Supplementary Material for SO-Pose: Exploiting Self-Occlusion for Direct 6D Pose Estimation

Yan Di[1], Fabian Manhardt[2], Gu Wang[3], Xiangyang Ji[3], Nassir Navab[1] and Federico Tombari[1,2]

[1]Technical University of Munich, [2]Google, [3]Tsinghua University

*shangbuhuan13@gmail.com, {fabian.manhardt, nassir.navab}@tum.de, wangg16@mails.tsinghua.edu.cn xyji@tsinghua.edu.cn, tombari@in.tum.de

Abstract

This document supplements our main paper entitled SO-Pose: Exploiting Self-Occlusion for Direct 6D Pose Estimation *by providing more details on involved equations, motivations, visualizations on the proposed self-occlusion information and showing additional qualitative results as well as a brief qualitative comparison with GDR-Net [2].*

1. Additional details with respect to Eq. 6

From Eq. 3 of the main paper, we get

$$P = Z_P K^{-1} \rho. \tag{1}$$

As P and Q_x both lie on the same line, we similarly get

$$Q_x = Z_{Q_x} K^{-1} \rho. \tag{2}$$

In addition, from Eq. 4. we we know that

$$(Rn_x)^T Q_x = (Rn_x)^T t. \tag{3}$$

When now substituting Eq. 2 into Eq. 3, we obtain

$$(Rn_x)^T (Z_{Q_x} K^{-1} \rho) = (Rn_x)^T t, \tag{4}$$

with

$$Z_{Q_x} = \frac{(Rn_x)^T t}{(Rn_x)^T (K^{-1} \rho)}, \tag{5}$$

Hence, we can finally derive

$$Q_x = Z_{Q_x} (K^{-1} \rho) = \frac{(Rn_x)^T t}{(Rn_x)^T (K^{-1} \rho)} (K^{-1} \rho). \tag{6}$$

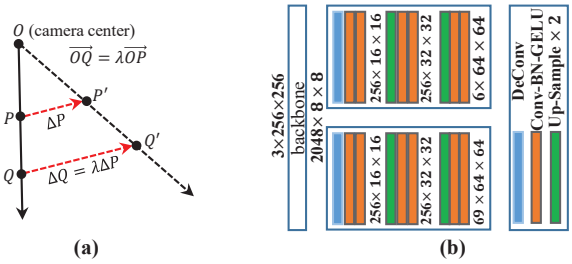


Figure 1. (a) illustrates how self-occlusion can help to reduce point matching noise. (b) Details on the architecture of the two branches.

2. Motivation on Leveraging Self-Occlusion

As shown in Fig. 1, P occludes Q, thus $\overrightarrow{OQ} = \lambda \overrightarrow{OP}$. If P moves to P' by ΔP due to matching noise, then to keep $\overrightarrow{OQ'} = \lambda \overrightarrow{OP'}$, Q must move $\Delta Q = \lambda \Delta P$. Since $\lambda > 1$ under occlusion, we have $\|\Delta Q\| > \|\Delta P\|$. Therefore, the matching error increased due to self-occlusion. If ΔP is small, then ΔQ may be significant enough for the network to identify the error. In the paper, reorganizing Eq. 6, λ can be directly calculated. Hence, via applying our training loss w.r.t. $\mathcal{L} = \|\Delta P\|_1 + \|\Delta Q\|_1 + \mathcal{L}_{cl-2D,3D}$, we can guide the network to learn more representative features, so to reduce matching noise. In the paper, we demonstrate that SO-Pose can surpass all other methods enforcing only the standard 2D-3D matching loss, *i.e.* $\mathcal{L} = \|\Delta P\|_1$.

In practice, we find that directly leveraging self-occlusion can lead to problems. First of all, renderers are required to compute the self-occlusion information, which can become time-consuming. In addition, the real self-occlusion information for *thin objects* (e.g. bananas, scissors), is hard to learn, leading to unstable results. To solve these issues, we instead calculate the occlusion information between the visible object surface and its coordinate frames. We still refer to it as self-occlusion, as it is determined by the object pose and shape. In addition, to further increase

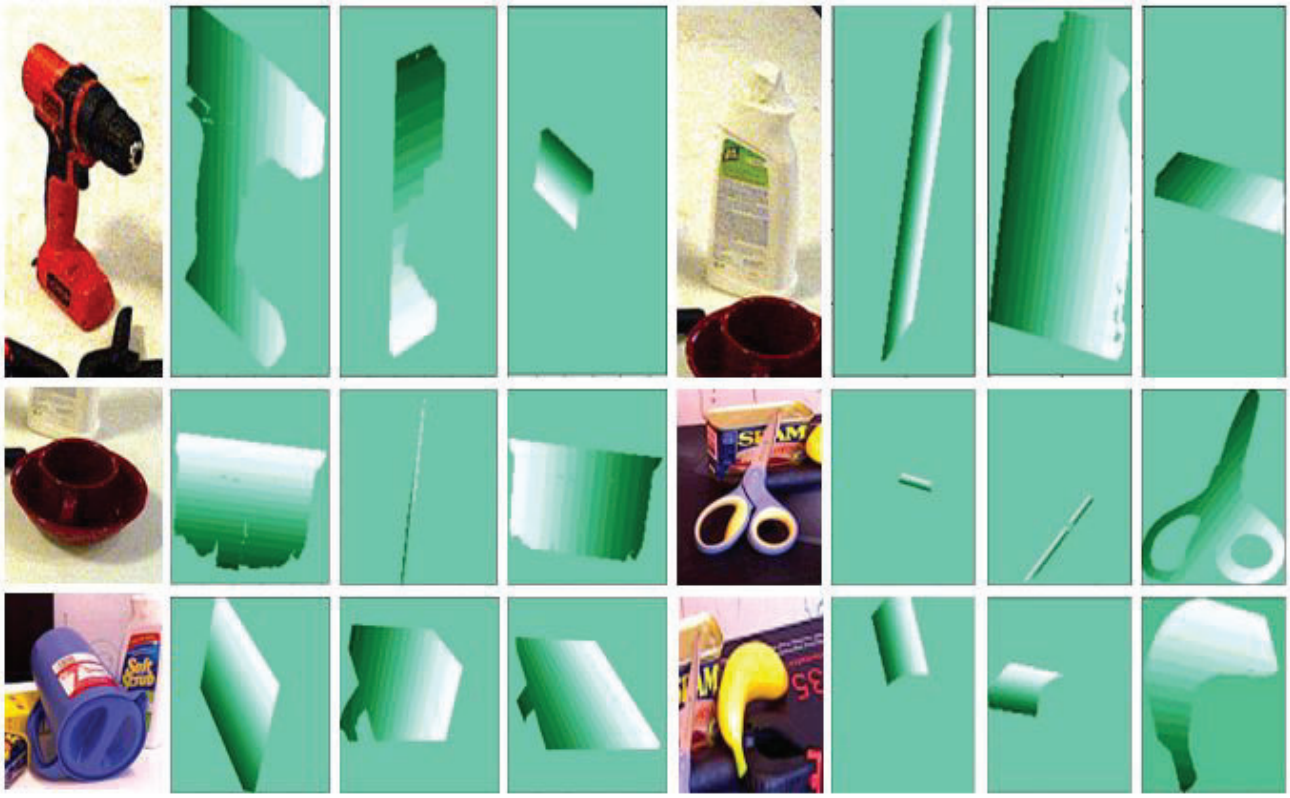


Figure 2. Self-occlusion coordinates of target objects.

training stability, we only consider self-occlusion information within a predefined region as shown in Fig. 3.

3. Details of Architecture

We demonstrate the basic structure of the two branches for self-occlusion and 2D-3D matching in Fig. 1 (b). For more details, please refer to the code on GitHub.

4. Self-Occlusion Coordinates

In Fig. 2, we provide additional results demonstrating the self-occlusion coordinates of the target objects. For an object, we present its self-occlusion information Q_x, Q_y and Q_z in the respective order.

5. More Results on LMO

In Fig. 3 – 6, we provide additional qualitative results for the 6D pose on LMO [1]. SO-Pose is able to estimate accurate and reliable 6D poses for all different kinds of objects even under strong occlusion.

6. More results on YCB-V

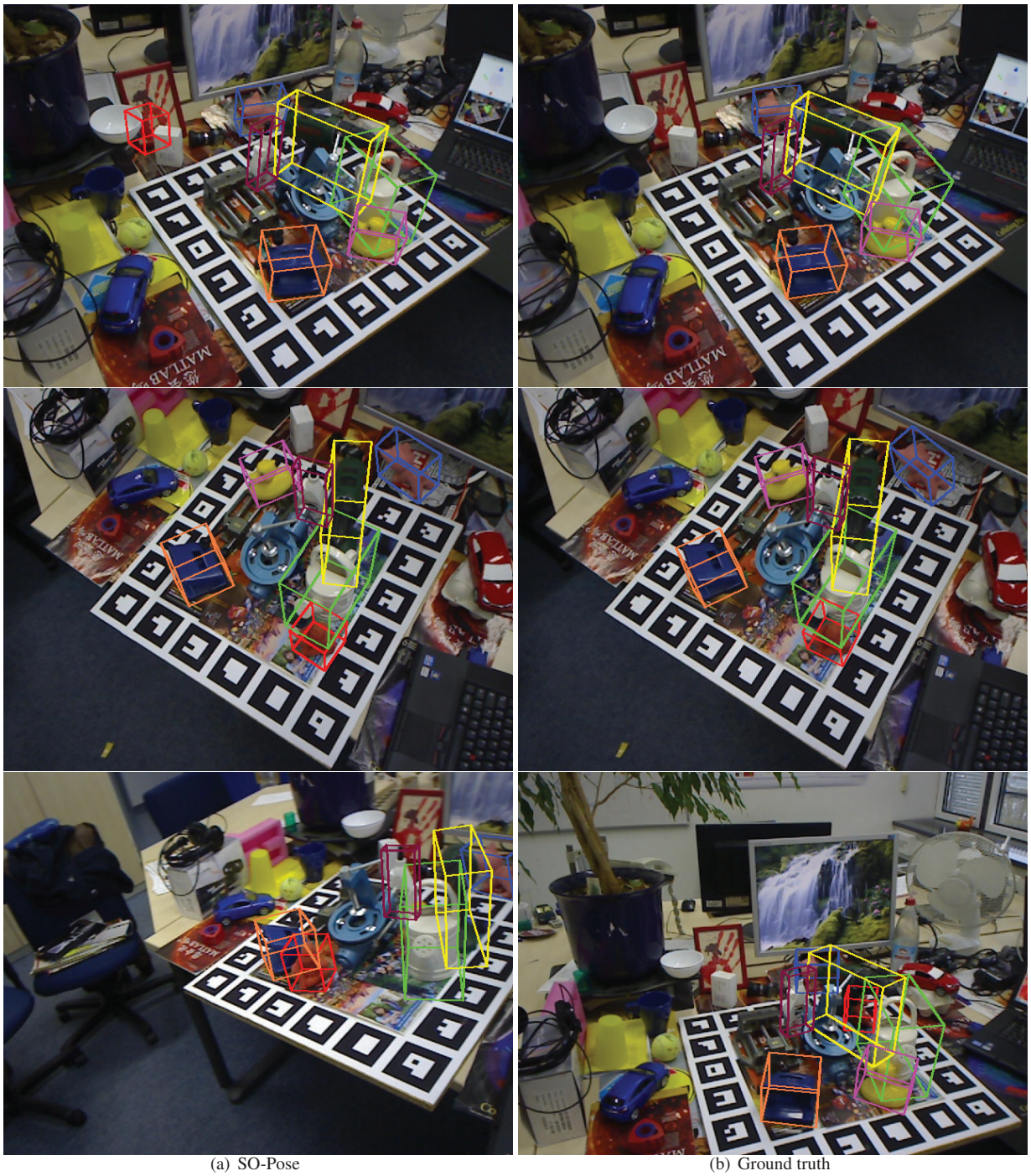
In Fig. 7 – 9, we illustrated qualitative results of SO-Pose and GDR-Net [2] for YCB-V [3]. While both methods perform very well even under occlusion, GDR-Net occasionally produces a few bad 6D poses whereas our method always computes proper estimates. An exemplary situation can be observed in the last row of Fig. 7. In contrast to GDR-Net, SO-Pose was able to estimate a good 6D pose candidate despite the bowl undergoing sever occlusion by the sugar box.

References

[1] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6D object pose estimation using 3D object coordinates. In *ECCV*, pages 536–551, 2014. 2, 3, 4, 5, 6

[2] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *CVPR*, June 2021. 1, 2, 7, 8, 9

[3] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. *RSS*, 2018. 2, 7, 8, 9

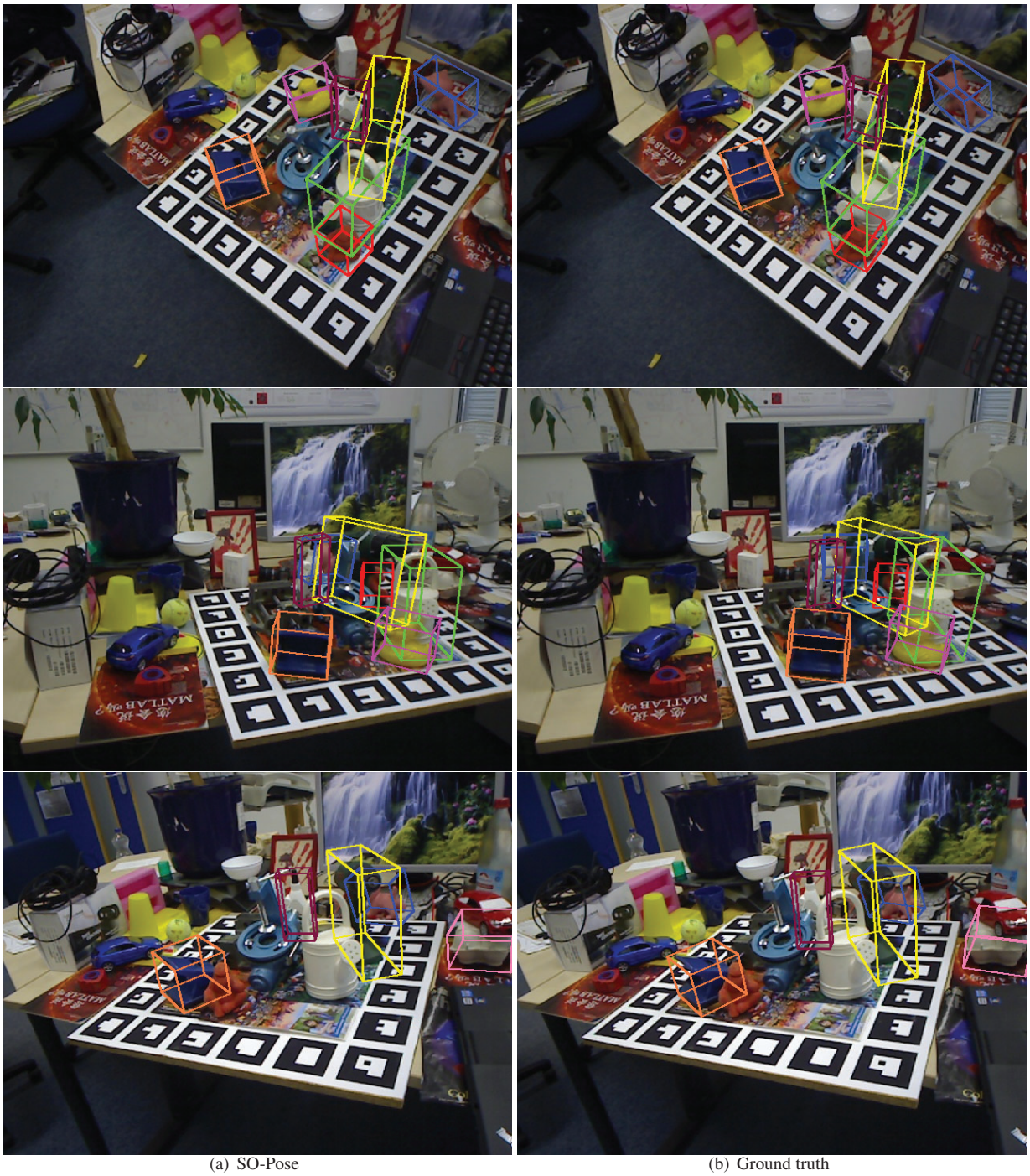


(a) SO-Pose (b) Ground truth

Figure 3. Qualitative pose estimation results on LMO [1].

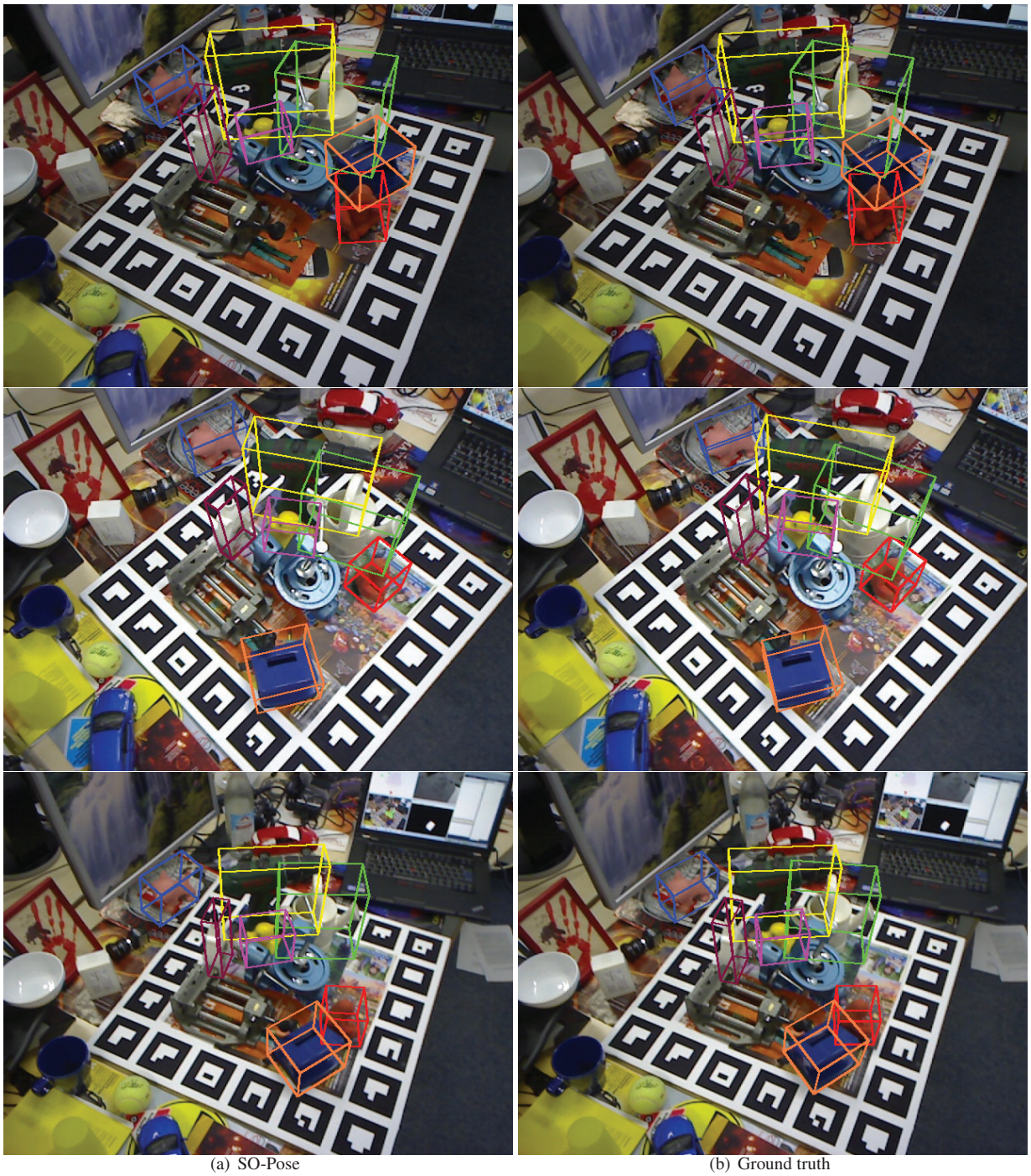
(a) SO-Pose (b) Ground truth

Figure 4. Qualitative pose estimation results on LMO [1].



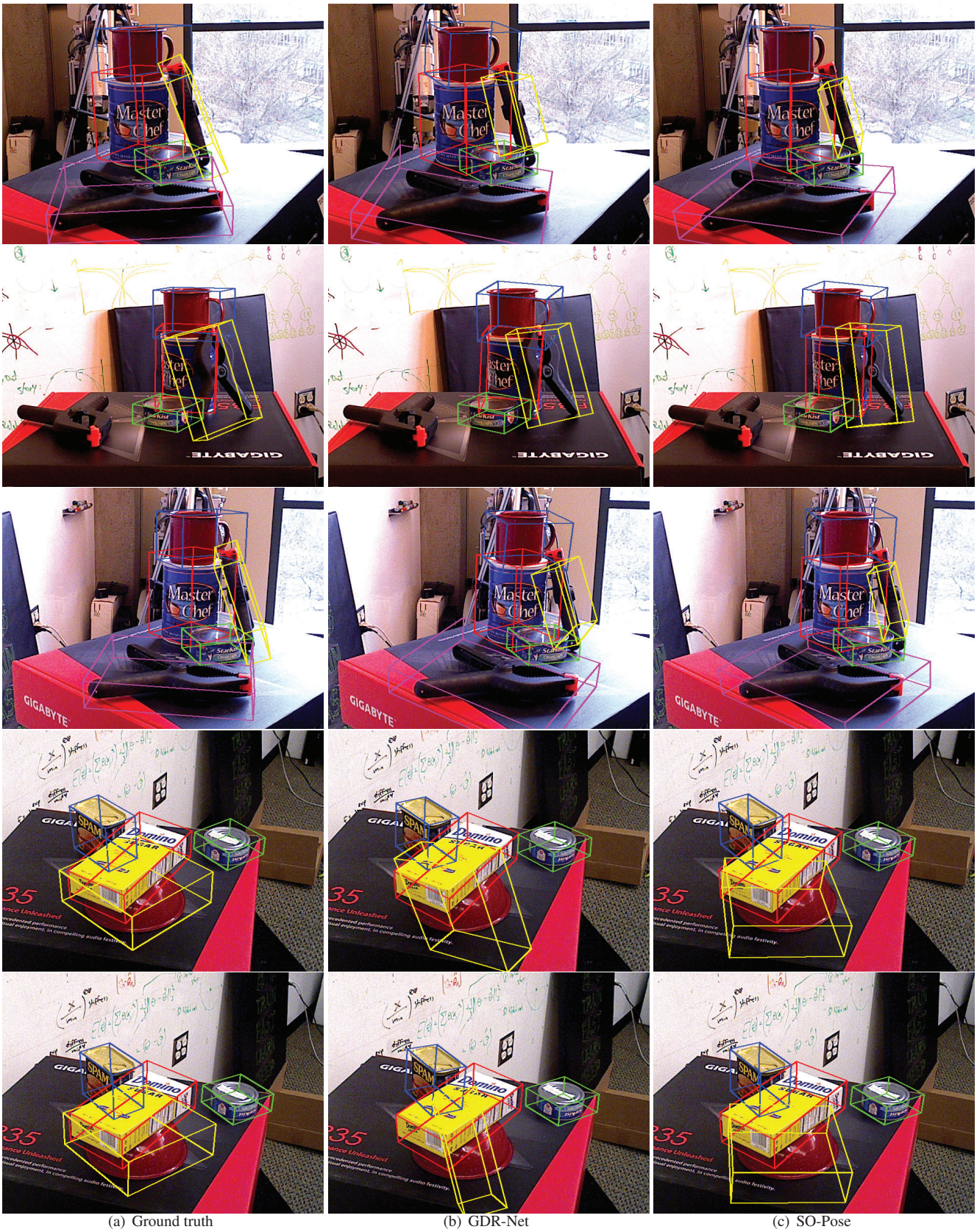
(a) SO-Pose (b) Ground truth

Figure 5. Qualitative pose estimation results on LMO [1].



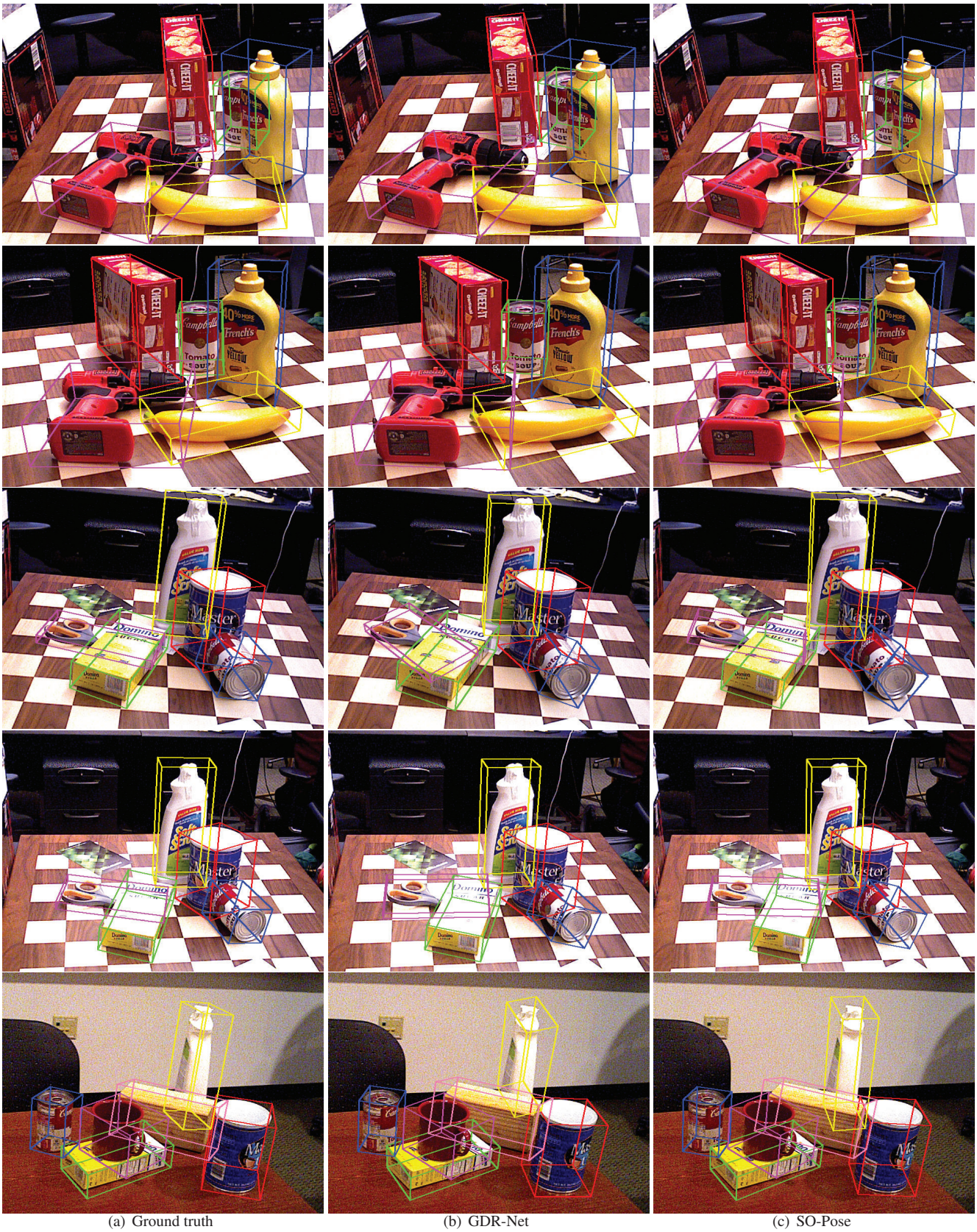
(a) SO-Pose (b) Ground truth

Figure 6. Qualitative pose estimation results on LMO [1].



(a) Ground truth (b) GDR-Net (c) SO-Pose

Figure 7. Qualitative pose estimation results on YCB-V [3]. We compare our method with GDR-Net [2].



(a) Ground truth (b) GDR-Net (c) SO-Pose

Figure 8. Qualitative pose estimation results on YCB-V [3]. We compare our method with GDR-Net [2].

(a) Ground truth (b) GDR-Net (c) SO-Pose

Figure 9. Qualitative pose estimation results on YCB-V [3]. We compare our method with GDR-Net [2].