Supplementary Material for PhysPose: Refining 6D Object Poses with Physical Constraints

Martin Malenický Martin Cífka Médéric Fourmy Louis Montaut Justin Carpentier
Josef Sivic Vladimir Petrik

A. Computation of analytic gradients

The proposed approach applies gradient descent to minimize the total cost defined in the main paper in Eq. (1). To achieve that, we derive analytical gradients for each partial cost defined in Sec 3.1.

The pose gradient for i-th object, denoted as $\boldsymbol{\nabla}\mathcal{P}_i$, guides the optimization process to maintain the object's pose close to the image-based estimate. It is computed as:

$$\boldsymbol{\nabla}\mathcal{P}_i = \boldsymbol{e}_i{}^T H_i J_i, \tag{1}$$

where $\boldsymbol{e}_i$ represents the residual vector between the optimized pose and the initial pose, defined as $\boldsymbol{e}_i = [\boldsymbol{t}_{C,Oi} - \tilde{\boldsymbol{t}}_{C,Oi}, \log\left(\tilde{R}^T_{C,Oi} R_{C,Oi}\right)]^T \in \mathbb{R}^6$, where $\boldsymbol{t}_{C,Oi}$ and $\tilde{\boldsymbol{t}}_{C,Oi}$ are the translation components of the optimized and estimated poses respectively, and $\log\left(\tilde{R}^T_{C,Oi} R_{C,Oi}\right)$ is the logarithmic mapping of the rotation difference. The term H_i is the precision matrix, which is the inverse of the covariance matrix, Σ_{Ci}. Finally, J_i is the Jacobian matrix, given by $J_i = \begin{bmatrix} R_{C,Oi} & \mathbf{0} \\ \mathbf{0} & \frac{\partial \log\left(R_{\tilde{O}i,Oi}\right)}{\partial R_{\tilde{O}i,Oi}} \end{bmatrix}$, where $R_{C,Oi}$ is the rotation matrix from the object O_i to the camera frame C, and the bottom right component is the partial derivative of the logarithmic rotation difference with respect to the rotation. An analytical formula for the jacobian of the SO(3) log map can be found in [3], Appendix B,C, Eq. (144), notated as $J_r(\theta)^{-1}$.

Collision gradient between two objects, denoted as $\mathcal{C}_{\mathcal{A},\mathcal{B}}$, aims to resolve overlapping shapes by moving them into a non-colliding state. To obtain the collision gradient $\boldsymbol{\nabla}\mathcal{C}_{\mathcal{A},\mathcal{B}}$, we need to differentiate the pairwise collision cost from Section 3.1 with respect to the pose of the object $\mathcal{A}$ (denoted as $T_\mathcal{A}$). The derivative is:

$$\boldsymbol{\nabla}\mathcal{C}_{\mathcal{A},\mathcal{B}} = \frac{\partial \mathcal{C}_{\mathcal{A},\mathcal{B}}}{\partial T_\mathcal{A}} = \frac{1}{n_\text{col}} \sum_{\mathcal{A}_i \in \mathcal{A}} \sum_{\mathcal{B}_j \in \mathcal{B}} \frac{\partial}{\partial T_\mathcal{A}} \left[-d(\mathcal{A}_i, \mathcal{B}_j)\right]_+ . \tag{2}$$

This can be further expressed as:

$$\boldsymbol{\nabla}\mathcal{C}_{\mathcal{A},\mathcal{B}} = \frac{1}{n_\text{col}} \sum_{\mathcal{A}_i \in \mathcal{A}} \sum_{\mathcal{B}_j \in \mathcal{B}} \begin{cases} 0 & \text{if } d(\mathcal{A}_i, \mathcal{B}_j) \geq 0 \\ -\frac{\partial d(\mathcal{A}_i, \mathcal{B}_j)}{\partial T_\mathcal{A}} & \text{if } d(\mathcal{A}_i, \mathcal{B}_j) < 0 \end{cases} \tag{3}$$

In our approach, the derivative of the signed distance, $\frac{\partial d(\mathcal{A}_i, \mathcal{B}_j)}{\partial T_\mathcal{A}}$, is obtained by the randomized smoothing approach described in [2].

Gravity gradient for object $\mathcal{A}$, denoted as $\mathcal{G}_\mathcal{A}$, prevents objects from levitating by encouraging them to move towards static objects in the direction of gravity. It is computed based on the closest convex subpart $\mathcal{B}$ of a static object and the average positive distance of the movable object $\mathcal{A}$'s convex subparts to $\mathcal{B}$ as defined in Section 3.1. The gravity gradient is then:

$$\boldsymbol{\nabla}\mathcal{G}_\mathcal{A} = \frac{\partial \mathcal{G}_\mathcal{A}}{\partial T_\mathcal{A}} = \frac{\partial}{\partial T_\mathcal{A}} \left(\delta_\mathcal{A} \frac{1}{|\mathcal{A}|} \sum_{\mathcal{A}_i \in \mathcal{A}} [d(\mathcal{A}_i, \mathcal{B})]_+ \right) . \tag{4}$$

Since $\delta_\mathcal{A}$ is a binary variable depending on the collision state and is assumed not to be directly dependent on the pose of object A (it depends on the state of other objects), and $|\mathcal{A}|$ is constant, we can rewrite the derivative as:

$$\boldsymbol{\nabla}\mathcal{G}_\mathcal{A} = \frac{\delta_\mathcal{A}}{|\mathcal{A}|} \sum_{\mathcal{A}_i \in \mathcal{A}} \frac{\partial}{\partial T_\mathcal{A}} [d(\mathcal{A}_i, \mathcal{B})]_+ , \tag{5}$$

where the partial derivative of the hinge loss is computed in the same way as for the collision gradient described above.

B. Additional qualitative results

In Fig. 1 and Fig. 2 we present additional qualitative results for enforcing physical pose estimation on the YCB-Video dataset. Please note, that the HOPE-Video dataset features levitating objects in its initial poses, a characteristic not readily apparent in static visualizations. Therefore, the static visualization of qualitative results for the HOPE-Video dataset has been omitted. Please refer to the supplementary video, described below, for a dynamic visualization of the optimization process on a HOPE-Video scene.

C. Supplementary video

The first part of the supplementary video demonstrates the optimization process for scenes from the YCB-Video and HOPE-Video datasets. YCB-Video scenes, which initially exhibit object-object and object-scene collisions, are successfully resolved by our PhysPose method. For the HOPE-Video datasets, the initial poses exhibit levitation of objects above the tabletop; our method successfully mitigates this issue by attracting the objects downwards towards the scene geometry.

The second section of the video presents our robotic grasping experiments, contrasting the performance of the baseline method [1] with our approach. In the first experiment, the baseline's insufficient pose accuracy prevents a firm grasp of objects, as exemplified by the mustard bottle. Conversely, our method, leveraging its refined pose estimates, achieves successful grasps. Subsequent experiments demonstrate the baseline attempting grasps of objects predicted to collide with the scene geometry, potentially damaging the detected Cheez-It box. Note the Cheez-It box is quickly removed by the robot operator just before it would be damaged by the robot. Our method, however, effectively avoids these collisions and grasps the Cheez-It box without incident. Finally, the baseline occasionally predicts objects as levitating above the surface, leading to grasping attempts that miss the object entirely. This issue is rectified by our more accurate pose estimates, as evidenced by our successful grasp of the sugar box.

References

[1] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. MegaPose: 6D Pose Estimation of Novel Objects via Render & Compare. In *CoRL*, 2022. 2, 3, 4

[2] Louis Montaut, Quentin Le Lidec, Antoine Bambade, Vladimir Petrik, Josef Sivic, and Justin Carpentier. Differentiable collision detection: a randomized smoothing approach. In *ICRA*, 2023. 1

[3] Joan Solà, Jeremie Deray, and Dinesh Atchuthan. A micro lie theory for state estimation in robotics. *arXiv:1812.01537*, 2021. 1

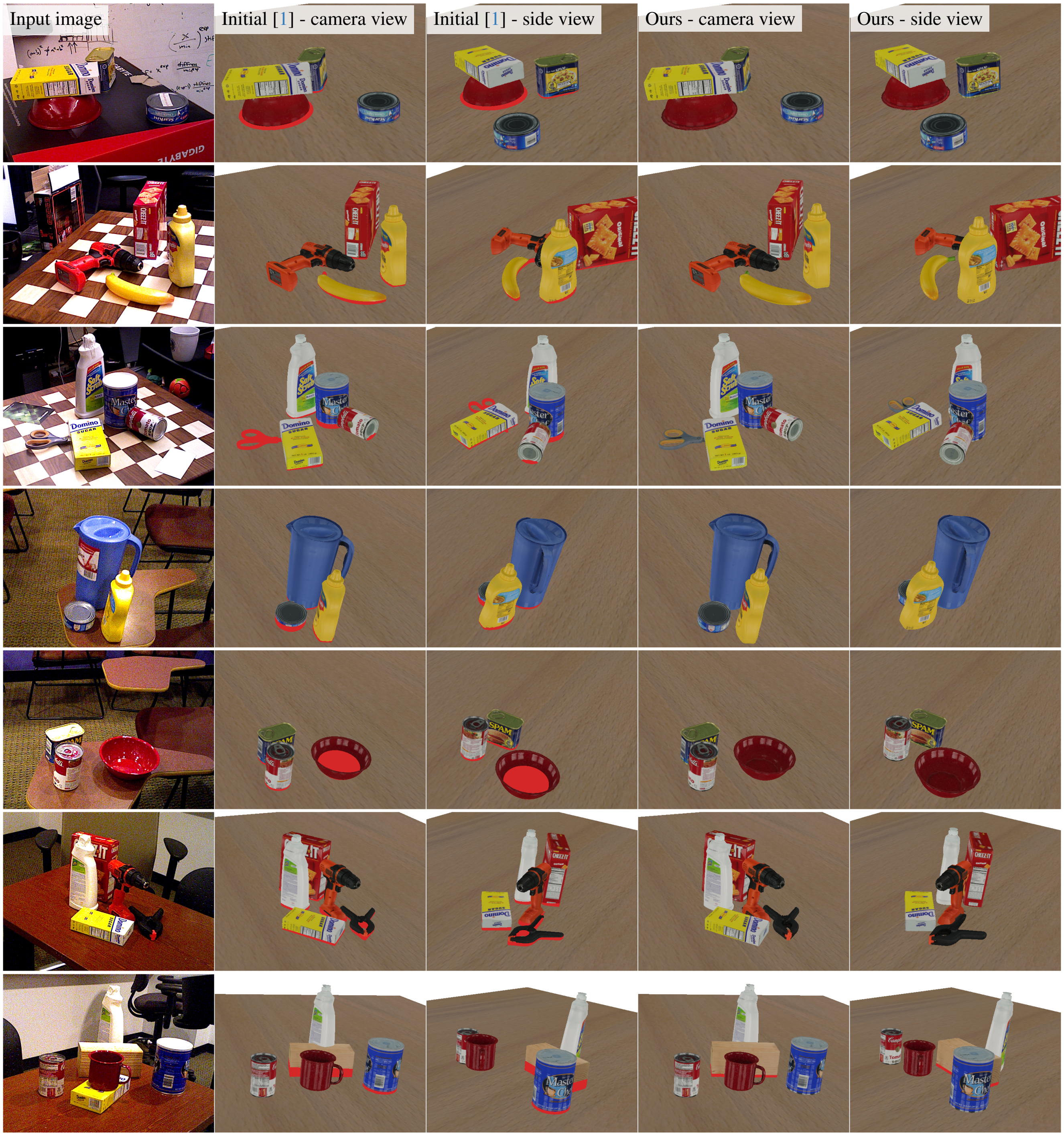

Figure 1. **Qualitative results on the YCB-V dataset.** We estimate initial object poses from an input image using MegaPose [1]. The resulting scene, shown from camera and side views, exhibits significant physical inconsistencies, with colliding parts highlighted in red. Our physical consistency optimization method significantly reduces these collisions, leading to a more plausible scene arrangement. Notice how our method successfully resolves collisions in the first two rows even though the objects are placed on top of each other.

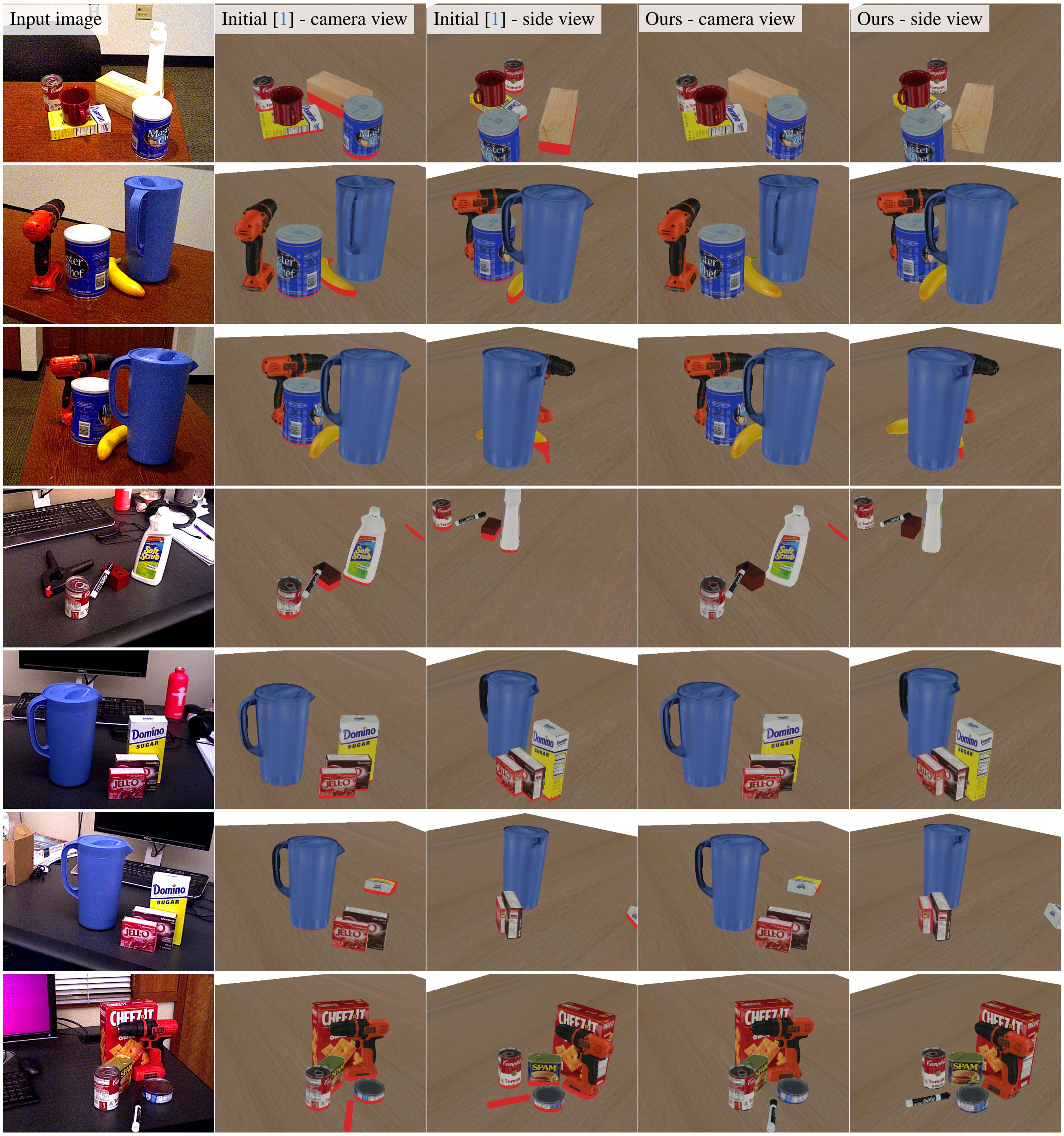

Figure 2. **Qualitative results on the YCB-V dataset.** We estimate initial object poses from an input image using MegaPose [1]. The resulting scene, shown from camera and side views, exhibits significant physical inconsistencies, with colliding parts highlighted in red. Our physical consistency optimization method significantly reduces these collisions, leading to a more plausible scene arrangement. Notice how our method successfully resolves collisions in the first two rows even though the objects are placed on top of each other.