GeoReF: Geometric Alignment Across Shape Variation for Category-level Object Pose Refinement

Linfang Zheng[1,3] Tze Ho Elden Tse* [3] Chen Wang* [1,2] Yinghan Sun[1] Hua Chen[1]
Aleš Leonardis[3] Wei Zhang† [1] Hyung Jin Chang[3]

[1]Shenzhen Key Laboratory of Control Theory and Intelligent Systems, School of System Design and Intelligent Manufacturing, Southern University of Science and Technology, China
[2]Department of Computer Science, the University of Hong Kong, China
[3]School of Computer Science, University of Birmingham, UK
{lxz948, txt994}@student.bham.ac.uk, cwang5@cs.hku.hk, sunyh2021@mail.sustech.edu.cn
{chenh6,zhangw3}@sustech.edu.cn,{a.leonadis,h.j.chang}@bham.ac.uk

Abstract

Object pose refinement is essential for robust object pose estimation. Previous work has made significant progress towards instance-level object pose refinement. Yet, category-level pose refinement is a more challenging problem due to large shape variations within a category and the discrepancies between the target object and the shape prior. To address these challenges, we introduce a novel architecture for category-level object pose refinement. Our approach integrates an HS-layer and learnable affine transformations, which aims to enhance the extraction and alignment of geometric information. Additionally, we introduce a cross-cloud transformation mechanism that efficiently merges diverse data sources. Finally, we push the limits of our model by incorporating the shape prior information for translation and size error prediction. We conducted extensive experiments to demonstrate the effectiveness of the proposed framework. Through extensive quantitative experiments, we demonstrate significant improvement over the baseline method by a large margin across all metrics. [1]

1. Introduction

Understanding an object's pose is crucial for a wide range of real-world applications, including robotic manipulation [18, 32, 52, 56], augmented reality [30, 33], and autonomous driving [19, 39]. Significant progress has been made for object pose estimation [5, 11, 15, 48, 53, 55] and pose refinement [2, 17, 22, 37, 51, 60] using the object's CAD model. Despite the promising performance, the reliance on accurate instance-level CAD models limits their generalizability to everyday objects. Category-level methods [7, 23–25, 29, 59, 61] is therefore been proposed to overcome this limitation. The objective of this line of work focuses on estimating object poses within a category given category-level shape priors. As a result, they face unique challenges as there exist diverse shape variations in each object category. We illustrate these shape variations in Fig. 1.

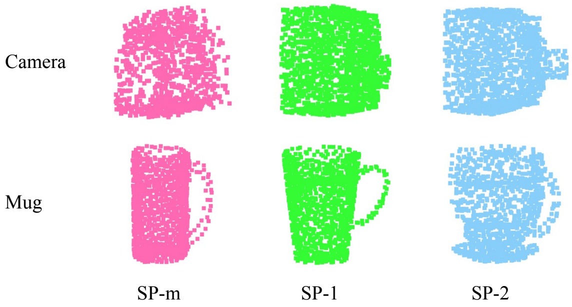


Figure 1. **Illustration of the shape variation.** *SP-m* represents the category's mean shape, *SP-1* and *SP-2* represents the randomly sampled object shapes from the CAMERA25 training set.

Recently, there have been remarkable advancements in category-level object pose estimation [9, 10, 25, 58, 61], primarily due to effective utilization of geometric information through 3D graph convolution [26]. In applications that require high precision, it is common to employ an object pose refinement procedure in conjunction with pose estimation. This involves an initial pose estimation algorithm determining the object pose, followed by a refinement step to further enhance the accuracy of the initially estimated pose by predicting and correcting its error. However, while instance-level object pose refinement has been extensively studied,

*Equal contribution, order by dice rolling.
†The corresponding author.
[1] Project page: https://lynne-zheng-linfang.github.io/georef.github.io

category-level pose refinement remained unexplored until the introduction of CATRE [29]. By leveraging initial object pose and size estimations, CATRE achieves category-level pose refinement by iteratively aligning the observed target object point cloud with the category-level shape prior. This pipeline is shown to be effective by improving the accuracy of the initial pose and size estimations.

While CATRE has proven to be effective in many scenarios, it is limited by the reliance on the PointNet [35] encoder, which is primarily designed for classification and segmentation tasks. This design choice limits its ability to capture essential and fine-grained geometric relationships for accurate pose estimation and refinement. This ability is particularly important in category-level pose refinement as there exists diverse shape variations between inputs. Consequently, CATRE obtains suboptimal performance by direct application of 3D graph convolution. In addition, as CATRE treats the point cloud and the shape prior features separately until a later stage of the network, they potentially miss out on the benefits of integrating these features earlier. Moreover, their approach does not incorporate shape prior information into the translation and scale estimation module, which presents another area for potential improvement.

In this paper, we introduce a novel architecture for category-level object pose refinement which aims to address the limitations mentioned above. To better extract both local and global geometric information, we incorporate an HS layer into our feature extraction process. We apply learnable affine transformations to the features to address the geometric discrepancies between the observed point cloud and the shape prior. This enables the network to align these features more effectively. In addition, we propose a cross-cloud transformation mechanism that is specifically designed to enhance the merging of information between the observed point clouds and the shape prior. This mechanism enables more efficient integration of information between the two sources. Finally, we push the limit of our model by incorporating shape prior information to more accurately predict errors in translation and size estimation.

Our extensive experimental results on two category-level object pose datasets demonstrate that our proposed model to be effective in addressing the problem of shape variations in category-level object pose refinement, and consequently outperforms the state-of-the-art significantly. To the best of our knowledge, our proposed method is the first to successfully address the shape variation issue which is common in category-level pose refinement. Specifically, to enable graph convolution to be effective in capturing geometric relationships between different shapes, we propose an adaptive affine transformation matrix that aligns the observed point clouds and the shape prior. Additionally, the proposed cross-cloud transformation mechanism effectively fuses features from different input point clouds and brings further performance improvements.

Our contributions are as follows:

- We introduce a novel architecture to specifically address the shape variations issue in category-level object pose refinement. Our proposed method results in consistent performance gain and exhibits better generalization ability.
- We propose a unique cross-cloud transformation mechanism which efficiently merges diverse information from observed point clouds and shape priors.
- We conduct extensive experiments on two category-level object pose datasets to validate our proposed method. On the REAL275 dataset, our method significantly outperforms SPD by 39.1% increase in the $5°5$cm metric. Additionally, we achieve 10.5% improvement in the $10°2$cm metric over the state-of-the-art method, CATRE.

2. Related Work

Instance-level object pose estimation and refinement. Instance-level approaches estimate the pose of the target object given known 3D CAD models. They can be briefly divided into correspondence matching methods and template matching methods. Correspondence matching methods [4, 8, 12, 13, 31, 36, 40, 41, 44, 46, 55, 62] matches the outstanding features of the observed object images with its model. Template matching methods [1, 14, 27, 34, 38, 42, 43, 47] compares the images or extracted features with the pre-generated templates. As the initial pose estimates can be noisy to various factors such as occlusions, object pose refinement [20, 22, 51] is shown to be useful in improving the performance of instance-level methods. Even though they achieved impressive over the target object, the reliance on object CAD models limited their generalizability for handling everyday objects. In this paper, we consider a more challenging problem setting where only the category-level shape prior is provided.

Category-level object pose estimation and refinement. Both tasks mainly focus on addressing the shape variation between the objects. The pioneering work NOCS [49] tackles the shape discrepancy by recovering the normalized visible shape of the target object and achieving the pose by point cloud matching. A series of methods extend this structure by leveraging different information such as domain adaptation [21], different reconstruction space [3], shape prior [3, 16, 21, 45], and structural similarities [6, 25]. However, this line of work is often limited in speed due to the iterative point matching. Another series of work starts with FS-Net [9], which adopts 3D graph convolution (3D-GC) [26] to obtain geometric sensitivity. Due to its effectiveness and real-time performance, graph convolution is widely adopted in recent methods with an enhancement in directions including loss function [10], bounding box voting [58], and shape deformation [59]. HS-Pose [61] extends the geometric feature extraction from local to global, which enhances the capability to handle objects with complex shapes. The research on category-level refinement began recently with the proposal of CATRE [29]. It introduced an effective pipeline that leverages shape priors and a focalization strategy for pose refinement and effectively improves the initial pose estimations. In this paper, we extend the CATRE and tackle the geometric variation issue within the framework of category-level pose refinement.

3. Methodology

3.1. Problem Formulation

In this paper, we tackle the problem of category-level object pose refinement. Given the initial pose and size estimation ($\mathbf{R}_0$, $\mathbf{t}_0$, $\mathbf{s}_0$), along with the observed point cloud $\mathcal{O} \in \mathbb{R}^{N^O \times 3}$ and the shape prior $\mathcal{P} \in \mathbb{R}^{N^P \times 3}$, we aim to predict the estimation error ($\Delta\mathbf{R}$, $\Delta\mathbf{t}$, $\Delta\mathbf{s}$) between the initial estimations and the ground truths. The pose refinement algorithm ϕ can be described as:

$$(\Delta\mathbf{R}, \Delta\mathbf{t}, \Delta\mathbf{s}) = \phi(\mathbf{R}_0, \mathbf{t}_0, \mathbf{s}_0, \mathcal{O}, \mathcal{P}). \tag{1}$$

This pose refinement algorithm ϕ can be applied iteratively to improve the refinement performance.

3.2. Preliminaries

Our proposed category-level object pose refinement framework builds upon two previous works, CATRE [29] and HS-layer [61], which we briefly review them in the following.

CATRE. CATRE is the first framework that considers the problem of category-level pose refinement. It predicts the error between the ground truth and the estimated poses by aligning the input point clouds and the categorical shape priors. Specifically, the network architecture of CATRE consists of four components: a) point cloud focalization, b) shared encoder, c) rotation prediction, and d) translation and size prediction. In point clouds focalization, the observed point clouds $\mathcal{O}$ and the shape prior $\mathcal{P}$ are first aligned with the initial pose and size estimation $[\mathbf{R}_0, \mathbf{t}_0, \mathbf{s}_0]$:

$$\begin{aligned} \hat{\mathcal{O}} &= \{\hat{o}_i | \hat{o}_i = o_i - \mathbf{t}_0, o_i \in \mathcal{O}\}, \\ \hat{\mathcal{P}} &= \{\hat{p}_i | \hat{p}_i = \text{diag}(\mathbf{s}_0)\mathbf{R}_0 p_i, p_i \in \mathcal{P}\}, \end{aligned} \tag{2}$$

where diag$(\cdot)$ converts a vector to a diagonal matrix. The focalized observed point cloud $\hat{\mathcal{O}}$ and the focalized shape prior points $\hat{\mathcal{P}}$ contain full information required to predict the estimation error ($\Delta\mathbf{R}, \Delta\mathbf{t}, \Delta\mathbf{s}$). First, a PointNet-based shared encoder is used to extract features from the two focalized point clouds independently. Then, both the extracted features are used for $\Delta\mathbf{R}$ estimation, while the global feature of the focalized observed point cloud along with the $\mathbf{s}_0$ are used for $\Delta\mathbf{t}$ and $\Delta\mathbf{s}$. The initial estimates are updated using the predicted error $(\Delta\mathbf{R}, \Delta\mathbf{t}, \Delta\mathbf{s})$. Finally, the updated estimates are used to predict the error again in which this process is iterative and the estimations are refined progressively and continuously.

HS-layer. The Hybrid-Scope Geometric Feature Extraction Layer (HS-layer) is a simple network structure based on 3D graph convolution. It consists of two parallel paths that extract different scopes of features from the point cloud. The first path encodes the size and translation information of the target object. Meanwhile, the second path extracts outlier-robust local and global geometric features by applying graph convolution with a strategy of *Receptive Fields with Feature Distance (RF-F)* metric, alongside an *Outlier Robust Feature Extraction Layer (ORL)*. These properties are particularly beneficial for category-level object pose estimation tasks. For more details, please refer to [61].

3.3. Overall Structure of GeoReF

The overall framework of our proposed object pose refinement approach is shown in Fig. 2. This framework comprises three principal components: 1) point cloud focalization, 2) feature extraction, and 3) pose error prediction. We follow CATRE and use the same point cloud focalization module. We apply focalization and extract features from both the observed point cloud and the shape prior by our Feature Extraction component. Then, we predict the estimation errors $(\Delta\mathbf{R}, \Delta\mathbf{t}, \Delta\mathbf{s})$ using the extracted features in the pose error prediction component.

3.4. Graph Convolution with Learnable Affine Transformation (LAT)

Geometric structural information is effective in estimating an object's pose for category-level object pose estimations. However, as shown in ablation study [AS-1], directly applying the 3D graph convolutions (*e.g.*, HS-Encoder [61] and 3DGCN-Encoder [26]) to category-level object pose refinement tasks results in poor performance. This is due to the differences in task nature between the pose estimation and the pose refinement. In the pose estimation task, the network only needs to extract geometric structural information from a single point cloud. However, in the pose refinement framework, it requires extracting the geometric structural information from the two input point clouds as well as establishing the geometric correspondences between different object shapes. This becomes challenging due to the issue of shape variation in category pose refinement.

To address the aforementioned problem, we propose to use learnable affine transformations (LATs). By employing LATs, the network can dynamically adjust the input point cloud and the point features which enables better establishment of geometric correspondences between the two different input shapes. Specifically, we apply three LATs (as shown in the bottom left of Fig. 2, where the Matrix Net

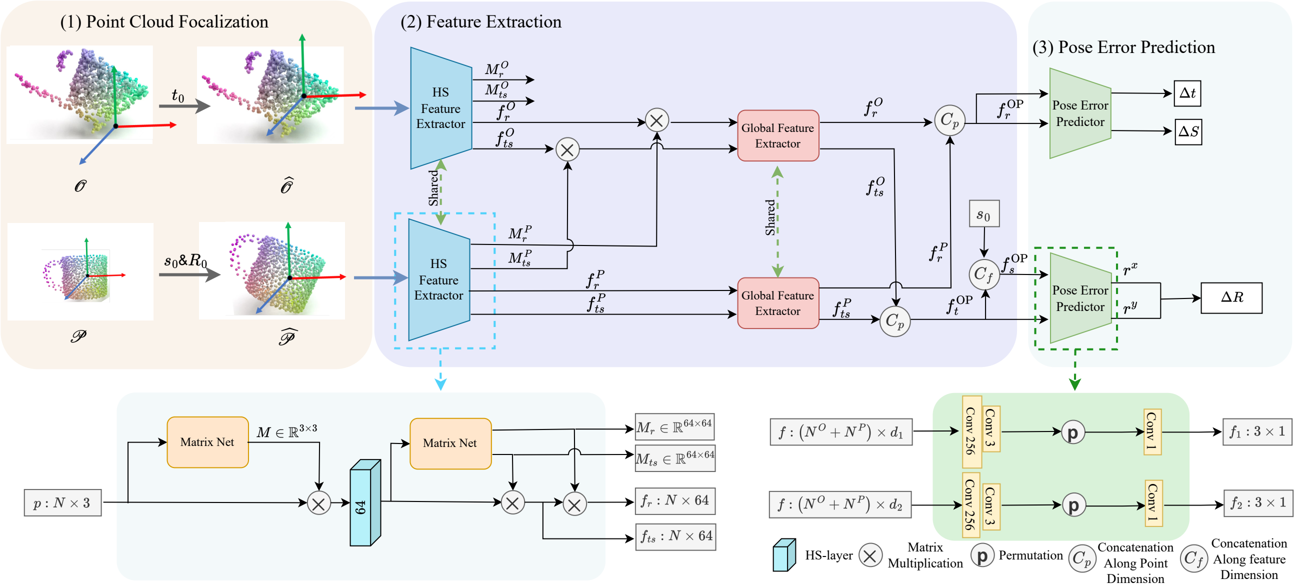


Figure 2. **Overall structure of the proposed method.** Our object pose refinement structure contains three main modules. Given the shape prior point cloud, the target object's observed point cloud, and the initial estimation, we first apply point cloud focalization on the input point clouds using the initial estimation. The focalized point clouds then go through a geometric-based feature extraction encoder to obtain geometric structural features. The extracted features are then fed into two branches for rotation error estimation, translation error, and size error estimation. Within the HS Feature Extractor, the Matrix Net models output the learnable affine transformations (LATs) for adaptive point and feature adjustment. The left output of the Matrix Net adjusts the input point clouds, while the right Matrix Net model outputs two affine transformations for adjusting the rotation features, and the translation and size features.

outputs the learnable affine transformations): The first LAT $M \in \mathbb{R}^3$ is applied to the input point cloud in the Euclidean space. The second LAT M_{ts} is applied to the extracted translation and size features f_{ts}. The third LAT M_r is applied to the extracted rotation feature f_r. With this approach, our method can better utilize the valuable geometric features in pose refinement.

3.5. Cross-Cloud Transformation (CCT) for Information Mixing

In pose refinement, effectively blending information from the focalized observed object and the shape prior is crucial for enabling the network to align them accurately. However, in CATRE, the data from the observed point cloud and the shape prior are processed independently until the late rotation prediction stage, where they are merely concatenated, limiting the effectiveness of the alignment. To address this problem, we introduce a novel cross-cloud transformation mechanism that effectively mixes the geometric information from the shape of prior features into the features of the observed point cloud. In particular, we use the feature transformation matrices M_r^P, M_{ts}^P of shape prior to transforming the features of the observed point cloud:

$$f_r^O = M_r^P f_r^O, \tag{3}$$

$$f_{\text{ts}}^O = M_{\text{ts}}^P f_{\text{ts}}^O. \tag{4}$$

3.6. Integrating Shape Prior in Pose Estimation

The information contained in the shape prior is crucial for the network to align the observed point cloud and the shape prior. For the rotation error prediction, the information contained in the shape prior is the essential information. For the translation and size prediction, this information can also be utilized by the network to adjust the learned geometric features accordingly. Therefore, unlike CATRE, which relied solely on features extracted from the observed point cloud to predict $(\Delta \mathbf{t}, \Delta \mathbf{s})$, our approach also incorporates the information from shape prior to predict them. In particular, we not only mix the information using previous CCT mechanism, but also concatenate the features from shape prior and observed point cloud to obtain mixed features in the similarly way as the rotation estimation. We utilize f_{t}^{OP} and f_{s}^{OP} which contain both information from shape prior and observed point cloud like f_r^{OP} to predict $(\Delta \mathbf{t}, \delta)$.

We use two pose error predictors of the same network architecture to predict the rotation error and the translation and size error, respectively. Note that the weights of these two pose error predictors are not shared. The network structure of the pose error predictor is shown in Fig. 2. The pose predictor takes in two features, passes them through two same paths separately, and obtains two vectors in $\mathbb{R}^3$. In the translation and size branch, the pose error predictor takes in f_{t}^{OP} and f_{s}^{OP} and passes them through the two paths, and the output two vectors are regarded as $\Delta \mathbf{t}$ and $\Delta \mathbf{s}$, respectively. For the rotation error prediction, the mixed rotation features f_r^{OP} are copied and passed through the two paths in the pose error predictor. The two output vectors are regarded as r_x and r_y, where r_x and r_y are the first and second axes of the rotation error matrix $\Delta \mathbf{R}$. The third column r_z of $\Delta \mathbf{R}$ can be found by:

$$r_z = r_x \times r_y. \tag{5}$$

4. Experiments

Implementation details. We implement and experiment with our method using an RTX 4090 GPU with a batch size of 12 and 150 training epochs. We follow CATRE [29] and adopt its loss functions and the basic data augmentation strategies including random dropping points, adding Gaussian noise, random pose perturbations, etc. We set the number of points for both the observed points and shape prior to be 512. We train the network using Ranger optimizer [28, 54, 57] with a base learning rate of 10^{-4} and anneal the learning rate from the 72% of the total epoch based on cosine schedule.

Baselines. We use CATRE as the baseline for our ablation study as it is the state-of-the-art category-level object refinement method. As CATRE did not provide the results of IoU$_{50}$, we obtained them by using their official pre-trained model and kept the rest of the reported metric scores consistent with the corresponding paper. For fair comparisons, we use the same initial estimations as CATRE, which is the pose estimation results of SPD [45]. The result of replacing PointNet with the 3DGC-Encoder in the ablation study is provided by CATRE. We also apply our method to other state-of-the-art category-level object pose estimation approaches [10, 25, 61] to demonstrate the effectiveness of our proposed refinement method. For the pose refinement on HS-Pose [61] and RBP-Pose [58], we compute the initial estimations using their official pre-trained models. The results of other methods are taken directly from their paper.

Datasets. As we focus on the problem of shape variation between input object point clouds, we choose two popular category-level object pose estimation benchmarks to verify our approach, *i.e.*, REAL275 [49] and CAMERA25 [49]. They both contain 6 object categories with multiple levels of shape complexities, *i.e.*, bowl, can, bottle, laptop, mug, and camera. REAL275 contains 36 objects in 13 real-world scenes with 7k RGB-D images in total. Among them, 16 objects in 7 scenes are used for training, resulting in 4.3k images in training. CAMERA25 is a large synthetic RGB-D dataset. It provides 1085 objects and 275k RGB-D images for training, and 184 objects and 25k images for testing.

Evaluation metrics. Following [29, 61], we evaluate our method using: 1) The mean average precision (mAP) of the *3D Intersection over Union (IoU)* at different thresholds (50% and 75%) to evaluate the pose and size estimation together[2]. 2) The pose metric at $n°m$cm defines a pose as correct if the rotation error is below $n°$ and the translation error is below m cm. Here, we use 5°, 10°, 2cm, and 5cm as the thresholds.

4.1. Ablation study

To verify the proposed architecture, we conducted comprehensive ablation studies on the REAL275 dataset using the initial pose estimations from SPD [45]. We present a quantitative comparison of our method with various key components disabled to motivate our design choices in Table 1. Full results of the ablation study are reported in the supplementary materials.

[AS-1] Using geometric features directly. To illustrate the limitations of existing geometric-based encoder under the problem of shape variations, we replace the encoder of CATRE with two robust geometric-based point cloud convolutional structures, namely 3DGCN-Encoder [26] and HS-Encoder [61]. 3DGCN is a widely adopted graph convolution in existing category-level object pose estimation algorithms, while HS-Encoder is a recent architecture that achieves state-of-the-art performance in category-level object pose estimation. However, as shown in Table 1 [C0, C1], even though both HS-Encoder and 3DGCN-Encoder are powerful in finding an object's pose from individual inputs, they failed to manage the pose refinement scenarios when there exist shape variations between the target object and the shape prior. We observed both encoders result in a performance drop when compared to the original CATRE with IoU$_{75}$ of 37.3% (HS-Encoder) *vs*. 43.7% (CATRE), 5°5cm of 43.4% (3DGCN-Encoder) *vs*. 53.3% (CATRE).

[AS-2] Use prior features in translation and scale estimation. To validate that the information of shape prior is also important in scale and translation error prediction, we add the features of the prior shape to the scale and translation branch by using the same network architecture as the rotation branch. As shown in Table 1 [D0], incorporating shape prior information in translation and size estimation enhanced the overall performance by 2.2% improvement on 5°2cm metric and 2.4% on 10°2cm metric.

[AS-3] Use learnable affine transformation (LAT) for geometric features. To demonstrate the effectiveness of LAT in addressing the shape variation issue, we conduct ablation studies on applying the proposed LAT to the input point cloud and the extracted geometric features in the feature space. The result is shown in Table 1 [E0]. Compared to using geometric features directly (Table 1 [C0]), LATs bring a significant boost on all the metrics, with IoU$_{75}$ improved by **20.9%**, 5°2cm improved by **10.5%**,

[2] Note that there was a small mistake with the IoU computation from the original benchmark evaluation code [49], we follow [29] to recalculate the IoU metrics for the SOTA methods.

Table 1. **Ablation studies on REAL275.**
Higher score indicates better performance. In the 'Row' column, the code in bold means the strategies taken in the final structure. In the 'Method' column, the notation 'X:Y' denotes module Y from structure X, 'X+Y' means add module Y to X, and '$X \rightarrow Y$' indicates replacing X with Y.

Row	Method	IoU$_{50}$	IoU$_{75}$	5°2cm	5°5cm	10°2cm	10°5cm	2cm	5°
A0	CATRE[45] (baseline)	77.0	43.6	45.8	54.4	61.4	73.1	75.1	58.0
B0	**Ours**: E0 + Cross-Cloud Transformation	**79.2** 2.2↑	**51.8** 8.2↑	**54.4** 8.6↑	**60.3** 5.9↑	**71.9** 10.5↑	**79.4** 6.3↑	**81.9** 6.8↑	**64.3** 6.3↑
C0	A0: PointNet → HS-Encoder	71.0	30.1	41.9	45.9	60.6	70.3	71.9	48.7
C1	A0: PointNet → 3DGCN-Encoder	-	28.4	36.0	43.4	-	-	68.0	47.7
D0	A0 + prior in ST branch	77.1	45.8	48.0	54.6	63.8	72.5	77.9	59.2
E0	D0: PointNet → HS-layer+LAT	79.4	51.0	52.4	58.6	69.4	77.7	80.4	62.4
E1	B0: No LAT on input points	76.1	39.3	46.6	53.0	65.4	74.8	78.0	58.2
E2	B0: No LAT on features	78.5	48.8	47.4	53.0	67.4	75.0	80.4	57.4
E3	B0: No LAT on the rotation feature	79.8	50.6	50.4	56.2	68.6	76.3	80.2	60.8
F0	E0+ Global Concatenation Fusion	77.7	48.4	47.8	54.5	67.1	75.2	80.1	59.4

and $5^\circ 5$cm improved by **12.7%**. The resulted network also significantly outperforms the PointNet-based encoder (Table 1[D0]) on all the metrics with IoU$_{75}$ and $10^\circ 5$cm improved by **5.2%**, $5^\circ 2$cm improved by **6.4%**, and $10^\circ 2$cm improved by **5.6%**. These results verified the effectiveness of LAT on geometric features.

[AS-4] The influence of each learnable affine transformation (LAT). To further demonstrate the influences of each LAT, we conducted three experiments by gradually disabling LAT from the framework: 1) without the LAT on the input point cloud, 2) without applying LATs on features, and 3) without independent LATs on the rotation features, where a single LAT is used for the rotation, translation, and scale features. The results are shown in Table 1 [E1-E3]. Compared to the directly using geometric features (Table 1 [C0]), we show that LAT can significantly enhance the network by around 10% improvements on IoU$_{75}$ and 7% on $5^\circ 5$cm metric in Table 1 [E1-E3]. We verify that the combination of them consistently results in better performance.

[AS-5] Cross-cloud transformation (CCT) based feature fusion To demonstrate that it is important to have a good feature fusion strategy in the feature extraction phase, we conducted experiments on two different feature fusion strategies. One of them is the widely adopted feature fusion strategy, where the global feature of one point cloud is concatenated with the features of another point cloud and then goes through convolutional layers for feature fusion. Another one is the proposed CCT-based feature fusion. As shown in Table 1 [F0], applying global concatenation-based fusion does not enhance the overall performance, and even results in worse performance in all the metrics with the $5^\circ 2$cm significantly decreased by 4.6%, and $5^\circ 5$cm decreased by 3.1% (compared to Table 1 [D0]). On the contrary, as shown in Table 1 [B0], our simple CCT-based feature fusion shows its effectiveness by improving all the pose metrics by around 2.0%. In Supplementary, we also visualize the influence of CCT in feature space.

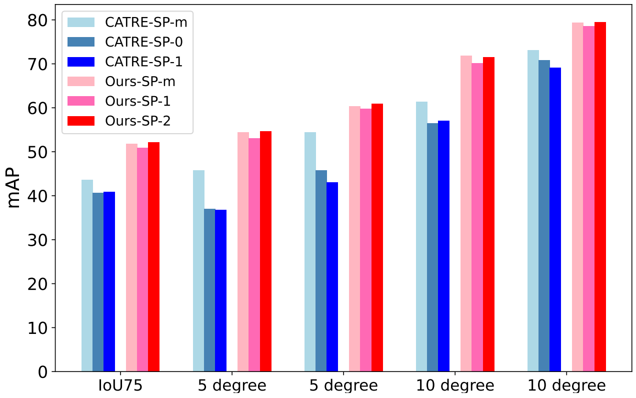


Figure 3. **Performance comparison of our method and CATRE under different shape priors.** *SP-m* denotes the category's mean shape, while *SP-1* and *SP-2* are randomly sampled object shapes from CAMERA25.

[AS-6] Handle shape variations. To demonstrate that the proposed method can handle the shape variations, we replaced the original shape prior with two randomly sampled models from CAMERA25 training set and trained on REAL275. Note that the original shape prior represents the mean shape of a category, and the new models are randomly sampled. Therefore, there are larger shape variations with certain target objects. As shown in Fig. 3, CATRE performs best when using the mean shape of the category, while its performance drops dramatically on the randomly sampled shape priors. In contrast, our method exhibits robustness to shape variations introduced by different shape priors and consistently delivers strong performance.

[AS-7] Refinement with different initial estimations. To demonstrate our model robustness to different initial estimations, we compare the proposed method with CATRE on different initial estimations generated by category-level

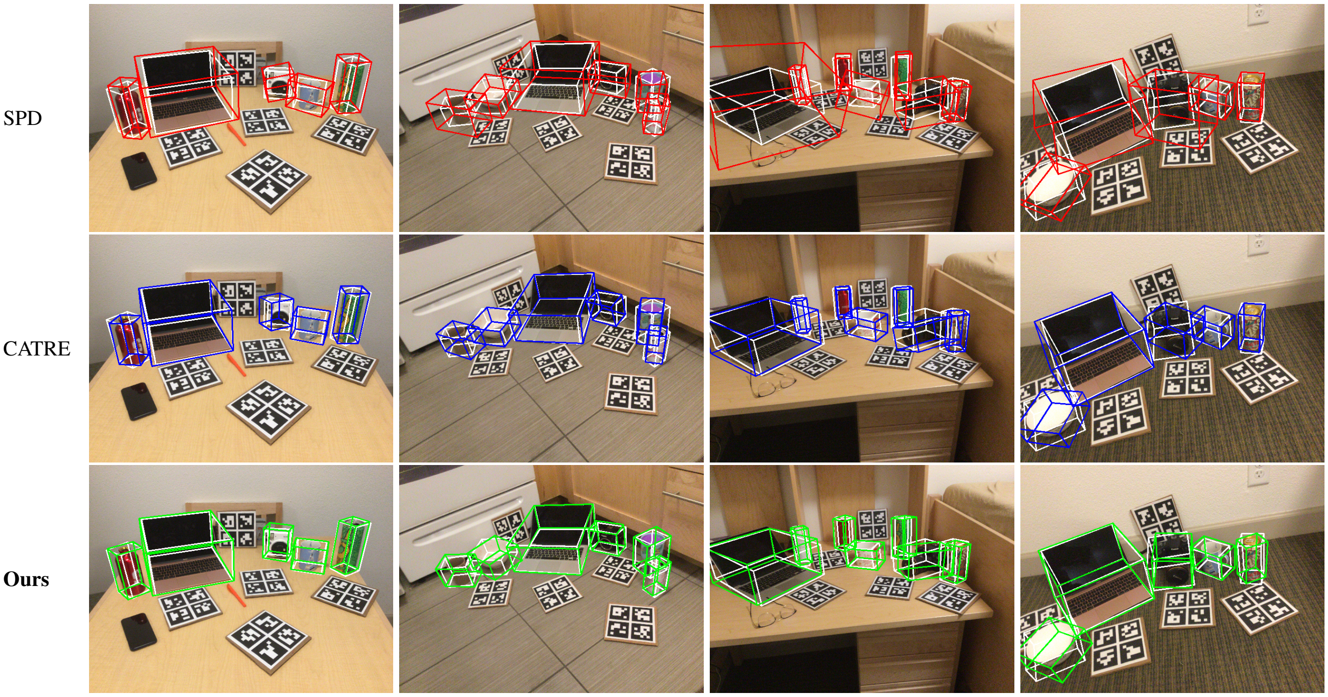


Figure 4. **Qualitative comparison of proposed (row #3) and baseline (row #2) methods using SPD (row #1) as initial estimation.** Ground truth shown with white lines. Note that the estimated rotations of symmetric objects (*e.g*. bowl, bottle, and can) are considered correct if the symmetry axis is aligned.

object pose estimation methods with varying performance, including HS-Pose [61], GPVPose [10], and the Self-DPDN [25]. As shown in Table 2, our method consistently improves the initial estimations by a large margin across different metrics. For example, we improved the SelfDPND and GPV-Pose on strict 5°2cm metric by 9.6% and 15.4%, respectively. We boosted the IoU$_{75}$ of the GPV-Pose and the HS-Pose by 26.5% and 15.2%, respectively. However, the baseline CATRE fails to refine the initial poses of Self-DPDN. Also, CATRE reaches its limit when refining high-accuracy initial estimations such as the initial poses generated by HS-Pose, resulting in performance drops on 10°5cm and 10°2cm. The comparison results demonstrated the robustness and capability of our method to different initial estimations. We report full details in the supplementary.

4.2. Generalizabily test on the CAMERA25 dataset

In real-world applications, category-level algorithms often need to generalize across diverse testing scenarios, encountering a larger number of objects than represented in their training sets. To simulate this problem setting, we choose CAMERA25 as it provides more than 25K RGB-D testing images. We train our model using only a mini set (2% and 4%) of the CAMERA25 training data, resulting in a training set of around 5K and 10K images from a total of 275K images. As shown in Table 3, the performance of the baseline method decreased dramatically when using 2% of the training images, with IoU$_{75}$ dropped by **12.9%** and 5°2cm dropped by **9.0%** (see [B0]). On the contrary, our method exhibits a much higher performance when using small datasets for training. Specifically, our method can already outperforms the fully-trained CATRE with only **2%** of images (see [B1]). We also report the performance using 4% images of the training set in Table 3. It can be seen that, despite CATRE use full CAMERA25 training set, our method outperforms it with IoU$_{75}$ and 5°2cm improved by 3.1% and 3.6%, respectively.

Table 2. **Comparison of the pose refinement with the baseline method CATRE on REAL275 with different initial estimations.** Each comparison group contains 3 methods: the initial pose estimation method, refinement using CATRE, and refinement using our method, respectively.

Method	IoU$_{75}$	5°2cm	5°5cm	10°2cm	10°5cm
Self-DPDN [25]	42.2	44.3	50.9	65.1	78.6
Self-DPDN + CATRE	0.0 42.2↓	0.3 44.0↓	5.1 45.8↓	0.4 64.7↓	6.9 71.7↓
Self-DPDN + Ours	49.7 7.5↑	53.9 9.6↑	60.1 9.2↑	75.0 9.9↑	82.8 4.2↑
GPV-Pose [10]	23.1*	32.0	42.9	55.0	73.3
GPV-Pose + CATRE	42.6 19.5↑	39.7 7.7↑	54.1 11.2↑	57.1 2.1↑	78.0 4.7↑
GPV-Pose + Ours	49.6 26.5↑	47.4 15.4↑	57.8 14.9↑	68.1 13.1↑	81.2 7.9↑
HS-Pose [61]	39.1*	46.5	55.2	68.6	82.7
HS-Pose + CATRE	47.1 8.0↑	48.7 2.2↑	59.1 3.9↑	67.8 0.8↓	81.2 1.5↓
HS-Pose + Ours	54.3 15.2↑	51.7 5.2↑	59.6 4.4↑	74.3 5.7↑	83.8 1.1↑

4.3. Comparison with state-of-the-arts

REAL275. We conduct pose refinement on SPD [45] using the proposed approach and compare the resulting performance with the state-of-the-art category-level object pose estimation and refinement methods. As shown in Table 4, our method significantly improves the performance of SPD on all the metrics, with 5°5cm enhanced

Table 3. **The generalizability test on the CAMERA25 dataset.**

Row	Method	Train Data Size	IoU$_{75}$	5°2cm	5°5cm	10°2cm	10°5cm	2cm	5°
A0	CATRE	275K	76.1	75.4	80.3	83.3	89.3	-	-
B0	CATRE	5K	63.2	66.4	72.3	79.4	87.4	88.8	73.3
B1	**Ours**	5K	77.5	75.4	81.1	83.4	90.0	91.0	82.3
C0	CATRE	10K	66.5	69.7	75.5	81.8	89.1	89.9	76.7
C1	**Ours**	10K	**79.2**	**77.9**	**84.0**	**83.8**	**90.5**	**92.0**	**85.4**

Table 4. **Comparison with other methods on REAL275.**

Method	IoU$_{75}$	5°2cm	5°5cm	10°2cm	10°5cm
NOCS [49]	9.4	7.2	10.0	13.8	25.2
DualPoseNet [24]	30.8	29.3	35.9	50.0	66.8
CR-Net [50]	33.2	27.8	34.3	47.2	60.8
SGPA [6]	37.1	35.9	39.6	61.3	70.7
RBP-Pose [58]	24.5	38.2	48.1	63.1	79.2
GPV-Pose [10]	23.1	32.0	42.9	55.0	73.3
HS-Pose [61]	39.1	46.5	55.2	68.6	**82.7**
SPD* [45]	27.0	19.1	21.2	43.5	54.0
SPD*+CATRE [29]	43.6	45.8	54.4	61.4	73.1
SPD*+**Ours**	**51.8**	**54.4**	**60.3**	**71.9**	79.4

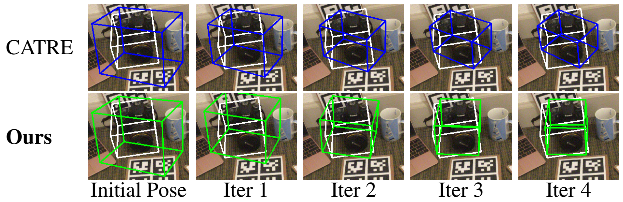


Figure 5. **Comparison of proposed (row #2) and baseline (row #1) methods) during a complete refinement iteration, both using SPD as initial estimation.** The ground truth is represented by white lines.

by **39.1%**, 5°5cm improved by **35.3%**, IoU$_{75}$ enhanced by **24.8%**, 10°2cm improved by **24.4%**, and 10°5cm improve by **25.4%**. In comparison to our baseline, CATRE, our proposed method demonstrates a substantial improvement across various performance metrics. Specifically, we observe a remarkable enhancement of **10.5%** in 10°2cm, **8.6%** in 5°2cm, **7.2%** in IoU$_{75}$, and **6.3%** in 10°5cm. In addition, compared with SOTA pose estimation methods, the pose estimation results of applying our proposed refinement method on SPD significantly outperformed these methods's results by a large margin. Specifically, the estimation results of applying our method on SPD rank top on 4 out of 5 metrics and rank second on the rest metric, and achieved a **7.9%** increase on the 5°2cm metric and **5.1%** enhancement on 5°5cm. It is worth noting that the purpose of this section is not to compare refinement and estimation methods, as they are designed to address different problems. Instead, our objective here is to demonstrate how our proposed refinement approach can improve the performance of existing pose estimation methods. Therefore, even though our refinement on HS-Pose produced better performance, as mentioned in the ablation study, we choose to refine weaker initial estimations to show the capability of our approach.

CAMERA25. Our method outperforms the state-of-the-art methods using only 2% of the training image set. For more details, please refer to the supplementary.

Qualitative examples. We provide qualitative comparisons of the pose estimation results by SPD, CATRE on SPD, and our method on SPD in Fig. 4 and 5. As shown in Fig. 4, our method on SPD achieves the best size and pose estimations. In particular, by considering the first column of Fig. 4, all of the comparison methods struggle to estimate the orientation of the camera category. We also provide qualitative examples on the iteration process in Fig. 5. Our method demonstrates faster convergence and more accurate final result than CATRE.

5. Conclusion

In this work, we proposed a novel category-level object pose refinement method which targeted at addressing the challenge of shape variation. We shown that the geometric structural information can be aligned by our adaptive affine transformations. We also demonstrated that the cross-cloud transformation mechanism can efficiently merges information from distinct point clouds. We further incorporated shape prior information and observed improvements in translation and size predictions. We verified that each of our technical components contributed meaningfully through extensive ablations. We believe our method sets a strong baseline for future study and opens up new possibilities to handling more complex shapes, *i.e.* articulated objects.

Acknowledgments

This research was supported in part by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-2020-0-01789) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), National Natural Science Foundation of China under Grant No.62073159, and the Shenzhen Key Laboratory of Control Theory and Intelligent Systems, under grant No.ZDSYS20220330161800001.

References

[1] Dingding Cai, Janne Heikkilä, and Esa Rahtu. Ove6d: Object viewpoint encoding for depth-based 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6803–6813, 2022. 2

[2] Pedro Castro and Tae-Kyun Kim. Crt-6d: Fast 6d object pose estimation with cascaded refinement transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5746–5755, 2023. 1

[3] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[4] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation, 2022. 2

[5] Hanzhi Chen, Fabian Manhardt, Nassir Navab, and Benjamin Busam. Texpose: Neural texture learning for self-supervised 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4841–4852, 2023. 1

[6] Kai Chen and Qi Dou. SGPA: Structure-guided prior adaptation for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2773–2782, 2021. 2, 8

[7] Kai Chen, Stephen James, Congying Sui, Yun-Hui Liu, Pieter Abbeel, and Qi Dou. Stereopose: Category-level 6d transparent object pose estimation from stereo images via back-view nocs. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2855–2861, 2023. 1

[8] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, and Ales Leonardis. G2L-Net: Global to Local Network for Real-Time 6D Pose Estimation With Embedding Vector Features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[9] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Ales Leonardis. Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1581–1590, 2021. 1, 2

[10] Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. GPV-Pose: Category-level object pose estimation via geometry-guided point-wise voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6781–6791, 2022. 1, 2, 5, 7, 8

[11] Yang Hai, Rui Song, Jiaojiao Li, Mathieu Salzmann, and Yinlin Hu. Rigidity-aware detection for 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8927–8936, 2023. 1

[12] Rasmus Laurvig Haugaard and Anders Glent Buch. Surfemb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings. *CoRR*, abs/2111.13489, 2021. 2

[13] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[14] Stefan Hinterstoisser, Vincent Lepetit, Naresh Rajkumar, and Kurt Konolige. Going further with point pair features. In *Computer Vision – ECCV 2016*, pages 834–848. Springer International Publishing, 2016. 2

[15] Tomas Hodan, Daniel Barath, and Jiri Matas. Epos: Estimating 6d pose of objects with symmetries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[16] Muhammad Zubair Irshad, Sergey Zakharov, Rares Ambrus, Thomas Kollar, Zsolt Kira, and Adrien Gaidon. Shapo: Implicit representations for multi object shape appearance and pose optimization. 2022. 2

[17] Shun Iwase, Xingyu Liu, Rawal Khirodkar, Rio Yokota, and Kris M Kitani. Repose: Fast 6d object pose refinement via deep texture rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3303–3312, 2021. 1

[18] Daniel Kappler, Franziska Meier, Jan Issac, Jim Mainprice, Cristina Garcia Cifuentes, Manuel Wüthrich, Vincent Berenz, Stefan Schaal, Nathan Ratliff, and Jeannette Bohg. Real-time perception meets reactive motion generation. *IEEE Robotics and Automation Letters*, 3(3):1864–1871, 2018. 1

[19] Nikunj Kothari, Misha Gupta, Leena Vachhani, and Hemendra Arya. Pose estimation for an autonomous vehicle using monocular vision. pages 424–431, 2017. 1

[20] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. CosyPose: Consistent multi-view multi-object 6D pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[21] Taeyeop Lee, Byeong-Uk Lee, Inkyu Shin, Jaesung Choe, Ukcheol Shin, In So Kweon, and Kuk-Jin Yoon. UDA-COPE: unsupervised domain adaptation for category-level object pose estimation. *CoRR*, abs/2111.12580, 2021. 2

[22] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[23] Haitao Lin, Zichang Liu, Chilam Cheang, Yanwei Fu, Guodong Guo, and Xiangyang Xue. Sar-net: Shape alignment and recovery network for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6707–6717, 2022. 1

[24] Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. DualPoseNet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision (ICCV), pages 3560–3569, 2021. 8

[25] Jiehong Lin, Zewei Wei, Changxing Ding, and Kui Jia. Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks, 2022. 1, 2, 5, 7

[26] Zhi-Hao Lin, Sheng-Yu Huang, and Yu-Chiang Frank Wang. Convolution in the cloud: Learning deformable kernels in 3d graph convolution networks for point cloud analysis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1797–1806, 2020. 1, 2, 3, 5

[27] Zheng Linfang, Leonardis Ales, Tse Tze Ho, Elden, Horanyi Nora, Chen Hua, Zhang Wei, and Chang Hyung Jin. Tp-ae: Temporally primed 6d object pose tracking with auto-encoders. In *2022 IEEE International Conference on Robotics and Automation (ICRA)*, 2022. 2

[28] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond, 2019. 5

[29] Xingyu Liu, Gu Wang, Yi Li, and Xiangyang Ji. CATRE: Iterative point clouds alignment for category-level object pose refinement. In *European Conference on Computer Vision (ECCV)*, pages 499–516. Springer, 2022. 1, 2, 3, 5, 8

[30] Eitan Marder-Eppstein. Project tango. pages 25–25, 2016. 1

[31] Nathaniel Merrill, Yuliang Guo, Xingxing Zuo, Xinyu Huang, Stefan Leutenegger, Xi Peng, Liu Ren, and Guoquan Huang. Symmetry and uncertainty-aware object slam for 6dof object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14901–14910, 2022. 2

[32] Andrew S. Morgan, Bowen Wen, Junchi Liang, Abdeslam Boularias, Aaron M. Dollar, and Kostas Bekris. Vision-driven Compliant Manipulation for Reliable; High-Precision Assembly Tasks. In *Proceedings of Robotics: Science and Systems*, Virtual, 2021. 1

[33] Andrew Nee, S K Ong, George Chryssolouris, and Dimitris Mourtzis. Augmented reality applications in design and manufacturing. *CIRP Annals - Manufacturing Technology*, 61:657–679, 2012. 1

[34] Van Nguyen Nguyen, Yinlin Hu, Yang Xiao, Mathieu Salzmann, and Vincent Lepetit. Templates for 3d object pose estimation revisited: Generalization to new objects and robustness to occlusions, 2022. 2

[35] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[36] Mahdi Rad and Vincent Lepetit. BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects Without Using Depth. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[37] Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. Generalized-icp. In *Robotics: science and systems (RSS)*, page 435. Seattle, WA, 2009. 1

[38] Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic. Osop: A multi-stage one shot object pose estimation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6835–6844, 2022. 2

[39] Yongzhi Su, Jason Rambach, Nareg Minaskan, Paul Lesur, Alain Pagani, and Didier Stricker. Deep multi-state object pose estimation for augmented reality assembly. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 222–227, 2019. 1

[40] Yongzhi Su, Mahdi Saleh, Torben Fetzer, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari. Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation, 2022. 2

[41] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6825–6834, 2022. 2

[42] Martin Sundermeyer, Maximilian Durner, En Yen Puang, Zoltan-Csaba Marton, Narunas Vaskevicius, Kai O. Arras, and Rudolph Triebel. Multi-path learning for object pose estimation across domains, 2019. 2

[43] Martin Sundermeyer, Zoltan Marton, Maximilian Durner, and Rudolph Triebel. Augmented Autoencoders: Implicit 3D Orientation Learning for 6D Object Detection. *International Journal of Computer Vision (IJCV)*, 128, 2019. 2

[44] Bugra Tekin, Sudipta Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. pages 292–301, 2018. 2

[45] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *European Conference on Computer Vision (ECCV)*, pages 530–546. Springer, 2020. 2, 5, 6, 7, 8

[46] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects, 2018. 2

[47] Joel Vidal, Chyi-Yeu Lin, and Robert Martí. 6d pose estimation using an improved method based on point pair features, 2018. 2

[48] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martin-Martin, Cewu Lu, Li Fei-Fei, and Silvio Savarese. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[49] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2642–2651, 2019. 2, 5, 8

[50] Jiaze Wang, Kai Chen, and Qi Dou. Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4807–4814. IEEE, 2021. 8

[51] Bowen Wen, Chaitanya Mitash, Baozhang Ren, and Kostas Bekris. se(3)-TrackNet: Data-driven 6D Pose Tracking by Calibrating Image Residuals in Synthetic Domains. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020. 1, 2

[52] Bowen Wen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal. You Only Demonstrate Once: Category-Level Manipulation from Single Visual Demonstration. In *Proceedings of Robotics: Science and Systems*, New York City, NY, USA, 2022. 1

[53] Heng Yang and Marco Pavone. Object pose estimation with statistical guarantees: Conformal keypoint detection and geometric uncertainty propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8947–8958, 2023. 1

[54] Hongwei Yong, Jianqiang Huang, Xiansheng Hua, and Lei Zhang. Gradient centralization: A new optimization technique for deep neural networks, 2020. 5

[55] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. DPOD: 6D Pose Object Detector and Refiner. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[56] Kun Zhang, Chen Wang, Hua Chen, Jia Pan, Michael Yu Wang, and Wei Zhang. Vision-based six-dimensional peg-in-hole for practical connector insertion. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1771–1777, 2023. 1

[57] Michael R. Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba. Lookahead optimizer: k steps forward, 1 step back, 2019. 5

[58] Ruida Zhang, Yan Di, Zhiqiang Lou, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. RBP-Pose: Residual bounding box projection for category-level pose estimation, 2022. 1, 2, 5, 8

[59] Ruida Zhang, Yan Di, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. SSP-Pose: Symmetry-aware shape prior deformation for direct category-level object pose estimation, 2022. 1, 2

[60] Zhongqun Zhang, Wei Chen, Linfang Zheng, Aleš Leonardis, and Hyung Jin Chang. Trans6d: Transformer-based 6d object pose estimation and refinement. In *Computer Vision – ECCV 2022 Workshops*, pages 112–128, Cham, 2023. Springer Nature Switzerland. 1

[61] Linfang Zheng, Chen Wang, Yinghan Sun, Esha Dasgupta, Hua Chen, Aleš Leonardis, Wei Zhang, and Hyung Jin Chang. HS-Pose: Hybrid scope feature extraction for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17163–17173, 2023. 1, 2, 3, 5, 7, 8

[62] Jun Zhou, Kai Chen, Linlin Xu, Qi Dou, and Jing Qin. Deep fusion transformer network with weighted vector-wise keypoints voting for robust 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13967–13977, 2023. 2

Supplementary material for GeoReF: Geometric Alignment Across Shape Variation for Category-level Object Pose Refinement

Linfang Zheng[1,3] Tze Ho Elden Tse[* 3] Chen Wang[* 1,2] Yinghan Sun[1] Hua Chen[1]
Aleš Leonardis[3] Wei Zhang[† 1] Hyung Jin Chang[3]

[1]Shenzhen Key Laboratory of Control Theory and Intelligent Systems, School of System Design and Intelligent Manufacturing, Southern University of Science and Technology, China
[2]Department of Computer Science, the University of Hong Kong, China
[3]School of Computer Science, University of Birmingham, UK
{lxz948, txt994}@student.bham.ac.uk, cwang5@cs.hku.hk, sunyh2021@mail.sustech.edu.cn
{chenh6,zhangw3}@sustech.edu.cn,{a.leonadis,h.j.chang}@bham.ac.uk

1. About the Runtime

On a machine with an Intel 13900k CPU and a Nvidia RTX 4090 GPU, the speed of our proposed method is 67.5 FPS for 1 iteration, and 22.3 FPS when using 4 iterations.

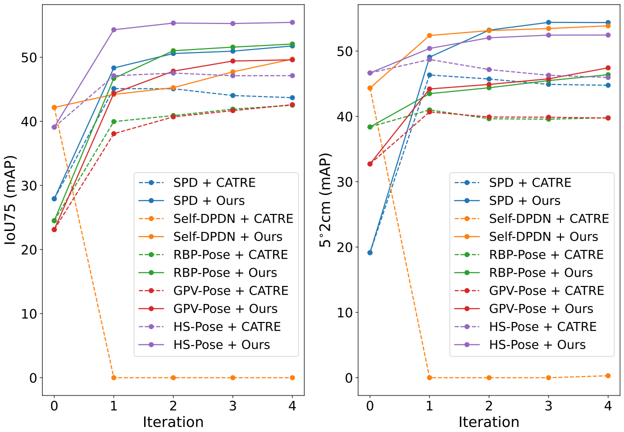
2. Effect of number of iterations

We find that the performance of our proposed method saturates after 4 iterations. Therefore, we set the iteration number to 4 for our experiments. We provide a line graph to show the performance changes of our method and CATRE [6] during the iteration in Fig. 1 We show that our proposed method consistently outperforms the baseline method and saturates after 4 iterations in both figures.

3. Ablation Studies

Refinement with different initial estimations. Apart from the table provided in the main paper, we visually show the robustness of our method on different initial estimations generated from 5 pose estimation methods [2, 5, 8, 11, 13] with ranging performance. As shown in Fig. 1, our method keeps improving the performance of the initial estimations, while CATRE [6] failed when refining the initial estimations from Self-DPDN [5]. Additionally, our method keeps improving during the iterations, while CATRE's performance starts to decrease after one iteration (see the dashed lines in Fig. 1).

The effect of CCT. To demonstrate the effect of CCT, we show a statistics plot of feature distances before and after CCT on objects with different shape complexities of the CAMERA25 test set. In this experiment, the initial pose of the shape prior is aligned with the ground truth pose to guarantee that the observed variations in feature distance are solely attributable to differences in shape. As shown in Fig 2, the feature distance between the shape prior and the input target shrinks significantly after applying CCT.



(a) (b)

Figure 1. **Comparison between CATRE and our method on different initial estimations across different refining iterations.** (a) IoU$_{75}$ performance comparison. (b) 5°2cm performance comparison. Our methods are shown in solid lines and CATRE's are in dashed lines. Iteration 0 shows the performance of the initial estimations.

4. Generalizability test on CAMERA25

More Results. To test the generalizability of our method when trained on a small dataset and tested on a large dataset, we randomly sample datasets from the CAMERA25 train-

[*] Equal contribution, order by dice rolling.
[†] The corresponding author.

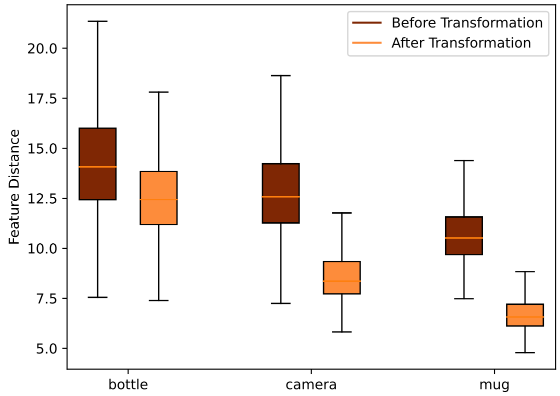


Figure 2. **Feature distances between the shape prior and the input point cloud before and after applying the cross-cloud transformation.**

ing dataset at different ratios (2%, 4%, and 6%). This yields training sizes of 5k, 10k, 15k. We show the results of the generalizability test in Table 1. We observe that our method, trained only using 2% of the train set, can already outperform a fully trained CATRE on all training data. Also, our performance becomes stable when using 4% of the train set (see Table 1 [C1, D1]), while CATRE requires additional training data for better performance. Since our performance became stable, we did not test on larger data sizes.

Experiment settings: To ensure the distribution of different categories in the sampled mini datasets, we control the image number of each object in the sampled datasets: 1) 5 images per object for the 2% train set, 2) 10 images for the 4% train set, and 3) 15 images for the 6% train set.

Table 1. **The generalizability test on the CAMERA25 dataset.** Higher score means better performance. Overall best results are in bold, and the second-best results are underlined. The training data size is denoted as *T. Size*.

Row	Method	T. Size	IoU_{75}	5°2cm	5°5cm	10°2cm	10°5cm
A0	CATRE	275k	76.1	75.4	80.3	83.3	89.3
B0	CATRE	5k	63.2	66.4	72.3	79.4	87.4
B1	Ours	5k	77.5	75.4	81.1	83.4	<u>90.0</u>
C0	CATRE	10k	66.5	69.7	75.5	81.8	89.1
C1	Ours	10k	**79.2**	<u>77.9</u>	<u>84.0</u>	**83.8**	**90.5**
D0	CATRE	15k	69.7	73.2	78.8	82.6	89.4
D1	Ours	15k	<u>78.1</u>	**78.0**	**84.1**	<u>83.6</u>	**90.5**

5. Detailed Network Architectures

The network structure of the HS Feature Extractor and the Pose Error Predictor is shown in Fig. 2 of the main paper. The structure of the Pose Error Predictor for $\Delta\mathbf{R}$ estimation and the $\Delta\mathbf{t}, \Delta\mathbf{s}$ estimation are identical, we follow the CATRE [6] and use 3 Convolution-1D layers with permutation before the final layer to generate the pose errors. For the Matrix Net, we follow PointNet [7] first use 3 Convolution-1D layers with $[64, 128, 1024]$ output dimensions and a kernel size of 1 to extract the dense point features, then the features going through a maximum pooling layer and 3 liner layers with [512, 256, $\boldsymbol{f}_{\text{LAT}}$] to generate the matrix. For the first Matrix Net that generates the adaptive affine transformation (LAT) for the input point cloud, $\boldsymbol{f}_{\text{LAT}}$ is 9. For the second Matrix Net, $\boldsymbol{f}_{\text{LAT}}$ is 8192, as it outputs two LATs with the matrix size of $\mathbb{R}^{64\times64}$. In the final structure of the GeoReF, we use two HS-layers to replace the first two Convolution-1D layers in the second Matrix Net, which in our experiments, show slightly better results than without HS-layers (See Table 2 [B0, G0] for the performance comparison). The structure of the Global Feature Extractor is shown in Fig. 3, we use 1 layer of HS-layer and 2 Convolution-1D layers with the output size of $[128, 512, 1024]$ to extract dense point features, and then apply maximum pooling to get the global feature. Finally, the global feature is concatenated with the input features for the outputs.

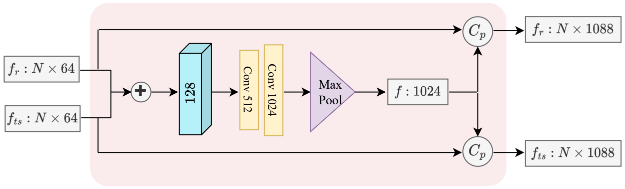


Figure 3. **Structure of the global extractor.**

6. Performance Comparion on CAMERA25

Table 3 compares the accuracy of our method with the state-of-the-arts. As discussed in Sec. 4, our performance stabilizes when using 4% of the full train set. Therefore, we present the results obtained with this training size. As shown in Table 3, we greatly enhanced the performance of SPD, resulting in a performance that outperformed state-of-the-art pose estimation methods. Specifically, we improved the performance of SPD [8] on IoU_{75} by 32.7%, 5°5cm by 25.2%, and 5°2cm by 23.8%. We also outperform the baseline CATRE on IoU_{75} by 3.1%, 5°5cm by 3.7%, and 5°2cm by 2.5%. Additionally, we show our results trained using 5k images (2%) of the train set, which already outperforms the state-of-the-art methods.

7. Per-category Performance

7.1. CAMERA25.

We present our per-category object pose refinement performance in Table 4. We use SPD [8] as the initial estimation method and report the performance after 4 refinement iterations. We show that our method largely improved the initial performance.

Table 2. **Ablation studies on REAL275.**
Higher score means better performance. Overall best results are in bold. Row's code in bold means the strategies taken in the final structure.

Row	Method	IoU_{50}	IoU_{75}	5°2cm	5°5cm	10°2cm	10°5cm	2cm	5°
A0	CATRE [6] (baseline)	77.0	43.6	45.8	54.4	61.4	73.1	75.1	58.0
B0	**Ours**: E0+Cross-Cloud Transformation	79.2 **2.2↑**	**51.8** **8.2↑**	**54.4** **8.6↑**	**60.3** **5.9↑**	**71.9** **10.5↑**	**79.4** **6.3↑**	**81.9** **6.8↑**	**64.3** **6.3↑**
C0	A0: PointNet → HS-Encoder	71.0	30.1	41.9	45.9	60.6	70.3	71.9	48.7
C1	A0: PointNet → 3DGCN-Encoder	-	28.4	36.0	43.4	-	-	68.0	47.7
D0	A0 + prior in ST branch	77.1	45.8	48.0	54.6	63.8	72.5	77.9	59.2
E0	D0: PointNet → HS-layer+LATs	79.4	51.0	52.4	58.6	69.4	77.7	80.4	62.4
E1	B0: No LAT on input points	76.1	39.3	46.6	53.0	65.4	74.8	78.0	58.2
E2	B0: No LATs on features	78.5	48.8	47.4	53.0	67.4	75.0	80.4	57.4
E3	B0: No LAT on the rotation feature	**79.8**	50.6	50.4	56.2	68.6	76.3	80.2	60.8
F0	E0+ Global Concatenation Fusion	77.7	48.4	47.8	54.5	67.1	75.2	80.1	59.4
G0	B0: No HS-layer in Matrix Net	77.8	50.2	54.1	60.1	70.5	78.0	81.2	63.6

Table 3. **Comparison with other methods on CAMERA25.**
Higher score means better performance. Overall best results are in bold. SPD* is the implementation results from CATRE, which is similar to the original SPD results.

Method	IoU_{75}	5°2cm	5°5cm	10°2cm	10°5cm
NOCS [9]	37.0	32.3	40.9	48.2	64.6
DualPoseNet [4]	71.7	64.7	70.7	77.2	84.7
CR-Net [10]	75.0	72.0	76.4	81.0	87.7
SGPA [1]	69.1	70.7	74.5	82.7	88.4
SAR-Net [3]	62.6	66.7	70.9	75.3	80.3
SSP-Pose [12]	-	64.7	75.5	-	87.4
RBP-Pose [11]	-	73.5	79.6	82.1	89.5
GPV-Pose [2]	-	72.1	79.1	-	89.0
HS-Pose [13]	-	73.3	80.5	80.4	89.4
SPD* [8]	46.9	54.1	58.8	73.9	82.1
SPD*+CATRE [6]	76.1	75.4	80.3	83.3	89.3
SPD*+**Ours** (2%)	77.5	75.4	81.1	83.4	90.0
SPD*+**Ours**	**79.2**	**77.9**	**84.0**	**83.8**	**90.5**

7.2. REAL275.

We present the per-category object pose refinement results in Table 5. We use SPD [8] as the initial estimation method and report the performance after 4 refinement iterations. We show that our method largely improved the initial performance.

8. Additional Qualitative Results

We show additional qualitative results of our method test on different REAL275 test scenes in Fig. 4 and Fig. 5. We highlight the performance differences with red arrows.

Table 4. **Per-category results of our method on CAMERA25 dataset.**

Method	Category	IoU$_{50}$	IoU$_{75}$	5°2cm	5°5cm	10°2cm	10°5cm	10°10cm	5°	2cm
SPD	bottle	88.9	64.5	63.8	82.8	69.2	92.4	97.3	86.8	69.8
SPD+**Ours**	bottle	89.4	73.8	73.8	94.2	74.2	95.1	99.4	98.4	74.2
SPD	bowl	95.9	80.6	83.4	83.7	95.8	96.3	96.3	83.7	99.2
SPD+**Ours**	bowl	96.0	94.7	97.9	98.2	99.5	99.8	99.8	98.2	99.6
SPD	camera	61.9	4.7	27.3	29.3	72.9	78.6	78.6	29.5	89.8
SPD+**Ours**	camera	81.6	67.7	83.1	87.2	90.8	95.2	95.2	87.2	93.9
SPD	can	90.2	87.2	98.1	98.2	99.4	99.6	99.6	98.2	99.6
SPD+**Ours**	can	90.3	89.8	99.9	100.0	99.9	100.0	100.0	100.0	99.9
SPD	laptop	93.3	17.7	35.0	41.9	61.0	80.5	84.5	43.7	65.5
SPD+**Ours**	laptop	95.3	81.3	74.0	85.5	77.4	91.8	95.8	89.1	77.9
SPD	mug	82.7	24.1	15.5	15.5	44.1	44.1	44.1	15.9	99.6
SPD+**Ours**	mug	89.8	67.7	39.0	39.0	61.0	61.0	61.0	39.4	99.9

Table 5. **Per-category results of our method on REAL275 dataset.**

Method	Category	IoU$_{50}$	IoU$_{75}$	5°2cm	5°5cm	10°2cm	10°5cm	10°10cm	5°	2cm
SPD	bottle	49.9	13.1	21.6	23.2	69.4	76.0	87.1	35.9	80.7
SPD+**Ours**	bottle	49.8	36.2	64.8	68.0	82.5	88.6	100.0	82.5	89.1
SPD	bowl	100.0	77.1	50.5	54.0	75.8	80.3	80.3	54.0	94.7
SPD+**Ours**	bowl	100.0	91.9	91.2	95.6	95.4	100.0	100.0	95.7	95.4
SPD	camera	43.4	3.4	0.0	0.0	0.2	0.2	0.2	0.0	34.8
SPD+**Ours**	camera	78.4	12.4	2.1	2.1	17.9	18.8	18.9	2.2	58.3
SPD	can	70.0	29.8	37.9	42.7	80.4	91.6	91.6	45.5	87.1
SPD+**Ours**	can	70.3	36.7	75.6	78.6	96.0	99.9	99.9	80.7	96.0
SPD	laptop	82.0	35.5	4.6	7.0	24.5	65.3	65.9	7.1	29.1
SPD+**Ours**	laptop	80.8	73.9	67.6	91.8	68.9	94.4	95.6	92.5	69.3
SPD	mug	66.5	8.7	0.3	0.3	10.3	10.4	10.4	0.3	85.2
SPD+**Ours**	mug	96.2	59.5	24.8	25.9	70.7	74.8	74.8	25.9	89.9

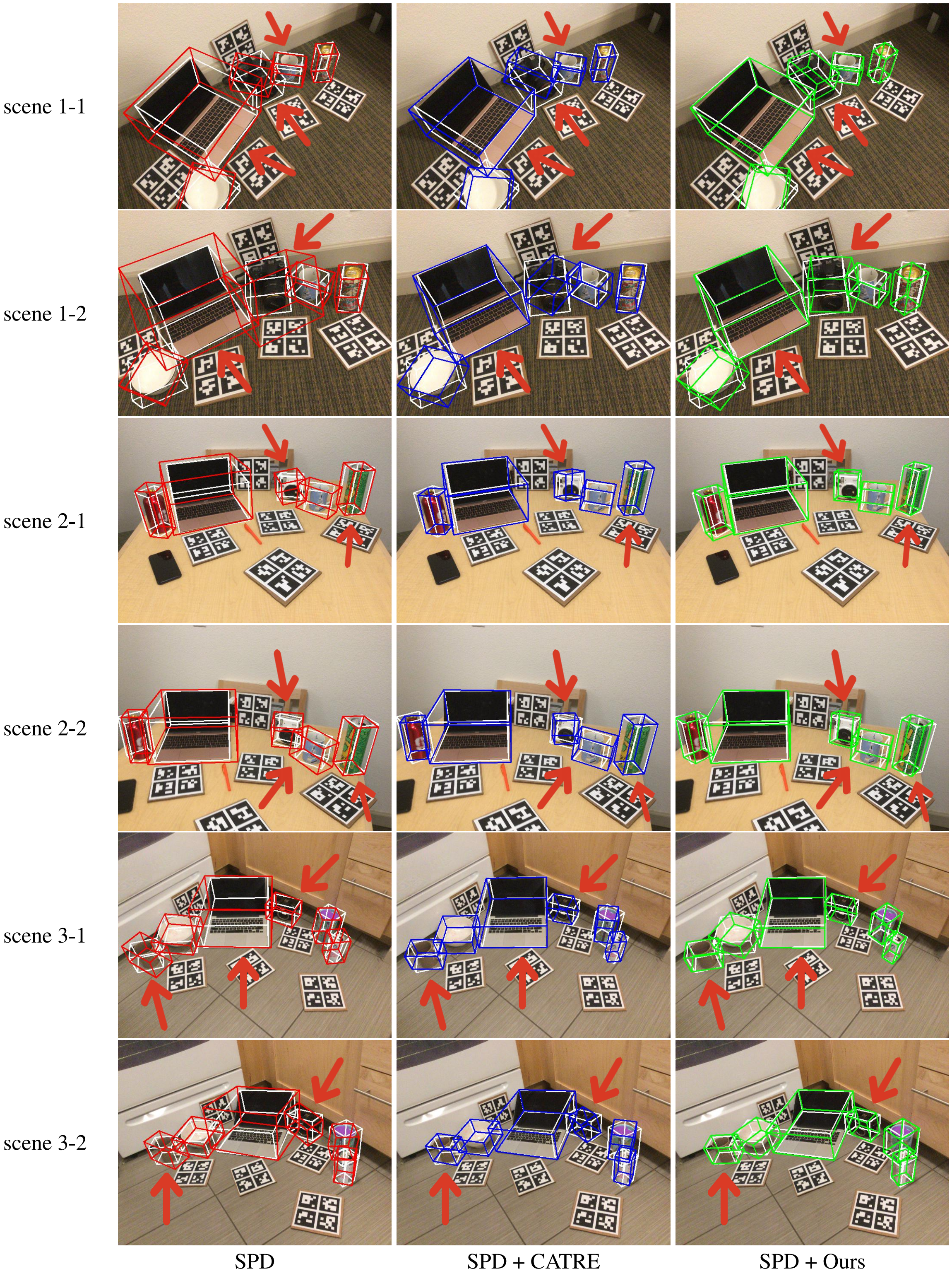


Figure 4. **More qualitative comparison between the proposed method (column #3) and the baseline method (column #2) use the SPD (column #1) as the initial estimation.** We choose two instances from each scene in REAL275 dataset. We show the ground truth with white lines. Note that the estimated rotations of symmetric objects (*e.g*. bowl, bottle, and can) are considered correct if the symmetry axis is aligned.

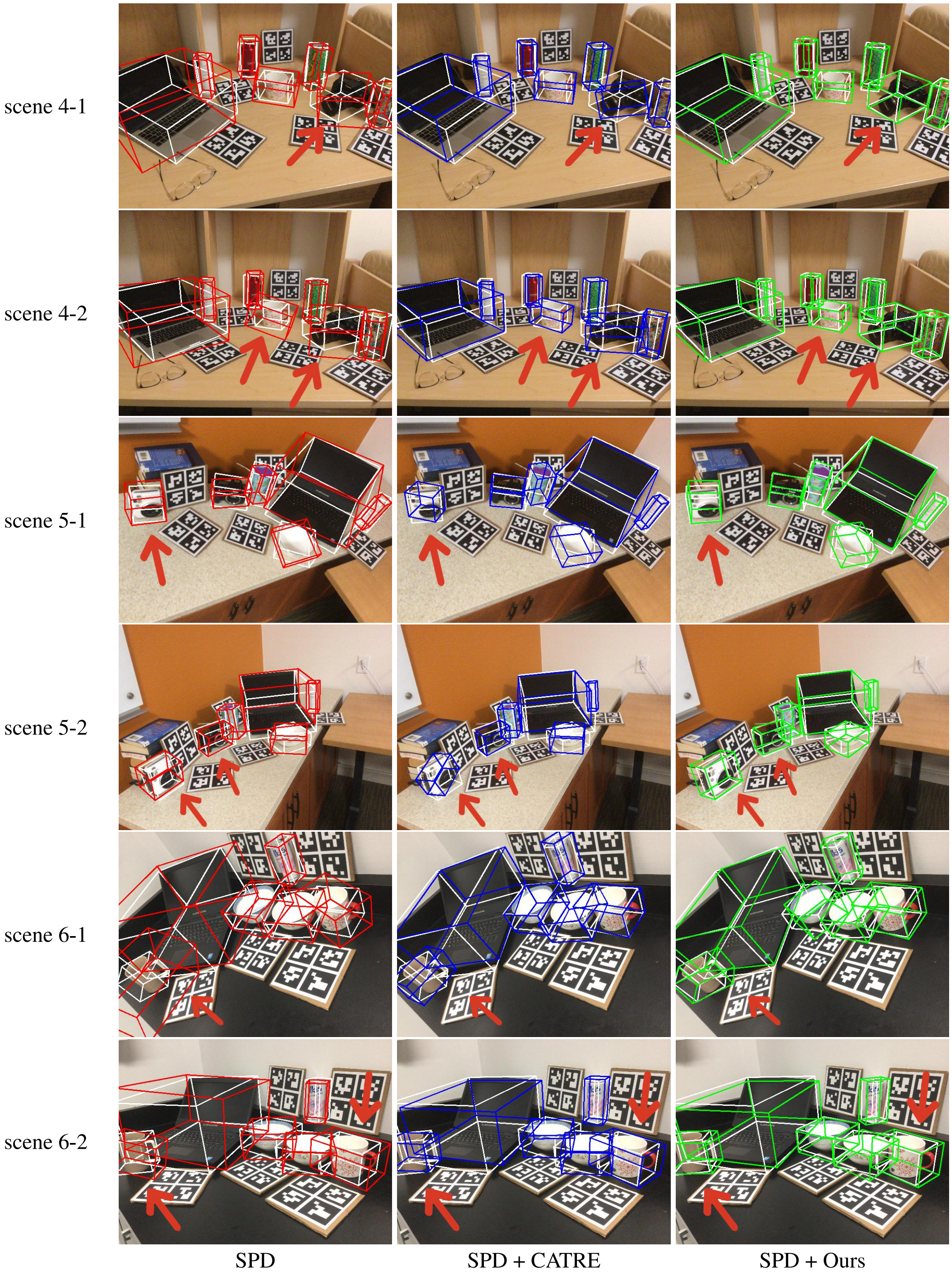


Figure 5. **More qualitative comparison between the proposed method (column #3) and the baseline method (column #2) use the SPD (column #1) as the initial estimation.** We choose two instances from each scene in REAL275 dataset. We show the ground truth with white lines. Note that the estimated rotations of symmetric objects (*e.g*. bowl, bottle, and can) are considered correct if the symmetry axis is aligned.

References

[1] Kai Chen and Qi Dou. SGPA: Structure-guided prior adaptation for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2773–2782, 2021. 3

[2] Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. GPV-Pose: Category-level object pose estimation via geometry-guided point-wise voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6781–6791, 2022. 1, 3

[3] Haitao Lin, Zichang Liu, Chilam Cheang, Yanwei Fu, Guodong Guo, and Xiangyang Xue. Sar-net: Shape alignment and recovery network for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6707–6717, 2022. 3

[4] Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. DualPoseNet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3560–3569, 2021. 3

[5] Jiehong Lin, Zewei Wei, Changxing Ding, and Kui Jia. Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks, 2022. 1

[6] Xingyu Liu, Gu Wang, Yi Li, and Xiangyang Ji. CATRE: Iterative point clouds alignment for category-level object pose refinement. In *European Conference on Computer Vision (ECCV)*, pages 499–516. Springer, 2022. 1, 2, 3

[7] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[8] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *European Conference on Computer Vision (ECCV)*, pages 530–546. Springer, 2020. 1, 2, 3

[9] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2642–2651, 2019. 3

[10] Jiaze Wang, Kai Chen, and Qi Dou. Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4807–4814. IEEE, 2021. 3

[11] Ruida Zhang, Yan Di, Zhiqiang Lou, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. RBP-Pose: Residual bounding box projection for category-level pose estimation, 2022. 1, 3

[12] Ruida Zhang, Yan Di, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. SSP-Pose: Symmetry-aware shape prior deformation for direct category-level object pose estimation, 2022. 3

[13] Linfang Zheng, Chen Wang, Yinghan Sun, Esha Dasgupta, Hua Chen, Aleš Leonardis, Wei Zhang, and Hyung Jin Chang. HS-Pose: Hybrid scope feature extraction for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17163–17173, 2023. 1, 3