# KMOPS: Keypoint-Driven Method for Multi-Object Pose and Metric Size Estimation from Stereo Images Supplementary Material

| Method | $3D_{25}$ | $3D_{50}$ | $3D_{75}$ | $5^\circ$2cm | $10^\circ$2cm |
|---|---|---|---|---|---|
| PnP | 93.1 | 82.9 | 50.3 | 14.8 | 29.3 |
| SVD | **95.6** | **90.5** | **70.5** | **18.0** | **39.4** |

Table A1. Ablations for different pose alignment methods on StereOBJ-1M validation set. "SVD" denotes the pose registration method used in the paper.

## 1. Additional Ablations

Here we provide additional ablations that further demonstrates the advantage of our design choice.

### 1.1. Solving Pose using PnP

In this ablation study, we compare our approach with a variant that uses the PnP algorithm [3] to solve for object pose. Similar to our original pipeline, we first construct canonical keypoints from triangulated predictions. In order to utilize PnP, we use the 2D keypoints estimated from the left image together with the 3D canonical keypoints to recover the pose. As shown in Table A1, there is an evident performance gap between PnP and our SVD-based method. While PnP relies on 2D–3D correspondences, our proposed method directly registers the triangulated 3D keypoints. This highlights the importance of exploiting geometric information from stereo images for accurate pose estimation.

## 2. Full-Category Evaluation on TOD

In this section, we extend our evaluation to all object categories in the TOD dataset including *ball*, *heart*, and *tree*, in addition to the previously considered categories *bottle*, *cup*, and *mug*. Following KVN [2], we report the following metrics:

- $<$ 2cm : the fraction of test samples whose mean keypoint error is below 2 cm. For symmetric objects, each predicted keypoint is matched to its nearest ground-truth counterpart to resolve ambiguity.
- **MAE** : the mean absolute error (in mm), i.e. the average Euclidean distance between each predicted and true 3D keypoint.
- **AUC** : the area under the MAE curve for thresholds from 0 cm to 10 cm, providing a single value that summarizes performance over varying tolerance levels.

### 2.1. Comparison with Stereo-Based Methods

We train on all categories and present the results in Table A2, comparing against several strong stereo approaches: KeyPose [4], KVN [2], GhostPose [1], and s-PVNet. KeyPose and KVN require a pre-cropped object patch prior to pose estimation, s-PVNet adapts the monocular PVNet [5] to stereo input and is similar to our approach, GhostPose directly processes the full image. Our approach achieves strong performance across most object classes. We achieves competitive results on the *mug* category, improving AUC by $\approx 3.1\%$, $< 2$ cm accuracy by $\approx 18\%$, and reducing MAE by $\approx 16\%$ overall. Similar gains are observed for *bottle* and *cup*, demonstrating the effectiveness of our keypoint-driven regression.

### 2.2. Failure Case: *Heart* Class

Although our approach surpasses previous methods on almost all categories, the *heart* class remains challenging. To assess whether the issue stems from the object's small size or visual ambiguity, we re-evaluate the model on tightly cropped object regions. This simple adjustment markedly improves scores across nearly every split (Table A3) and narrows the gap for *heart*. The predicted keypoints are often confused between the tip and back of the *heart*, leading to a performance drop. This confusion may stem from label noise or annotation inconsistencies in this class (see Fig. A1) and we will leave this in future work for further exploration.

## References

[1] Jaesik Chang, Minju Kim, Seongmin Kang, Heungwoo Han, Sunpyo Hong, Kyunghun Jang, and Sungchul Kang. Ghostpose*: Multi-view pose estimation of transparent objects for robot hand grasping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, pages 5749–5755. IEEE, 2021. 1

| Method | Metric | $ball_0$ | $bottle_0$ | $bottle_1$ | $bottle_2$ | $cup_0$ | $cup_1$ | $mug_0$ | $mug_1$ | $mug_2$ | $mug_3$ | $mug_4$ | $mug_5$ | $mug_6$ | $heart_0$ | $tree_0$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| KeyPose | AUC | 96.1 | **95.4** | **94.9** | 90.7 | 93.1 | 92.0 | 91.0 | 78.1 | 89.7 | 88.6 | 87.8 | 91.0 | 90.3 | 84.3 | 87.1 |
| | $< 2$ cm | **100.0** | **99.8** | 99.7 | 91.4 | 97.8 | 95.3 | 94.6 | 63.6 | 90.1 | 94.8 | 93.4 | 93.1 | 92.2 | 77.2 | 82.5 |
| | MAE | 3.8 | **4.6** | **5.1** | 9.3 | 6.8 | 7.1 | 8.9 | 21.9 | 10.1 | 10.8 | 10.2 | 9.0 | 9.7 | 15.6 | 12.8 |
| GhostPose | AUC | – | 94.5 | 92.4 | 90.2 | 92.7 | 91.8 | 90.1 | 90.0 | 90.1 | 90.1 | 89.7 | 90.3 | 89.0 | 88.2 | – |
| | $< 2$ cm | – | 96.4 | 95.2 | 91.6 | 92.6 | 94.0 | 93.4 | **94.8** | 93.4 | 93.4 | 92.9 | 95.8 | 93.6 | **93.1** | – |
| | MAE | – | 5.3 | 7.8 | 9.4 | 7.2 | 8.4 | 10.1 | 10.8 | 10.2 | 10.1 | 10.1 | 9.0 | 10.6 | 11.5 | – |
| s-PVNet | AUC | 96.6 | 94.7 | 93.7 | 89.9 | 93.1 | 94.3 | 85.8 | 79.7 | 88.4 | 87.3 | 89.4 | 86.0 | 87.2 | 89.5 | 90.5 |
| | $< 2$ cm | 97.7 | 95.2 | 96.6 | 91.1 | 96.8 | 97.1 | 84.4 | 71.9 | 88.7 | 88.5 | 91.8 | 90.3 | 89.5 | 85.5 | 92.8 |
| | MAE | 3.9 | 5.8 | 6.7 | 10.7 | 7.5 | 6.1 | 15.8 | 21.7 | 12.8 | 14.7 | 11.6 | 16.3 | 14.1 | 10.9 | 10.1 |
| KVN | AUC | **97.4** | 95.0 | 94.0 | **92.5** | 94.2 | 94.4 | 91.1 | 82.6 | **91.4** | **91.5** | 91.9 | 89.8 | 88.3 | **90.4** | 91.1 |
| | $< 2$ cm | 99.5 | 95.4 | 96.8 | 94.8 | 97.6 | 97.3 | 93.6 | 77.4 | **94.2** | **93.3** | 96.2 | 93.2 | 90.7 | 87.2 | 93.4 |
| | MAE | **3.1** | 5.4 | 6.4 | **8.1** | 6.3 | 6.0 | 9.5 | 18.3 | **9.5** | **10.1** | 8.9 | 12.0 | 13.2 | **10.0** | 9.4 |
| Ours | AUC | 94.3 | 93.1 | 94.6 | 91.9 | **96.8** | **96.3** | **93.4** | **90.1** | 88.9 | 90.0 | **92.8** | **95.6** | **95.4** | 39.5 | **92.8** |
| | $< 2$ cm | 95.0 | 96.3 | **100.0** | **96.5** | **100.0** | **99.7** | **97.5** | 91.9 | 86.3 | 91.6 | **98.4** | **100.0** | **100.0** | 5.3 | **97.9** |
| | MAE | 6.2 | 7.3 | 5.8 | 8.5 | **3.7** | **4.2** | **7.1** | **10.2** | 12.8 | 12.9 | **7.6** | **4.8** | **5.1** | 102.9 | **7.7** |

Table A2. Per-object pose estimation metrics (%). For AUC and "$< 2$ cm" higher is better; for MAE (pixels) lower is better.

| Method | Metric | $ball_0$ | $bottle_0$ | $bottle_1$ | $bottle_2$ | $cup_0$ | $cup_1$ | $mug_0$ | $mug_1$ | $mug_2$ | $mug_3$ | $mug_4$ | $mug_5$ | $mug_6$ | $heart_0$ | $tree_0$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ours | AUC | 94.3 | 93.1 | 94.6 | 91.9 | **96.8** | 96.3 | 93.4 | 90.1 | 88.9 | 90.0 | 92.8 | 95.6 | 95.4 | 39.5 | **92.8** |
| | $< 2$ cm | 95.0 | 96.3 | **100.0** | 96.5 | **100.0** | **99.7** | 97.5 | 91.9 | 86.3 | 91.6 | 98.4 | **100.0** | **100.0** | 5.3 | **97.9** |
| | MAE | 6.2 | 7.3 | 5.8 | 8.5 | **3.7** | 4.2 | 7.1 | 10.2 | 12.8 | 12.9 | 7.6 | 4.8 | 5.1 | 102.9 | **7.7** |
| Ours (Cropping) | AUC | **98.0** | **97.8** | **98.1** | **96.4** | **96.6** | **96.7** | **97.0** | **93.4** | **95.6** | **95.9** | **96.0** | **96.8** | **96.8** | **63.1** | 92.6 |
| | $< 2$ cm | **100.0** | **100.0** | **100.0** | **99.7** | 99.1 | 99.1 | **100.0** | **97.2** | **98.4** | **99.4** | **100.0** | 98.1 | **100.0** | **44.7** | 93.8 |
| | MAE | **2.4** | **2.6** | **2.3** | **4.1** | 3.9 | **3.8** | **3.5** | **6.0** | **4.8** | **4.6** | **4.4** | **3.6** | **3.7** | **45.0** | 7.9 |

Table A3. Comparison of cropping and non-cropping experiment (%). For AUC and "$< 2$ cm" higher is better; for MAE (pixels) lower is better.


[2] Ivano Donadi and Alberto Pretto. KVN: keypoints voting network with differentiable RANSAC for stereo pose estimation. *IEEE Robotics Autom. Lett.*, 9(4):3498–3505, 2024. 1

[3] Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A. Vela, and Stan Birchfield. Single-stage keypoint- based category-level object pose estimation from an RGB image. In *ICRA*, 2022. 1

[4] Xingyu Liu, Shun Iwase, and Kris M. Kitani. Stereobj-1m: Large-scale stereo image dataset for 6d object pose estimation. In *ICCV*, 2021. 1

[5] Georgios Pavlakos, Xiaowei Zhou, Aaron Chan, Konstantinos G. Derpanis, and Kostas Daniilidis. 6-dof object pose from semantic keypoints. In *ICRA*, pages 2011–2018, 2017. 1

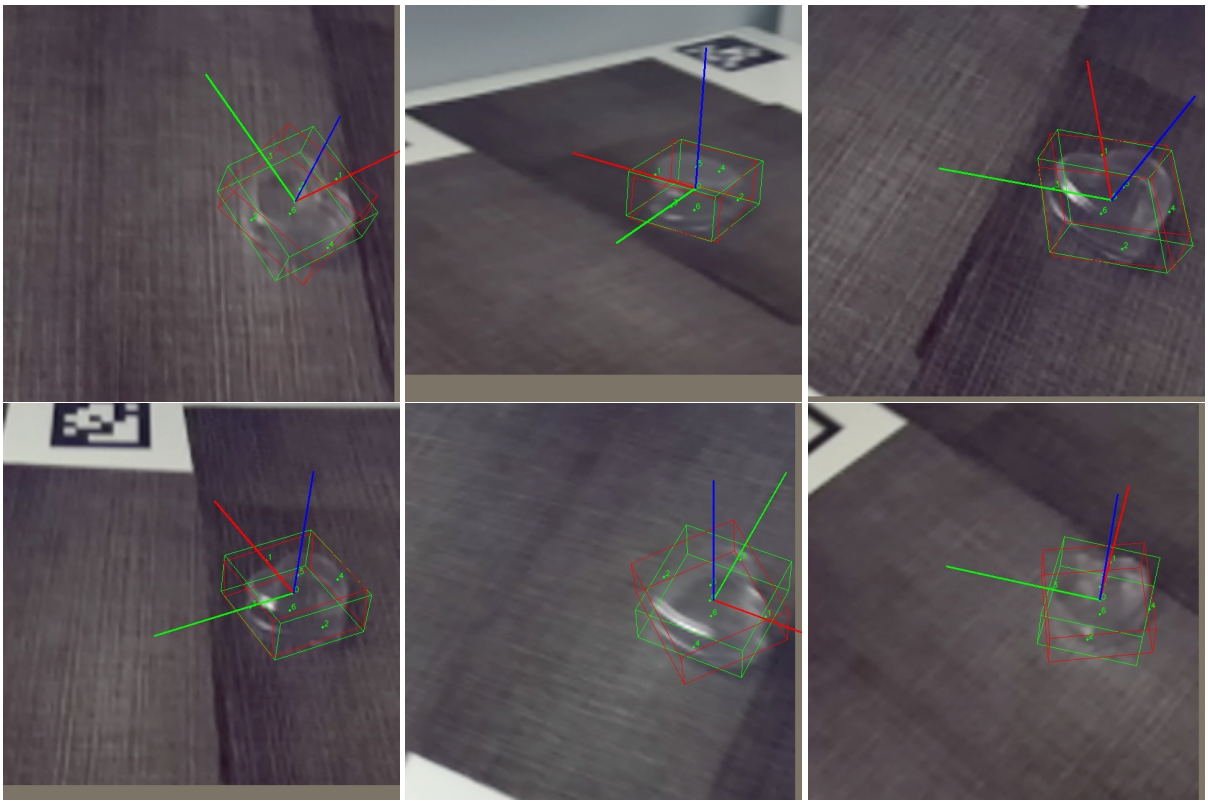

Figure A1. Vistualization of result for prediction(Green) and ground truth(Red) label for heart object. The labels for the *heart* object vary across different poses and images, indicating potential inconsistencies in the annotation.