^3 and the relative penetration is a fraction of the object volume. Lower values are better. We outperform all baselines, also indicating that our approach generalizes well to the unseen data in Synpick.

		Abs. Pen.		Rel. Pen.	
		Avg.	Max.	Avg.	Max.
Test Pick Untargeted	MF$^{\mathrm{GT}}$	$6.46 \cdot 10^{3}$	$5.39 \cdot 10^{5}$	$1.27 \cdot 10^{-2}$	$7.84 \cdot 10^{-1}$
	EN	$8.14 \cdot 10^{3}$	$5.29 \cdot 10^{5}$	$1.56 \cdot 10^{-2}$	$9.43 \cdot 10^{-1}$
	SA-EN	$3.56 \cdot 10^{3}$	$4.42 \cdot 10^{5}$	$9.54 \cdot 10^{-3}$	$9.25 \cdot 10^{-1}$
	CIR	$1.47 \cdot 10^{3}$	$4.54 \cdot 10^{5}$	$2.95 \cdot 10^{-3}$	$7.12 \cdot 10^{-1}$
	Ours	$\underline{1.26 \cdot 10^{2}}$	$\underline{1.56 \cdot 10^{4}}$	$\underline{4.83 \cdot 10^{-4}}$	$\mathbf{1.92 \cdot 10^{-1}}$
	Ours$^{-\mathrm{bg}}$	$\mathbf{1.11 \cdot 10^{2}}$	$\mathbf{1.40 \cdot 10^{4}}$	$\mathbf{4.25 \cdot 10^{-4}}$	$\underline{3.00 \cdot 10^{-1}}$
Test Pick Targeted	MF$^{\mathrm{GT}}$	$3.09 \cdot 10^{3}$	$2.37 \cdot 10^{5}$	$9.10 \cdot 10^{-3}$	$7.28 \cdot 10^{-1}$
	EN	$6.91 \cdot 10^{3}$	$8.14 \cdot 10^{5}$	$1.46 \cdot 10^{-2}$	$8.16 \cdot 10^{-1}$
	SA-EN	$2.72 \cdot 10^{3}$	$3.08 \cdot 10^{5}$	$7.48 \cdot 10^{-3}$	$8.87 \cdot 10^{-1}$
	CIR	$8.78 \cdot 10^{2}$	$3.10 \cdot 10^{5}$	$3.32 \cdot 10^{-3}$	$4.77 \cdot 10^{-1}$
	Ours	$\underline{7.25 \cdot 10^{1}}$	$\underline{1.14 \cdot 10^{4}}$	$\mathbf{4.31 \cdot 10^{-4}}$	$\mathbf{1.10 \cdot 10^{-1}}$
	Ours$^{-\mathrm{bg}}$	$\mathbf{7.11 \cdot 10^{1}}$	$\mathbf{1.11 \cdot 10^{4}}$	$\underline{4.79 \cdot 10^{-4}}$	$\underline{1.64 \cdot 10^{-1}}$

while our approach combines geometry and texture cues. On YCB Video, for recall scores on the AD measure (Table 3 left), our method and CIR outperform the comparison approaches (except for MF using ground truth segmentation). When restricting the evaluation to objects with a ground-truth visibility of less than 50%, we observe a larger improvement by our method for the recall AD measure below two percent of object diameter (AD< 0.02), indicating that our method manages to improve the accuracy when objects are not well visible. We also observe this trend in Fig. 7. While our method performs on-par with the baseline CIR in most cases, we see a larger gap for AD< 0.02 below a visibility of 60%. Our method further consistently outperforms SA-EN with the same initialization as our method and only falls behind SA-PC and MF at the



(a) Test Pick Targeted (b) Test Pick Untargeted

Fig. 5: Penetration recall curves on the Synpick dataset. A recall of 1.0 means that the average/maximum penetration in all frames is below the threshold on the x-axis. Our method resolves a large fraction of penetration between objects, even when the background box is not used for optimization (Ours$^{-\mathrm{box}}$). SA-EN and MF do not resolve penetrations as well.

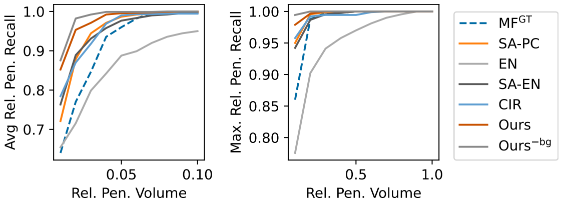


Fig. 6: Penetration recall curves on the YCB Video dataset. A recall of 1.0 means that the avg./max. penetration in all frames is below the threshold on the x-axis. Our approach demonstrates the best performance.

Table 2: Pose evaluation results on Synpick. We report average recall scores for the AD measure at different thresholds. Higher values are better. In these cluttered scenes, our method consistently improves the results obtained by CIR. Using the background box as additional cue does not have a large effect on the results. EN, SA-EN, and MF with GT masks perform worse than our approach, indicating that our approach generalizes better to the unseen dataset. All approaches are trained on YCB Video.

	Test Pick Untargeted			Test Pick Targeted		
	AD	AD	AD	AD	AD	AD
	$< 0.02d$	$< 0.05d$	$< 0.1d$	$< 0.02d$	$< 0.05d$	$< 0.1d$
MF$^{\text{GT}}$	0.408	0.511	0.582	0.394	0.494	0.595
EN	0.007	0.052	0.159	0.009	0.065	0.173
SA-EN	0.125	0.423	0.601	0.150	0.433	0.579
CIR	0.670	0.730	0.748	0.622	0.702	0.730
Ours	0.674	**0.733**	0.748	**0.628**	**0.708**	**0.733**
Ours$^{-\text{bg}}$	**0.675**	0.732	**0.748**	0.626	0.707	0.732

higher distance thresholds of 5 and 10% object diameter. Note that SA-PC used PoseCNN detections with different recall properties than EN, and MF used ground truth detections and is thus expected to perform better. Under the AUC measure for the ADD, AD, and ADI measures (Table 3 right), our method and CIR achieves slightly lower scores than SA-PC. Note that this evaluation uses the pipeline from PoseCNN [27], which differs in some points from the evaluation using the BOP toolkit. In Table 3 (left), the AD scores are normalized by object diameter and poses deviating more than 10% of the object diameter from the ground truth are rejected as invalid. By contrast the PoseCNN evaluation in Table 3 (right) does not normalize the poses and rejects poses further than 10cm away from the ground truth, *i.e.*, being less strict.

5 Conclusion

We develop a novel approach to 6-DoF object pose refinement that includes a penetration constraint directly in a neural render-and-compare pipeline. Our method improves the results of the base approach CIR [14], especially in clut-

Table 3: Pose evaluation on the YCB Video Dataset. Our method slightly improves over the CIR baseline for AD recall scores. The improvement becomes more noticeable at the lowest distance threshold when focusing on objects with a visibility lower than 50% (highlighted in green). In the AUC recall evaluation, we perform on-par with the baseline CIR and outperform the SA-EN. Note that SA-PC uses a different detector.

Recall	all objects			visibility < 0.5		
	AD $< 0.02d$	AD $< 0.05d$	AD $< 0.1d$	AD $< 0.02d$	AD $< 0.05d$	AD $< 0.1d$
$\mathrm{MF}^{\mathrm{GT}}$	0.628	0.865	**0.940**	**0.379**	**0.500**	**0.603**
SA-PC	0.369	0.784	0.867	0.224	0.379	0.388
EN	0.057	0.260	0.522	0.000	0.017	0.043
SA-EN	0.401	0.811	0.886	0.181	0.216	0.216
CIR	0.704	0.904	0.915	0.147	0.224	0.233
Ours	**0.708**	**0.908**	0.916	0.216	0.224	0.233
$\mathrm{Ours}^{-\mathrm{bg}}$	0.703	0.906	0.916	0.147	0.224	0.233

	ADD AUC	AD AUC	ADI AUC
$\mathrm{MF}^{\mathrm{GT}}$	**84.3**	**93.7**	**97.0**
SA-PC	79.0	88.8	93.6
EN	60.1	69.7	79.7
SA-EN	73.9	84.5	90.1
CIR	75.5	86.3	90.9
Ours	75.6	86.4	90.9
$\mathrm{Ours}^{-\mathrm{bg}}$	75.5	86.3	90.9

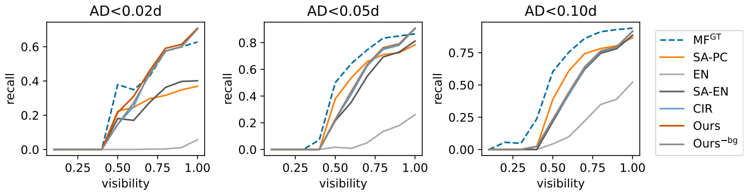


Fig. 7: Pose recall depending on visibility on YCB Video. Our method performs on-par with the baseline CIR and outperforms SA-EN consistently. At higher distance thresholds, SA-PC and MF with ground-truth segmentation outperform our method. We observe a larger improvement of our method over CIR at the lowest distance threshold when the object visibility is lower than 60%.

tered scenes when the proximity of the objects aids pose estimation. Compared to other approaches exploiting physical plausibility for 6 DoF pose estimation (MoreFusion and SporeAgent), we observe that we resolve a larger fraction of the penetrations between objects. Analyzing the pose accuracy, we observe that our approach performs on par with the base approach, yielding slight improvements in cluttered scenes when the surroundings of objects are helpful for pose estimation. Limitations of our method are currently memory and runtime requirements. In future work, further improvements in memory consumption might be achievable by using neural implicit representations like DeepSDF [16] for representing the objects. This might however come at the cost of higher computational effort for querying SDF values compared to trilinear interpolation lookups in our precomputed grids. We anticipate that our results will be helpful in downstream tasks, as the reduced penetrations allow for easier embedding of the estimated objects, *e.g.*, in physical simulations.

Acknowledgements

This work was supported by Cyber Valley, the Max Planck Society, and Amazon through the Max Planck Amazon Science Hub. This work was partially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project no. 466606396 (STU 771/1-1).

References

1. Bauer, D., Patten, T., Vincze, M.: Physical plausibility of 6D pose estimates in scenes of static rigid objects. In: Bartoli, A., Fusiello, A. (eds.) Computer Vision – ECCV 2020 Workshops. pp. 648–662. Springer International Publishing, Cham (2020)
2. Bauer, D., Patten, T., Vincze, M.: VeREFINE: Integrating object pose verification with physics-guided iterative refinement. IEEE Robotics and Automation Letters **5**(3), 4289–4296 (jul 2020). https://doi.org/10.1109/lra.2020.2996059
3. Bauer, D., Patten, T., Vincze, M.: SporeAgent: Reinforced scene-level plausibility for object pose refinement. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 654–662 (January 2022)
4. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988. IEEE (oct 2017). https://doi.org/10.1109/iccv.2017.322
5. He, Y., Huang, H., Fan, H., Chen, Q., Sun, J.: FFB6D: A full flow bidirectional fusion network for 6d pose estimation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2021). https://doi.org/10.1109/cvpr46437.2021.00302
6. He, Y., Sun, W., Huang, H., Liu, J., Fan, H., Sun, J.: PVN3D: A deep pointwise 3d keypoints voting network for 6dof pose estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2020). https://doi.org/10.1109/cvpr42600.2020.01165
7. Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N.: Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes, pp. 548–562. Springer Berlin Heidelberg (2013). https://doi.org/10.1007/978-3-642-37331-2_42
8. Hodaň, T., Michel, F., Brachmann, E., Kehl, W., Buch, A.G., Kraft, D., Drost, B., Vidal, J., Ihrke, S., Zabulis, X., Sahin, C., Manhardt, F., Tombari, F., Kim, T.K., Matas, J., Rother, C.: BOP: Benchmark for 6D object pose estimation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018, vol. 11214, pp. 19–35. Springer International Publishing (2018). https://doi.org/10.1007/978-3-030-01249-6_2, https://doi.org/10.1007/978-3-030-01249-6_2
9. Hodaň, T., Sundermeyer, M., Drost, B., Labbé, Y., Brachmann, E., Michel, F., Rother, C., Matas, J.: BOP challenge 2020 on 6d object localization. In: Computer Vision – ECCV 2020 Workshops, pp. 577–594. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-66096-3_39
10. Horaud, R., Conio, B., Leboulleux, O., Lacolle, B.: An analytic solution for the perspective 4-point problem. Computer Vision, Graphics, and Image Processing **47**(1), 33–44 (Jul 1989). https://doi.org/10.1016/0734-189x(89)90052-2, https://inria.hal.science/inria-00589990/document

11. Labbé, Y., Carpentier, J., Aubry, M., Sivic, J.: CosyPose: Consistent multi-view multi-object 6d pose estimation. In: Computer Vision – ECCV 2020, pp. 574–591. Springer International Publishing (Aug 2020). https://doi.org/10.1007/978-3-030-58520-4_34, http://arxiv.org/abs/2008.08465, arXiv:2008.08465 [cs] type: article
12. Lepetit, V., Moreno-Noguer, F., Fua, P.: EPnP: An accurate o(n) solution to the PnP problem. International Journal of Computer Vision **81**(2), 155–166 (jul 2008). https://doi.org/10.1007/s11263-008-0152-6, https://doi.org/10.1007/s11263-008-0152-6
13. Li, Y., Wang, G., Ji, X., Xiang, Y., Fox, D.: DeepIM: Deep iterative matching for 6D pose estimation. International Journal of Computer Vision **128**(3), 657–678 (nov 2019). https://doi.org/10.1007/s11263-019-01250-9
14. Lipson, L., Teed, Z., Goyal, A., Deng, J.: Coupled iterative refinement for 6D multi-object pose estimation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2022). https://doi.org/10.1109/cvpr52688.2022.00661
15. Lowe, D.: Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision. vol. 2, pp. 1150–1157 vol.2. IEEE (Sep 1999). https://doi.org/10.1109/iccv.1999.790410
16. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: DeepSDF: Learning continuous signed distance functions for shape representation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2019). https://doi.org/10.1109/cvpr.2019.00025
17. Periyasamy, A.S., Schwarz, M., Behnke, S.: SynPick: A dataset for dynamic bin picking scene understanding. In: 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE). pp. 488–493. IEEE (aug 2021). https://doi.org/10.1109/case49439.2021.9551599
18. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3D deep learning with PyTorch3D. arXiv:2007.08501 (2020)
19. Strecke, M., Stueckler, J.: Where does it end? - reasoning about hidden surfaces by object intersection constraints. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2020). https://doi.org/10.1109/cvpr42600.2020.00961
20. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning, 2019 (May 2019)
21. Teed, Z., Deng, J.: RAFT: Recurrent all-pairs field transforms for optical flow. In: European Conference on Computer Vision (ECCV). pp. 402–419. Springer International Publishing (Mar 2020). https://doi.org/10.1007/978-3-030-58536-5_24
22. Tremblay, J., Wen, B., Blukis, V., Sundaralingam, B., Tyree, S., Birchfield, S.: Diff-DOPE: Differentiable deep object pose estimation (Sep 2023). https://doi.org/10.48550/ARXIV.2310.00463
23. Wada, K., Sucar, E., James, S., Lenton, D., Davison, A.J.: MoreFusion: Multi-object reasoning for 6D pose estimation from volumetric fusion. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2020). https://doi.org/10.1109/cvpr42600.2020.01455
24. Wang, G., Manhardt, F., Tombari, F., Ji, X.: GDR-Net: Geometry-guided direct regression network for monocular 6D object pose estimation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2021). https://doi.org/10.1109/cvpr46437.2021.01634
25. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: CVPR. pp. 2642–2651 (2019)

26. Wang, P.S., Liu, Y., Tong, X.: Dual octree graph networks for learning adaptive volumetric shape representations. ACM Transactions on Graphics (SIGGRAPH) **41**(4), 1–15 (Jul 2022). https://doi.org/10.1145/3528223.3530087

27. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. In: Robotics: Science and Systems XIV. Robotics: Science and Systems Foundation (jun 2018). https://doi.org/10.15607/rss.2018.xiv.019