Center-Based Decoupled Point Cloud Registration for 6D Object Pose Estimation

Haobo Jiang[1], Zheng Dang[2], Shuo Gu[1], Jin Xie*[1], Mathieu Salzmann[2], and Jian Yang*[1]

[1]PCA Lab, Nanjing University of Science and Technology, China
[2]CVLab, EPFL, Switzerland
{jiang.hao.bo, shuogu, csjxie, csjyang}@njust.edu.cn
{zheng.dang, mathieu.salzmann}@epfl.ch

Abstract

In this paper, we propose a novel center-based decoupled point cloud registration framework for robust 6D object pose estimation in real-world scenarios. Our method decouples the translation from the entire transformation by predicting the object center and estimating the rotation in a center-aware manner. This center offset-based translation estimation is correspondence-free, freeing us from the difficulty of constructing correspondences in challenging scenarios, thus improving robustness. To obtain reliable center predictions, we use a multi-view (bird's eye view and front view) object shape description of the source-point features, with both views jointly voting for the object center. Additionally, we propose an effective shape embedding module to augment the source features, largely completing the missing shape information due to partial scanning, thus facilitating the center prediction. With the center-aligned source and model point clouds, the rotation predictor utilizes feature similarity to establish putative correspondences for SVD-based rotation estimation. In particular, we introduce a center-aware hybrid feature descriptor with a normal correction technique to extract discriminative, part-aware features for high-quality correspondence construction. Our experiments show that our method outperforms the state-of-the-art methods by a large margin on real-world datasets such as TUD-L, LINEMOD, and Occluded-LINEMOD. Code is available at https://github.com/Jiang-HB/CenterReg.

*Corresponding authors
Haobo Jiang, Shuo Gu, Jin Xie, and Jian Yang are with PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, China.

1. Introduction

Accurate 6D object pose estimation (position and orientation in 3D space) is a crucial task in many real-world applications, such as robotics grasping [13, 59, 73], augmented reality [43, 44], and autonomous navigation [8, 22, 62]. While great progress has been made when exploiting RGB or RGB-D data as input [34, 54, 51, 59, 49, 69, 58], the advances in 3D sensors and deep point-cloud learning architectures have led to the development of increasingly accurate point cloud registration algorithms [61, 67, 32, 19, 14].

Nevertheless, the current state-of-the-art object-level 3D registration methods [56, 67, 19] prioritize achieving high performance on synthetic data, and yet still struggle with the challenges present in real-world data [25, 24, 6], such as full-range transformations, natural noise interference, and severe occlusions. A promising direction to alleviate this consists of decoupling the rotation and translation solutions, so as to reduce their interference. This was first investigated in [41, 42, 57] via the use of handcrafted rotation-invariant and translation-invariant feature descriptors. More recently, this idea was translated to the deep learning realm by aiming to learn representations that disentangle rotation and translation [10].

In this paper, we introduce a drastically different approach to decoupled registration for robust 6D object pose estimation in real-world scenarios. We advocate using the center offset between the source and model point clouds to decouple the translation from the entire transformation. In contrast to common correspondence-based translation estimation [61, 67, 14], our method is correspondence-free. It just requires regressing the independent center of the point cloud itself, freeing us from the difficulty of correspondence construction in challenging scenarios and thus significantly enhancing robustness. As the model point cloud can always be centered at the referential origin, our translation decoupling can be further simplified to only predicting the position of the center in the source point cloud. We therefore

decouple 3D registration into a center prediction (for translation estimation) and a center-aware rotation prediction.

Specifically, our method consists of a multi-view center predictor and a center-aware rotation predictor. The center predictor extracts a bird's eye view (BEV) and a front view description of the object shape from the source point cloud features, which jointly vote for the object center. Notably, we augment the source-point features with rich object-shape information from the model point cloud via a model-shape embedding module. This lets us largely complement the missing shape information of the partially-scanned source point cloud, thus facilitating the center prediction. With the center-aligned source and model point clouds, the rotation predictor utilizes feature similarity to establish putative correspondences for SVD-based rotation estimation [61]. In this context, we develop a center-aware hybrid feature descriptor to inhibit wrong correspondences caused by mismatched points with similar local structures. Specifically, this descriptor characterizes a part-aware representation of points, highlighting the intra-object location of each point (i.e., which part of the object the point belongs to), so as to distinguish locally-similar yet mismatched points. In addition, we propose a center-aware normal-orientation correction technique to account for normal consistency in the construction of the correspondences.

To summarize, our main contributions are as follows:

- We propose a novel center-based decoupled registration framework for robust real-world 6D object pose estimation, which first identifies the object center to decouple translation and rotation prediction.
- We develop a robust multi-view center predictor, using BEV and front-view source-feature projections to jointly vote for the object center. Notably, the source features are augmented with a proposed model-shape embedding module to complement the missing object-shape information due to partial scanning.
- We propose an effective center-aware rotation estimation method, designing a center-aware hybrid feature descriptor and a normal correction strategy to improve the robustness of extracted feature correspondences.

Our extensive experimental results on the real-world TUD-L [25], LINEMOD [24] and Occluded-LINEMOD [6] datasets evidence that our method outperforms the state of the art by a large margin.

2. Related Work

Traditional Registration Methods. Iterative Closest Point (ICP) [5] is a widely used and reliable fine registration algorithm. Its objective is to iteratively construct nearest-neighbor correspondences and perform least-squares optimization. However, the non-convexity of this overall objective function can lead to poor solutions depending on the initial pose. To address this sensitivity to pose initialization, Go-ICP [66] employs the branch-and-bound (BnB) scheme to globally search for the optimal transformation in a discretized 6D transformation parameter space. For improved registration reliability in the presence of partial observations, trimmed ICP [11] chooses minimal error subsets rather than optimizing all transformations. Other ICP variants [55, 18, 3, 23, 16] also show competitive performance in fine registration. By contrast, RANSAC-based coarse registration methods utilize an hypothesis verification approach to search for the optimal transformation. For instance, 4PCS [1] utilizes intersectional diagonal ratios to constrain the correspondence (four-point sets) sampling, and Super4PCS [45] further improves the computational efficiency of 4PCS, requiring only linear complexity. Other RANSAC variants such as [47, 46, 63, 27, 21] have led to impressive registration ability.

End-to-end Deep Registration Methods. With the progress of deep learning in 3D vision [52, 53], learning-based end-to-end registration approaches have emerged as a promising alternative to traditional methods. These method learn to directly regress the rotation and translation from an input point cloud pair. DCP [61], a pioneer in this area, uses deep closest points as pseudo-corresponding target points to establish putative correspondences for SVD-based transformation prediction. To handle the partial-to-partial registration case, PRNet [60] further performs keypoint detection and Gumbel softmax-based correspondence identification in an iterative manner. RPMNet [67] further incorporates a Sinkhorn layer and an annealing scheme for better outlier rejection. Furthermore, PointNetLK [2] and FMR [29] use the Lucas & Kanada (LK) [4] and inverse compositional (IC) [4] algorithms, respectively, to iteratively search for the optimal transformation via feature alignment. RGM [19] and RIENet [56] propose to leverage graph matching and geometric difference of the neighborhood, respectively, to improve the robustness to outliers. [32, 31] integrate the cross-entropy method into the deep model for robust rigid registration. Many other models, such as [60, 36, 12, 48, 39, 38, 72, 30] have further been developed, achieving impressive registration performance.

Transformation Decoupling. A particularly promising research direction in point cloud registration consists of decoupling the rotational and the translational parts of the transformation. This reduces interference between these two terms, and simplifies the solution space, leading to more robust 3D registration. Various decomposition strategies have been proposed. Straub et al. [57] use the surface normal distribution of a point cloud as translation-invariant feature and apply the BnB algorithm to search for the optimal rotation and translation. Liu et al. [41] design a rotation-invariant feature, allowing for the translation to be decoupled first using the BnB algorithm. Chen et al. [7]

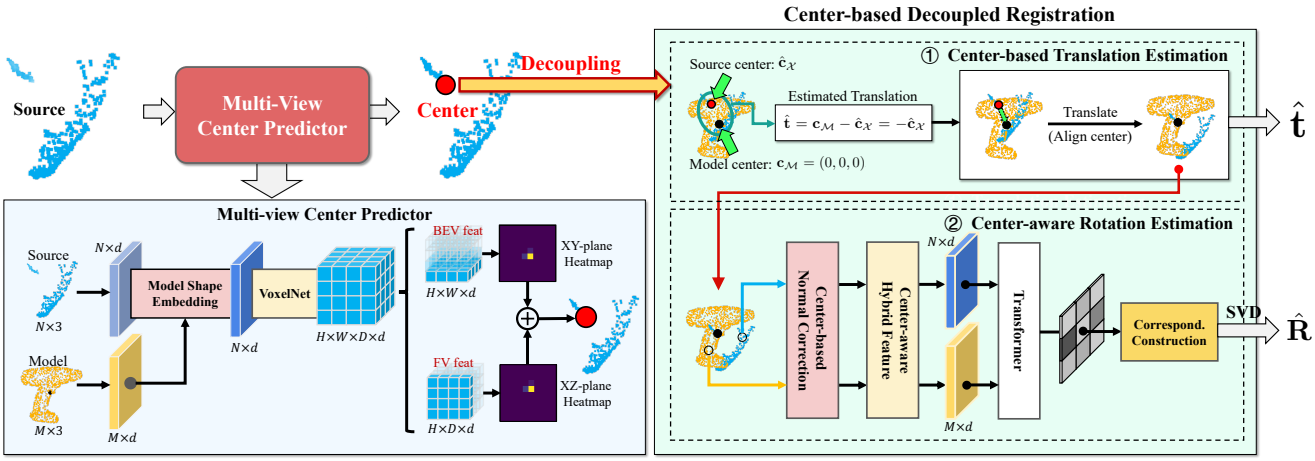

Figure 1. **Detailed pipeline of the proposed center-based decoupled registration.** Given the source and model point clouds, $\mathcal{X}$ and $\mathcal{M}$, our multi-view center predictor exploits a model shape embedding to augment the source point cloud features, and extracts bird's eye view (BEV) and front view (FV) source features that jointly vote for the object center $\hat{\mathbf{c}}_{\mathcal{X}}$. Then, the translation can be estimated using the center offset $\hat{\mathbf{t}} = \mathbf{c}_{\mathcal{M}} - \hat{\mathbf{c}}_{\mathcal{X}} = -\hat{\mathbf{c}}_{\mathcal{X}}$. After translation, the center-aligned point cloud pair is fed into our center-aware rotation predictor, relying on center-aware hybrid feature descriptors with normal correction to extract discriminative features.

further reduce the high non-linearity of the 3 rotational degrees of freedom (DoF) by predicting rotation as (2+1) and (1+2) DoF. [65] presents a framework based on graph theory to separate the estimation of scale, rotation, and translation. [37, 35] propose to leverage line vectors for transformation decomposition. [42] separates the rotation prediction by maximizing the correlations between two Extended Gaussian Images (EGI) [26] of the surface normals with the spherical Fourier Transform [17]. DetarNet [10] uses iterative Coherent Feature Drift modules and an attention-enhanced weighted SVD for transformation decoupling. In this work, we also propose a novel transformation decoupling framework by predicting the object center and estimating the rotation in a center-aware manner. Unlike previous methods, our translation decoupling is correspondence-free, freeing us from the difficulty of constructing correspondences in challenging scenarios, thus significantly improving our robustness on translation estimation.

3. Approach

3.1. Problem Formulation

Point cloud-based 6D object pose estimation can be defined as follows. Given a partially-scanned source point cloud of an object $\mathcal{X} = \{\mathbf{x}_i \in \mathcal{R}^3\}_{i=1}^{N}$ and a corresponding complete model point cloud $\mathcal{M} = \{\mathbf{m}_j \in \mathcal{R}^3\}_{j=1}^{M}$, the objective is to estimate the optimal rotation $\mathbf{R}^* \in \mathrm{SO}(3)$ and translation $\mathbf{t}^* \in \mathcal{R}^3$ such that the transformed source point cloud aligns with the model point cloud precisely. The source point cloud $\mathcal{X}$ is obtained by a masked depth map captured by a depth sensor with known camera intrinsic parameters. The model point cloud $\mathcal{M}$ is generated by uniform sampling from a mesh model.

3.2. Multi-view Center Prediction with Semantic Augmentation

We advocate leveraging the center offset between the source and model point clouds to decouple the translation from the entire transformation. As the model point cloud can always be centered at the referential origin, we can simplify the translation decoupling process to only predicting the object center in the source point cloud. Unlike common translation estimation methods [61, 67, 14] that heavily depend on constructing robust correspondences, predicting the object center can be treated as a regression task, freeing us from the challenging task of establishing correspondences between misaligned, noisy, and partial point clouds. To this end, we introduce a multi-view center predictor that exploits a model shape embedding for reliable center regression of partially-scanned, incomplete source point clouds.

Geometric Feature Extraction. To capture the local geometry of $\mathcal{X}$ and $\mathcal{M}$ well, we adopt the deep hybrid feature descriptor of [67], yielding features denoted as $\mathcal{F}_{\mathcal{X}} = \{\mathbf{f}_{x_i} \in \mathcal{R}^d\}_{i=1}^{N}$ and $\mathcal{F}_{\mathcal{M}} = \{\mathbf{f}_{m_j} \in \mathcal{R}^d\}_{j=1}^{M}$. In detail, for each point $\mathbf{x}_i \in \mathcal{X}$, its neighboring points within a ball of radius δ are first grouped in a set $\mathcal{N}(\mathbf{x}_i) = \{\mathbf{x}_k \mid \|\mathbf{x}_k - \mathbf{x}_i\| \leq \delta\}$. The local geometry of the neighborhood is described by the coordinate offsets $\{\Delta\mathbf{x}_{i,k} = \mathbf{x}_k - \mathbf{x}_i\}$ and PPF features $\{\mathrm{PPF}(\mathbf{x}_i, \mathbf{x}_k)\}$. Then, the deep features of each point $\mathbf{x}_i$ are obtained by fusing its spatial coordinates and its neighborhood's geometric features using a Multi-Layer Perceptron (MLP). This yields

$$\mathbf{f}_{x_i} = \mathrm{MLP}(\mathbf{x}_i, \{\Delta\mathbf{x}_{i,k}\}, \{\mathrm{PPF}\,(\mathbf{x}_i, \mathbf{x}_k)\}). \quad (1)$$

Model Shape Embedding for Source Augmentation. To address the challenge of accurately predicting the object center from the partially-scanned source point cloud $\mathcal{X}$,

which may lack important shape information due to its incompleteness, we introduce a Model Shape Embedding (MoSE) module to augment the source-point features. It is learned to retrieve and prioritize relevant shape cues from the complete model point cloud $\mathcal{M}$, which significantly compensates for the missing shape information due to partial scanning, thus facilitating the center prediction.

Specifically, to characterize the local similarity between $\mathcal{X}$ and $\mathcal{M}$, we first compute their similarity map $S \in \mathcal{R}^{N\times M}$ between features $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{M}$ using the vector inner product. That is, each entry in this map is computed as $S_{i,j} = \langle \mathbf{f}_{x_i}, \mathbf{f}_{m_j} \rangle$. Then, we normalize each row $S_{i,:} \in \mathcal{R}^M$ through a softmax function $\bar{S}_{i,:} = \text{softmax}(S_{i,:})$. As such, the elements in $\bar{S}_{i,:}$ are constrained to the $[0, 1]$ range and their sum is equal to 1. Such feature correlations are used to guide the retrieval of object cues, expressed as:

$$\mathbf{f}_{i,j} = \text{MLP}([\bar{S}_{i,j} \cdot \mathbf{f}_{m_j}; \mathbf{m}_j]), \tag{2}$$

where $\mathbf{f}_{i,j} \in \mathcal{R}^d$ represents the retrieved object information from the j-th model point for i-th source point, and $[\cdot;\cdot]$ indicates vector concatenation. In essence, we weight the effect of the model feature $\mathbf{f}_{m_j}$ by the feature similarity $\bar{S}_{i,j}$. We then summarize the retrieved object cues for point $\mathbf{x}_i$ via the max-pooling operator as $\bar{\mathbf{f}}_{x_i} = \text{MaxPool}_j \mathbf{f}_{i,j} \in \mathcal{R}^d$. Finally, we employ an MLP to fuse the retrieved object features, along with the original source features and coordinates, to produce the augmented source features, denoted by $\tilde{\mathcal{F}}_\mathcal{X} = \{\tilde{\mathbf{f}}_{x_i} = \text{MLP}([\mathbf{x}_i; \mathbf{f}_{x_i}; \bar{\mathbf{f}}_{x_i}]) \in \mathcal{R}^d\}_{i=1}^N$.

Multi-view Center Prediction. To estimate the object center, we extract dense bird's eye view (BEV) and front-view (FV) object shape descriptions from the augmented source features, and make them jointly vote for the center position. Specifically, inspired by [71, 68], we divide the augmented source features along with their associated point coordinates $\{[\mathbf{x}_i; \tilde{\mathbf{f}}_{x_i}] \in \mathcal{R}^{3+d}\}_{i=1}^N$ into an equally-spaced voxel grid of size $H = \frac{H'}{v}$, $W = \frac{W'}{v}$ and $D = \frac{D'}{v}$, where H', W', and D' represent the range of the source point cloud $\mathcal{X}$ along the X, Y, and Z axes, while v denotes the voxel size. We employ a mini-PointNet [52] extract voxel-wise shape features, followed by a series of 3D convolutional middle layers [71] that hierarchically aggregate the voxel features for global context encoding.

The resulting voxel features are then max-pooled along the Z and Y axes to generate the BEV and FV feature maps, respectively. Intuitively, by compressing the 3D voxel features into 2D features, we can significantly increase the proportion of non-empty voxels and thus obtain dense, informative features for more robust center prediction. The generated 2D feature maps are then mapped to XY-plane and XZ-plane keypoint heatmaps, $\hat{\mathbf{H}}_{\text{XY}} \in [0, 1]^{H\times W\times 1}$ and $\hat{\mathbf{H}}_{\text{XZ}} \in [0, 1]^{H\times D\times 1}$, along with offset maps $\hat{\mathbf{O}}_{\text{XY}} \in \mathcal{R}^{H\times W\times 2}$ and $\hat{\mathbf{O}}_{\text{XZ}} \in \mathcal{R}^{H\times D\times 2}$. As such, the integer center coordinates $\hat{\mathbf{c}}_{\text{XY}}$ and $\hat{\mathbf{c}}_{\text{XZ}}$ can be determined by selecting the peak (the coordinate with the highest response) in each heatmap. We then refine these integer center coordinates using the corresponding continuous coordinate offsets $\hat{\mathbf{o}}_{\text{XY}}$ and $\hat{\mathbf{o}}_{\text{XZ}}$, yielding two 2D center predictions, $\hat{\mathbf{c}}' = (u', v') = \hat{\mathbf{c}}_{\text{XY}} + \hat{\mathbf{o}}_{\text{XY}}$ and $\hat{\mathbf{c}}'' = (u'', v'') = \hat{\mathbf{c}}_{\text{XZ}} + \hat{\mathbf{o}}_{\text{XZ}}$. Finally, the 2D center predictions in the two planes jointly vote for the 3D object center in $\mathcal{X}$ as

$$\hat{\mathbf{c}}_\mathcal{X} = \Big(\frac{1}{2}(u' + u''), v', v''\Big). \tag{3}$$

3.3. Center-based Decoupled Registration

Center-based Translation Estimation. With the predicted object center $\hat{\mathbf{c}}_\mathcal{X}$ in the source point cloud as detailed in Sec. 3.2, we can directly decouple the translation from the rigid transformation using the proposed center offset-based decoupling strategy. As the model point cloud can always be centered at the referential origin, that is its center coordinate $\mathbf{c}_\mathcal{M} = (0, 0, 0)$, the estimated translation can be computed as

$$\hat{\mathbf{t}} = \mathbf{c}_\mathcal{M} - \hat{\mathbf{c}}_\mathcal{X} = -\hat{\mathbf{c}}_\mathcal{X}. \tag{4}$$

Center-aware Rotation Estimation. With the estimated translation $\hat{\mathbf{t}} = -\hat{\mathbf{c}}_\mathcal{X}$, we translate the source point cloud to align the centers of source and model point clouds, obtaining $\tilde{\mathcal{X}} = \{\tilde{\mathbf{x}}_i = \mathbf{x}_i + \hat{\mathbf{t}}\}_{i=1}^N$. We then propose a center-aware rotation predictor to estimate the rotation transformation between $\tilde{\mathcal{X}}$ and $\mathcal{M}$. Specifically, our rotation predictor utilizes the feature similarity to establish point-to-point correspondences between $\tilde{\mathcal{X}}$ and $\mathcal{M}$ for end-to-end SVD-based rotation estimation [61]. Notably, while center alignment can remove the translation-caused feature difference between corresponding points, constructing high-quality correspondences in challenging real-world scenarios for rotation estimation remains nontrivial. We address this by introducing an effective center-aware hybrid feature descriptor to improve the discriminative power of the features.

Our approach is inspired by the hybrid feature descriptor of [67]. However, this descriptor heavily relies on PPF features as local descriptors, PPF features have two main drawbacks in correspondence learning. **(i)** First, they solely capture local properties, which tend to make the resulting features ambiguous. For example, when non-corresponding source and model points have consistent local structures, their PPF features also are consistent, thereby generating wrong correspondences. **(ii)** Second, PPF features heavily rely on surface normals to describe the local geometry, which suffer from orientation ambiguities, thereby degrading the feature consistency of corresponding points.

To mitigate issue **(i)**, we propose a *Center-aware Feature* (CF) descriptor that characterizes a part-aware representation of points (see Fig 2). By encoding the relative position with respect to the center point, our CF descriptor highlights the intra-object location of each point (i.e., which part of the object the point belongs to), so as to distinguish

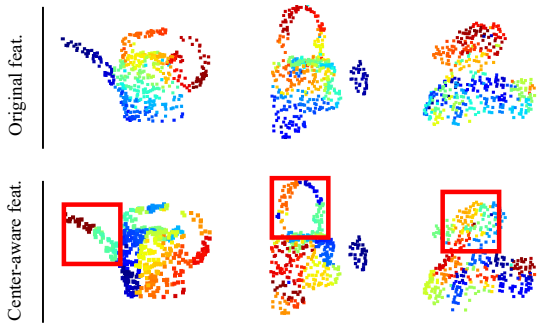

Figure 2. Visualized point features with/without our *center-aware feature* (CF) descriptor. Our CF descriptor extracts part-aware features, which significantly vary across the different object parts, allowing for better distinguishing the locally-similar yet non-corresponding points.

locally-similar yet non-corresponding points. Without loss of generality, we take the source point cloud $\mathcal{X}$ as an example. Our CF descriptor is computed as

$$\begin{aligned}\text{CF}(\mathbf{x}_i,\mathbf{x}_k,\mathbf{c}_\mathcal{X}) = (&\angle(\mathbf{n}_i,\mathbf{x}_i-\mathbf{c}_\mathcal{X}),\angle(\mathbf{n}_i,\mathbf{x}_k-\mathbf{c}_\mathcal{X})\\ &\angle(\mathbf{n}_k,\mathbf{x}_k-\mathbf{c}_\mathcal{X}),\|\mathbf{x}_i-\mathbf{c}_\mathcal{X}\|_2,\|\mathbf{x}_k-\mathbf{c}_\mathcal{X}\|_2),\end{aligned} \tag{5}$$

where $\mathbf{n}_i$ and $\mathbf{n}_k$ denote the normal vectors of point $\mathbf{x}_i$ and of its neighboring point $\mathbf{x}_k$ (see Sec. 3.2). The first three items in Eq. 5 describe angles between the normal and the coordinate offset vector relative to the center, and the last two items describe the Euclidean distance relative to the center. For corresponding points in the source and model point clouds, both the angle and distance features should be consistent, but this is typically not the case for non-corresponding points, thereby helping to identify the correct correspondences. Our final center-aware hybrid features are then represented by

$$\begin{aligned}\check{\mathbf{f}}_{x_i} = \text{MLP}(\mathbf{x}_i,&\{\Delta\mathbf{x}_{i,k}\},\{\text{PPF}\,(\mathbf{x}_i,\mathbf{x}_k)\},\\ &\{\text{CF}(\mathbf{x}_i,\mathbf{x}_k,\mathbf{c}_\mathcal{X})\}).\end{aligned} \tag{6}$$

We note that our CF descriptor enjoys the same transformation-invariance property as the PPF descriptor, which can be easily proved as follows. For the angle-wise features (the first three items in Eq. 5), we take the first one as an example as the proofs for the other two are analogous. Given an arbitrary rotation $\mathbf{R}\in\text{SO}(3)$ and translation $\mathbf{t}\in\mathcal{R}^3$, we have

$$\begin{aligned}&\angle(\mathbf{R}\mathbf{n}_i,\mathbf{R}\mathbf{x}_i+\mathbf{t}-(\mathbf{R}\mathbf{c}_\mathcal{X}+\mathbf{t}))\\ =&\angle(\mathbf{R}\mathbf{n}_i,\mathbf{R}(\mathbf{x}_i-\mathbf{c}_\mathcal{X}))=\angle(\mathbf{n}_i,\mathbf{x}_i-\mathbf{c}_\mathcal{X})\ .\end{aligned} \tag{7}$$

For the distance-wise features (the last two items in Eq. 5), we take the fourth item as an example. In this case, we have

$$\begin{aligned}&\|\mathbf{R}\mathbf{x}_i+\mathbf{t}-(\mathbf{R}\mathbf{c}_\mathcal{X}+\mathbf{t})\|_2\\ =&\|\mathbf{R}(\mathbf{x}_i-\mathbf{c}_\mathcal{X})\|_2=\|\mathbf{x}_i-\mathbf{c}_\mathcal{X}\|_2.\end{aligned} \tag{8}$$

Its confirms the transformation-invariance of our CF descriptor.

Let us now turn to issue **(ii)** discussed above. To mitigate it, we further leverage the centers of the source and model point clouds as viewpoints to correct their normal orientations. Without loss of generality, we take the source point cloud $\mathcal{X}$ as an example. We define a corrected normal vector of point $\mathbf{x}_i\in\mathcal{X}$ as $\tilde{\mathbf{n}}_i=\text{sgn}(\langle\mathbf{n}_i,\mathbf{x}-\mathbf{c}_\mathcal{X}\rangle)\mathbf{n}_i$, where $\mathbf{n}_i$ is its original normal; $\langle\cdot,\cdot\rangle$ denotes the inner product and $\text{sgn}(x)=|x|/x\ (x\neq 0)$ is the sign function. Lemma 1 below shows the consistency in normal orientation of corresponding points. Please refer to Appendix A for the proof.

Lemma 1 *Let $\mathbf{x}_i$ and $\mathbf{m}_j$ be corresponding points in $\mathcal{X}$ and $\mathcal{M}$, and $\mathbf{n}_i$ and $\mathbf{n}_j$ be their original normal vectors. Because of the normal orientation ambiguity, $\mathbf{n}_j$ is equal to either $\mathbf{R}^*\mathbf{n}_i$ or $-\mathbf{R}^*\mathbf{n}_i$, with $\mathbf{R}^*$ the true rotation between $\mathcal{X}$ and $\mathcal{M}$. The corrected normals, $\tilde{\mathbf{n}}_i=\text{sgn}(\langle\mathbf{n}_i,\mathbf{x}-\mathbf{c}_\mathcal{X}\rangle)\mathbf{n}_i$ and $\tilde{\mathbf{n}}_j=\text{sgn}(\langle\mathbf{n}_j,\mathbf{n}_j-\mathbf{c}_\mathcal{M}\rangle)\mathbf{n}_j$, have a consistent orientation, that is, $\tilde{\mathbf{n}}_\mathbf{m}=\mathbf{R}^*\tilde{\mathbf{n}}_\mathbf{x}$.*

We then use this strategy to correct all point normals.

After learning the point-wise features of the center-aligned source and model point clouds, $\tilde{\mathcal{X}}$ and $\mathcal{M}$, we use feature similarity to obtain the corresponding point $\hat{\mathbf{m}}_{(i)}$ in $\mathcal{M}$ of each $\tilde{\mathbf{x}}_i$, as in [61], and then use SVD [61] to solve for the rotation matrix $\hat{\mathbf{R}}$. In detail, we first calculate the cross-covariance matrix $\mathbf{H}=\sum_{i=1}^N(\tilde{\mathbf{x}}_i-\bar{\mathbf{x}})(\hat{\mathbf{m}}_i-\bar{\mathbf{m}})^\top$, where $\bar{\mathbf{x}}=\frac{1}{N}\sum_{i=1}^N\tilde{\mathbf{x}}_i$ and $\bar{\mathbf{m}}=\frac{1}{N}\sum_{i=1}^N\hat{\mathbf{m}}_{(i)}$ indicate the centroids of the source points and their corresponding model points. The rotation matrix $\hat{\mathbf{R}}$ can then be estimated via SVD of $\mathbf{H}$ as

$$\hat{\mathbf{R}}=\mathbf{U}diag(1,1,-1)\mathbf{V}^T,\ \mathbf{H}=\mathbf{U}\mathbf{S}\mathbf{V}^\top. \tag{9}$$

3.4. Loss Function

Center Supervision. Following [20], we generate the ground-truth XY-plane and XZ-plane keypoint heatmaps $\mathbf{H}_{\text{XY}}$ and $\mathbf{H}_{\text{XZ}}$ as supervision. The center coordinate is assigned a value of 1, while coordinates within the spatial range of the point cloud are set to $\frac{1}{d+1}$ (where d represents the Euclidean distance between the coordinate and the center). All other coordinates are set to 0. To balance the positive and negative samples more effectively, we utilize the focal loss [40]

$$\begin{aligned}\mathcal{L}_{\mathbf{XY}}=-\sum&\mathbb{I}\left[\mathbf{H}_{\text{XY}}^{(i,j)}=1\right]\left(1-\hat{\mathbf{H}}_{\text{XY}}^{(i,j)}\right)^\alpha\log\left(\hat{\mathbf{H}}_{\text{XY}}^{(i,j)}\right)+\\ &\mathbb{I}\left[\mathbf{H}_{\text{XY}}^{(i,j)}\neq 1\right]\left(1-\mathbf{H}_{\text{XY}}^{(i,j)}\right)^\beta\left(\hat{\mathbf{H}}_{\text{XY}}^{(i,j)}\right)^\alpha\log(1-\hat{\mathbf{H}}_{\text{XY}}^{(i,j)}),\end{aligned} \tag{10}$$

where the hyper-parameters α and β control the loss weights for positive and negative samples. The loss $\mathcal{L}_{\mathbf{XZ}}$ on the XZ-plane heatmap is analogous to $\mathcal{L}_{\mathbf{XY}}$.

Offset Supervision. Furthermore, taking the XY-plane as an example, we supervised offset regression with the loss

$$\mathcal{L}_{\mathbf{OXY}}=\sum_{\Delta x=-r}^{r}\sum_{\Delta y=-r}^{r}\left|\hat{\mathbf{O}}_{\text{XY}}^{\tilde{c}+(\Delta x,\Delta y)}-\mathbf{O}_{\text{XY}}^{\tilde{c}+(\Delta x,\Delta y)}\right|, \tag{11}$$

where $\tilde{c}$ indicates the discrete center coordinate and $\mathbf{O}_{\mathrm{XY}}^{\tilde{c}+(\Delta x,\Delta y)}$ denotes the ground-truth coordinate offset from the discrete coordinate $\tilde{c}+(\Delta x,\Delta y)$ to the continuous center coordinate. The offset regression $\mathcal{L}_{\mathbf{OXZ}}$ on the XZ plane is analogous. As such, the loss function for optimizing the center predictor is formulated as $\mathcal{L}_{\mathbf{C}} = \mathcal{L}_{\mathbf{XY}} + \mathcal{L}_{\mathbf{XZ}} + \mathcal{L}_{\mathbf{OXY}} + \mathcal{L}_{\mathbf{OXZ}}$.

Rotation Supervision. To supervise the rotation predictor, we first translate the source point cloud $\mathcal{X}$ using the translation vector $(\mathbf{R}^*)^{-1}\mathbf{t}^*$ (since $\mathbf{R}^*\mathbf{x}_i + \mathbf{t}^* = \mathbf{R}^*(\mathbf{x}_i + (\mathbf{R}^*)^{-1}\mathbf{t}^*)$) to align the center coordinates of the source and model point clouds. Then, we optimize the network parameters of the rotation predictor by minimizing the L_1 distance between the rotated source points using the predicted rotation matrix $\hat{\mathbf{R}}$ and the ground-truth one $\mathbf{R}^*$, i.e.,

$$\mathcal{L}_{\mathbf{R}} = \frac{1}{N}\sum_{i}^{N}\left\|\mathbf{R}^*(\mathbf{x}_i + (\mathbf{R}^*)^{-1}\mathbf{t}^*) - \hat{\mathbf{R}}(\mathbf{x}_i + (\mathbf{R}^*)^{-1}\mathbf{t}^*)\right\|_1. \tag{12}$$

4. Experiments

4.1. Experimental Settings

Implementation Details. We set the number of points in the source and model point clouds to $N = 512$ and $M = 1024$, respectively. The feature dimension d, voxel size v, and radius δ of the neighborhood ball are set to 32, 0.15, and 0.3. The hyper-parameters α and β in loss function 10 are set to 2 and 4. We optimize the overall loss function for 20 epochs with a batch size of 32, and use ADAM [33] as our optimizer with an initial learning rate of 0.001 decayed by a factor 0.2 every 6 epochs. We employ PyTorch[50] to implement our approach and conduct all experiments on a server with an Intel i5 2.2 GHz CPU and one TITAN RTX GPU with 24 GB of memory. We refer to our **Center**-based **Reg**istration framework as **CenterReg**.

Evaluation Metrics. Following [14], we evaluate performance using the rotation and translation errors between the predicted rotation and translation transformations $\hat{\mathbf{R}}$ and $\hat{\mathbf{t}}$ and the ground-truth ones $\mathbf{R}^*$ and $\mathbf{t}^*$ as defined below:

$$\mathrm{RE}(\hat{\mathbf{R}}) = \arccos\frac{\mathrm{Tr}\left(\hat{\mathbf{R}}^\top\mathbf{R}^*\right)-1}{2}, \mathrm{TE}(\hat{\mathbf{t}}) = \left\|\hat{\mathbf{t}}-\mathbf{t}^*\right\|_2^2. \tag{13}$$

As in [15, 14], we summarize these errors via mean average precision (mAP) under varying accuracy thresholds.

4.2. Comparison with Existing Methods

Evaluation on TUD-L. We first evaluate our method on the TUD-L dataset [25], which is a real-world dataset containing three moving household objects that are placed under eight different lighting settings (including ambient and directional light). We follow the data split given in [25] for model training and testing. We compare our approach with ten state-of-the-art (SOTA) methods, including representative traditional methods, such as ICP [5], FGR [70], TEASER++[65], and Super4PCS[45], as well as SOTA deep learning-based methods, such as DCP [61], IDAM [36], FMR [28], RPMNet [67], MN-IDAM [14] and MN-DCP [14]. Note that we do not include the deep decoupling model DetarNet [10] in our comparisons because

Models	Rotation mAP 5°	10°	20°	Translation mAP 1*cm*	2*cm*	5*cm*
ICP [5]	0.02	0.02	0.02	0.01	0.14	0.57
FGR [70]	0.00	0.01	0.01	0.04	0.25	0.63
TEASER++ [64]	0.13	0.17	0.19	0.03	0.22	0.56
Super4PCS [45]	0.30	0.50	0.56	0.05	0.40	0.92
DCP [61]	0.00	0.01	0.02	0.02	0.07	0.55
IDAM [36]	0.03	0.05	0.10	0.02	0.08	0.49
FMR [28]	0.02	0.09	0.13	0.02	0.06	0.19
RPMNet [67]	0.71	0.93	0.99	0.87	0.96	0.98
MN-IDAM [14]	0.36	0.46	0.53	0.23	0.47	0.57
MN-DCP [14]	0.70	0.81	0.87	0.71	0.86	0.97
CenterReg	**0.81**	**0.97**	**0.99**	**0.97**	**0.99**	**0.99**
CenterReg+ICP	0.89	0.98	0.99	0.95	0.97	0.99

Table 1. Performance comparisons with SOTA methods on the **TUD-L** benchmark dataset [25].

Models	Rotation mAP 5°	10°	20°	Translation mAP 1*cm*	2*cm*	5*cm*
ICP [5]	0.00	0.01	0.01	0.04	0.27	0.82
FGR [70]	0.00	0.00	0.00	0.05	0.31	0.89
TEASER++ [64]	0.01	0.03	0.05	0.03	0.21	0.73
Super4PCS [45]	0.02	0.09	0.15	0.04	0.31	0.89
DCP [61]	0.00	0.00	0.01	0.05	0.24	0.83
IDAM [36]	0.00	0.01	0.05	0.03	0.16	0.71
FMR [28]	0.00	0.01	0.04	0.07	0.17	0.42
RPMNet [67]	0.04	0.24	0.56	0.26	0.51	0.82
MN-IDAM [14]	0.01	0.07	0.15	0.13	0.38	0.87
MN-DCP [14]	0.10	0.27	0.49	0.26	0.60	0.95
CenterReg	**0.20**	**0.47**	**0.71**	**0.83**	**0.95**	**0.99**
CenterReg+ICP	0.66	0.79	0.85	0.92	0.96	0.99

Table 2. Performance comparisons with SOTA methods on the **LINEMOD** benchmark dataset [24].

Models	Rotation mAP 5°	10°	20°	Translation mAP 1*cm*	2*cm*	5*cm*
ICP [5]	0.01	0.01	0.01	0.07	0.36	0.85
FGR [70]	0.00	0.00	0.00	0.08	0.43	0.85
TEASER++ [64]	0.01	0.02	0.05	0.04	0.26	0.77
Super4PCS [45]	0.01	0.03	0.06	0.06	0.31	0.83
DCP [61]	0.00	0.00	0.01	0.03	0.30	0.83
IDAM [36]	0.00	0.02	0.06	0.07	0.26	0.76
FMR [28]	0.00	0.00	0.02	0.09	0.18	0.43
RPMNet [67]	0.04	0.24	**0.56**	0.26	0.51	0.82
MN-IDAM [14]	0.02	0.08	0.18	0.15	0.44	0.84
MN-DCP [14]	0.07	0.19	0.36	0.24	0.57	**0.88**
CenterReg	**0.11**	**0.29**	0.52	**0.57**	**0.76**	0.87
CenterReg+ICP	0.41	0.59	0.68	0.66	0.76	0.85

Table 3. Performance comparisons with SOTA methods on the **Occluded-LINEMOD** benchmark dataset [6].

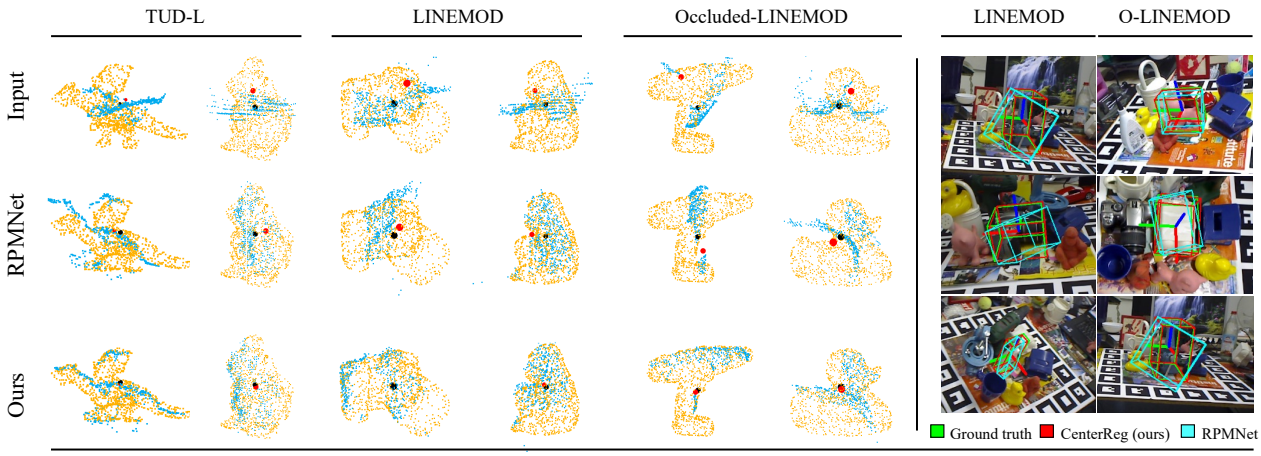

Figure 3. Qualitative comparison on the TUD-L [25], LINEMOD [24] and Occluded-LINEMOD [6] benchmark datasets.

it is designed for large-scale scene-level registration and is not directly applicable to object-level registration. We also tested other algorithms, such as RIENet [56], RGM [19], and SC2_PCR [9]. However, even with careful tuning, we are unable to achieve their reasonable scores. Therefore, we have included their scores in Appendix B. As shown in Table 1, traditional methods appear to be unable to produce reasonable results due to the real-world challenges (particularly for ICP and FGR), while deep models such as MNDCP and RPMNet tend to stick to the limited registration precision. Our method achieves the best performance on all translation and rotation criteria by a considerable margin, particularly impressive for the mAP@5° (10%↑) and mAP@1cm (9%↑). Such impressive performance is mainly due to our effective center-based decoupling mechanism for reliable translation estimation and the discriminative center-assisted feature extraction for high-quality correspondence construction in rotation estimation. We also observe that with ICP-based pose refinement, the rotation accuracy can be further improved while the translation precision tends to decrease. We conjecture this to be due to the good precision of our translation regression, with ICP, being less robust to outliers, having a negative impact. Qualitative results are provided in Fig. 3.

Evaluation on LINEMOD. We conduct further experiments on the LINEMOD dataset [24], a widely-used real-world dataset for 6D object pose estimation, which consists of 15 texture-less household objects in cluttered scenes. Following the setup in [14], we utilize the PBR dataset provided by the BOP Benchmark for model training and the testing split of the BOP 2019 challenge for performance evaluation. We compare our approach to the same methods as in the TUD-L case. The comparison results are provided in Table 2. Our approach consistently outperforms both traditional and state-of-the-art deep registration methods on all criteria, even in the face of the significant challenges posed by the LINEMOD dataset. Our method exhibits particularly impressive performance gaps, with precision advantages of up to 20% and 56% in mAP@10° and mAP@1cm. Such significant performance gain strongly supports the excellent robustness to real-world challenges of our center-based translation decoupling strategy. Moreover, Table 2 also shows that with just the low-cost ICP refinement, our registration precision can be further boosted, particularly in terms of rotation mAP. We also visualize 6D pose estimation results in Fig. 3 (right part).

Evaluation on Occluded-LINEMOD. We finally perform comparisons on the Occluded-LINEMOD dataset [6], which is a subset of the LINEMOD dataset containing 8 texture-less objects with varying levels of occlusion. As explicit training data for Occluded-LINEMOD is lacking, we directly use the model trained on LINEMOD for evaluation on this dataset. We include the same methods as before in our comparisons. The results are presented in Table 3. Due to severe occlusion, the registration precision of all methods tends to degrade compared to their performance on LINEMOD. Nonetheless, our method still achieves the best scores on most criteria, except for a slightly weaker performance compared to RPMNet on mAP@20° and MNDCP on mAP@5cm. Moreover, our method's translation prediction scores are particularly impressive, which can be attributed to the shape embedding module used in the center predictor that can retrieve valuable object cues from the object model. This allows the missing shape information in the source point cloud, resulting from the severe occlusions, to be largely completed for better center localization.

4.3. Ablation Studies and Analysis

Multi-view Center Predictor. We first test the performance contribution of our multi-view center predictor.

(1) Center-based translation decoupling. We take the rotation predictor as our *Baseline*, which additionally predicts the translation transformation using SVD. As demonstrated in the second row of Table 4, this *Baseline* model without our decoupling mechanism struggles with real-world chal-

Models	TUD-L 5°	TUD-L 10°	TUD-L 1*cm*	TUD-L 2*cm*	LINEMOD 5°	LINEMOD 10°	LINEMOD 1*cm*	LINEMOD 2*cm*	Occluded-LINEMOD 5°	Occluded-LINEMOD 10°	Occluded-LINEMOD 1*cm*	Occluded-LINEMOD 2*cm*	Sec.
Baseline	0.20	0.73	0.36	0.68	0.02	0.11	0.08	0.23	0.01	0.08	0.08	0.24	0.029
Baseline+Center	0.72	0.88	**0.97**	**0.99**	0.05	0.23	**0.83**	**0.95**	0.04	0.13	**0.57**	**0.76**	0.048
Baseline+Center+NC	0.77	0.94	**0.97**	**0.99**	0.11	0.34	**0.83**	**0.95**	0.07	0.21	**0.57**	**0.76**	0.050
Baseline+Center+NC+CF*	**0.81**	**0.97**	**0.97**	**0.99**	**0.20**	**0.47**	**0.83**	**0.95**	**0.11**	**0.29**	**0.57**	**0.76**	0.051
CenterReg (*w/o* MoSE)	**0.81**	**0.97**	**0.97**	**0.99**	0.19	0.45	0.76	0.90	0.11	0.27	0.48	0.63	0.045
CenterReg (*w/* sparsemax)	**0.81**	**0.97**	**0.97**	**0.99**	0.19	**0.47**	**0.85**	**0.95**	**0.12**	**0.30**	0.56	0.75	0.055
CenterReg (*w/* softmax)*	**0.81**	**0.97**	**0.97**	**0.99**	**0.20**	**0.47**	0.83	**0.95**	0.11	0.29	**0.57**	**0.76**	0.051

Table 4. Ablation studies on the TUD-L [25], LINEMOD [24] and Occluded-LINEMOD [6] benchmark datasets.

lenges and fails to give reasonable registration results, especially on the challenging Occluded-LINEMOD dataset. Instead, equipped with our center-based translation decoupling, *Baseline+Center* achieves a huge precision improvement in terms of translation mAP, with an increase of 74% in mAP@1*cm* on the LINEMOD dataset and similar improvements on the other benchmark datasets. These results demonstrate the impressive robustness and effectiveness of our center-based translation decoupling mechanism in complex real-world scenarios. Moreover, after decoupling the translation estimation, the rotation prediction also obtains some level of precision gain. This arises from the fact that that, after our translation decoupling, the feature differences, caused by the translation, between the center-aligned source and model point clouds have effectively been reduced, thus improving feature consistency and yielding more reliable correspondences.

(2) Model shape embedding. Furthermore, we test the effectiveness of our model shape embedding module (*MoSE*). As shown in the last block of Table 4, CenterReg without *MoSE* suffers from a significant precision degradation, particularly on Occluded-LINEMOD, where severe occlusions often cause missing object information in the source point cloud. By contrast, CenterReg with *MoSE* presents more reliable center prediction, such as 9% and 13% improvements in mAP@1*cm* and mAP@2*cm*, respectively. In addition, we also try to replace the softmax operator in *MoSE* with the sparsemax operator. Different from softmax which assigns small weights to the irrelevant shape cues, sparsemax enables to adaptively reduce such weights to zero. While reasonable, compared to CenterReg with *softmax*, CenterReg with *sparsemax* tends to bring limited precision gain (even some degradation) and also needs additional time for computing its sparse distribution. Thus, in our implementation, we just use the simple softmax function to guide the shape information retrieval in the *MoSE* module. Finally, in Fig. 4, we highlight that our center predictor presents robust center prediction under different subsampling ratios in addition to some extremely sparse settings.

Center-aware Rotation Predictor. We next justify the effectiveness of the proposed rotation predictor.

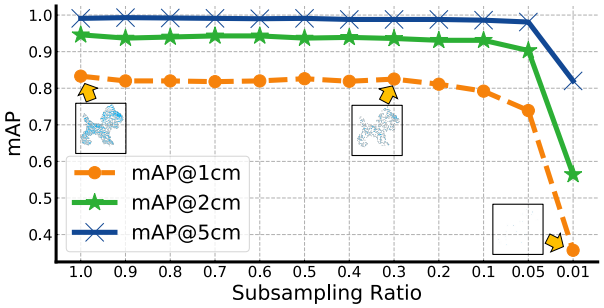

Figure 4. Translation mAPs under different subsampling ratios on LINEMOD benchmark dataset [24].

(1) Normal correction. We first test the performance gain brought by the proposed normal correction (*NC*). The third row in Table 4 shows that, owing to the orientation-refined normal using *NC*, the prediction precision of the rotation can be consistently improved by significant margins on all tested datasets, e.g., 6%↑ and 11%↑ in mAP@5° and mAP@10° on the LINEMOD dataset. This comes from the fact that the orientation-refined normal vectors obtained by *NC* promote more consistent local descriptions of the corresponding points, which help establish reliable correspondences for more precise rotation estimation.

(2) Center-aware feature descriptor. Then, we further test the precision contribution of the proposed center-aware feature descriptor (*CF*, Eq. 5). The fourth row in Table 4 also shows consistent performance improvements. Our *CF* descriptor enhances the point features with part-aware representations, which allow our method to effectively distinguish the locally-similar yet non-corresponding points, and thus establish more robust correspondences. We have included a discussion on the limitations of our study and the future work in Appendix C.

5. Conclusion

In this paper, we have proposed a novel and effective object center-based deep decoupling registration framework for robust point clouds based object pose estimation in real-world scenarios. As the object model can always be centered at the referential origin, we have converted translation estimation to the problem of object-center localization in

the source point cloud so as to directly decouple the translation from the transformation. We have then developed a center-aware rotation predictor to estimate the rotation from the center-aligned source and model point clouds. To construct high-quality correspondences for reliable rotation estimation, we have introduced center-aware feature descriptors and a center-based normal correction technique. Our extensive experiments on challenging real-world datasets have verified the outstanding performance of our method.

6. Acknowledgments

This work was supported by the National Science Fund of China (Grant Nos. U1713208, U62276144, 62106106) and the Swiss Innovation Agency (Innosuisse).

References

[1] Dror Aiger, Niloy Jyoti Mitra, and Daniel Cohen-Or. 4-points congruent sets for robust pairwise surface registration. *ACM SIGGRAPH 2008 papers*, 2008. 2

[2] Yasuhiro Aoki, Hunter Goforth, Rangaprasad Arun Srivatsan, and Simon Lucey. PointnetLK: Robust & efficient point cloud registration using pointnet. In *CVPR (2019)*. 2

[3] Kwang-Ho Bae and Derek D Lichti. A method for automated registration of unorganised point clouds. *Journal of Photogrammetry and Remote Sensing (2008)*, 63(1):36–54. 2

[4] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *International journal of computer vision*, 2004. 2

[5] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, 1992. 2, 6

[6] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *ECCV*, 2014. 1, 2, 6, 7, 8

[7] Wen Chen, Haoang Li, Qiang Nie, and Yun-Hui Liu. Deterministic point cloud registration via novel transformation decomposition. In *CVPR*, 2022. 2

[8] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *CVPR*, 2017. 1

[9] Zhi Chen, Kun Sun, Fan Yang, and Wenbing Tao. Sc2-pcr: A second order spatial compatibility for efficient and robust point cloud registration. In *CVPR*, 2022. 7

[10] Zhi Chen, Fan Yang, and Wenbing Tao. Detarnet: Decoupling translation and rotation by siamese network for point cloud registration. In *AAAI*, 2022. 1, 3, 6

[11] Dmitry Chetverikov, Dmitry Svirko, Dmitry Stepanov, and Pavel Krsek. The trimmed iterative closest point algorithm. In *Object recognition supported by user interaction for service robots (2002)*, volume 3, pages 545–548. 2

[12] Christopher Choy, Wei Dong, and Vladlen Koltun. Deep global registration. In *CVPR (2020)*, pages 2514–2523. 2

[13] Alvaro Collet, Manuel Martinez, and Siddhartha S Srinivasa. The moped framework: Object recognition and pose estimation for manipulation. *The international journal of robotics research*, 2011. 1

[14] Zheng Dang, Lizhou Wang, Yu Guo, and Mathieu Salzmann. Learning-based point cloud registration for 6d object pose estimation in the real world. In *ECCV*, 2022. 1, 3, 6, 7

[15] Zheng Dang, Kwang Moo Yi, Yinlin Hu, Fei Wang, Pascal Fua, and Mathieu Salzmann. Eigendecomposition-free training of deep networks with zero eigenvalue-based losses. In *ECCV*, 2018. 6

[16] Haowen Deng, Tolga Birdal, and Slobodan Ilic. PPFNet: Global context aware local features for robust 3D point matching. In *CVPR (2018)*. 2

[17] James R Driscoll and Dennis M Healy. Computing fourier transforms and convolutions on the 2-sphere. *Advances in applied mathematics*, 1994. 3

[18] Andrew W Fitzgibbon. Robust registration of 2D and 3D point sets. *Image and vision computing (2003)*, 21(13-14):1145–1153. 2

[19] Kexue Fu, Shaolei Liu, Xiaoyuan Luo, and Manning Wang. Robust point cloud registration framework based on deep graph matching. In *CVPR*, 2021. 1, 2, 7

[20] Runzhou Ge, Zhuangzhuang Ding, Yihan Hu, Yu Wang, Sijia Chen, Li Huang, and Yuan Li. Afdet: Anchor free one stage 3d object detection. *arXiv preprint arXiv:2006.12671*, 2020. 5

[21] Xuming Ge. Automatic markerless registration of point clouds with semantic-keypoint-based 4-points congruent sets. *ISPRS Journal of Photogrammetry and Remote Sensing (2017)*, 130:344–357. 2

[22] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1

[23] Adrien Gressin, Clément Mallet, Jérôme Demantké, and Nicolas David. Towards 3D lidar point cloud registration improvement using optimal neighborhood knowledge. *ISPRS journal of photogrammetry and remote sensing (2013)*, 79:240–251. 2

[24] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *ACCV*, 2013. 1, 2, 6, 7, 8

[25] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *ECCV*, 2018. 1, 2, 6, 7, 8

[26] Berthold Klaus Paul Horn. Extended gaussian images. *Proceedings of the IEEE*, 1984. 3

[27] Jida Huang, Tsz-Ho Kwok, and Chi Zhou. V4PCS: Volumetric 4PCS algorithm for global registration. *Journal of Mechanical Design (2017)*, 139(11). 2

[28] Xiaoshui Huang, Guofeng Mei, and Jian Zhang. Feature-metric registration: A fast semi-supervised approach for robust point cloud registration without correspondences. In *CVPR (2020)*. 6

[29] Xiaoshui Huang, Guofeng Mei, and Jian Zhang. Feature-metric registration: A fast semi-supervised approach for robust point cloud registration without correspondences. In *CVPR*, June 2020. 2

[30] Haobo Jiang, Zheng Dang, Zhen Wei, Jin Xie, Jian Yang, and Mathieu Salzmann. Robust outlier rejection for 3d registration with variational bayes. In *CVPR*, 2023. 2

[31] Haobo Jiang, Jianjun Qian, Jin Xie, and Jian Yang. Planning with learned dynamic model for unsupervised point cloud registration. *IJCAI (2021)*. 2

[32] Haobo Jiang, Yaqi Shen, Jin Xie, Jun Li, Jianjun Qian, and Jian Yang. Sampling network guided cross-entropy method for unsupervised point cloud registration. In *ICCV*, 2021. 1, 2

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[34] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *ECCV*, 2020. 1

[35] Jiayuan Li, Qingwu Hu, and Mingyao Ai. Point cloud registration based on one-point ransac and scale-annealing biweight estimation. *IEEE Transactions on Geoscience and Remote Sensing*, 2021. 3

[36] Jiahao Li, Changhao Zhang, Ziyao Xu, Hangning Zhou, and Chi Zhang. Iterative distance-aware similarity matrix convolution with mutual-supervised point elimination for efficient point cloud registration. In *ECCV (2020)*. 2, 6

[37] Jiayuan Li, Pengcheng Zhao, Qingwu Hu, and Mingyao Ai. Robust point cloud registration based on topological graph and cauchy weighted lq-norm. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020. 3

[38] Xiang Li, Lingjing Wang, and Yi Fang. PC-Net: Unsupervised point correspondence learning with neural networks. In *3DV (2019)*. 2

[39] Xiang Li, Lingjing Wang, and Yi Fang. Unsupervised partial point set registration via joint shape completion and registration. *arXiv preprint arXiv:2009.05290 (2020)*. 2

[40] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 5

[41] Yinlong Liu, Chen Wang, Zhijian Song, and Manning Wang. Efficient global point cloud registration by matching rotation invariant features through translation search. In *ECCV*, 2018. 1, 2

[42] Ameesh Makadia, Alexander Patterson, and Kostas Daniilidis. Fully automatic registration of 3d point clouds. In *CVPR*, 2006. 1, 3

[43] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 2015. 1

[44] Eitan Marder-Eppstein. Project tango. In *ACM SIGGRAPH 2016 Real-Time Live!* 2016. 1

[45] Nicolas Mellado, Dror Aiger, and Niloy J Mitra. Super 4pcs fast global pointcloud registration via smart indexing. In *Computer graphics forum*, 2014. 2, 6

[46] Mustafa Mohamad, Mirza Tahir Ahmed, David Rappaport, and Michael Greenspan. Super generalized 4pcs for 3D registration. In *3DV (2015)*. 2

[47] Mustafa Mohamad, David Rappaport, and Michael Greenspan. Generalized 4-points congruent sets for 3D registration. In *3DV (2014)*, volume 1, pages 83–90. 2

[48] G Dias Pais, Srikumar Ramalingam, Venu Madhav Govindu, Jacinto C Nascimento, Rama Chellappa, and Pedro Miraldo. 3DRegNet: A deep neural network for 3D point registration. In *CVPR (2020)*. 2

[49] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *ICCV*, 2019. 1

[50] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019. 6

[51] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *CVPR*, 2019. 1

[52] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 2, 4

[53] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NIPS*, 2017. 2

[54] 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 *ICCV*, 2017. 1

[55] Gregory C Sharp, Sang W Lee, and David K Wehe. ICP registration using invariant features. *IEEE Transactions on Pattern Analysis and Machine Intelligence (2002)*, 24(1):90–102. 2

[56] Yaqi Shen, Le Hui, Haobo Jiang, Jin Xie, and Jian Yang. Reliable inlier evaluation for unsupervised point cloud registration. In *AAAI*, 2022. 1, 2, 7

[57] Julian Straub, Trevor Campbell, Jonathan P How, and John W Fisher. Efficient global point cloud alignment using bayesian nonparametric mixtures. In *CVPR*, 2017. 1, 2

[58] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *ECCV*, 2018. 1

[59] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. *arXiv preprint arXiv:1809.10790*, 2018. 1

[60] Yue Wang and Justin M Solomon. PRNet: Self-supervised learning for partial-to-partial registration. *arXiv preprint arXiv:1910.12240 (2019)*. 2

[61] Yue Wang and Justin M Solomon. Deep closest point: Learning representations for point cloud registration. In *ICCV*, 2019. 1, 2, 3, 4, 5, 6

[62] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *CVPR*, 2018. 1

[63] Yusheng Xu, Richard Boerner, Wei Yao, Ludwig Hoegner, and Uwe Stilla. Pairwise coarse registration of point clouds in urban scenes using voxel-based 4-planes congruent sets. *ISPRS journal of photogrammetry and remote sensing (2019)*, 151:106–123. 2

[64] Heng Yang and Luca Carlone. A polynomial-time solution for robust registration with extreme outlier rates. *arXiv preprint arXiv:1903.08588*, 2019. 6

[65] Heng Yang, Jingnan Shi, and Luca Carlone. Teaser: Fast and certifiable point cloud registration. *IEEE Transactions on Robotics*, 2020. 3, 6

[66] Jiaolong Yang, Hongdong Li, and Yunde Jia. Go-ICP: Solving 3D registration efficiently and globally optimally. In *ICCV (2013)*. 2

[67] Zi Jian Yew and Gim Hee Lee. Rpm-net: Robust point matching using learned features. In *CVPR*, 2020. 1, 2, 3, 4, 6

[68] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, 2021. 4

[69] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *ICCV*, 2019. 1

[70] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In *ECCV*, 2016. 6

[71] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*, 2018. 4

[72] Jing Zhu and Yi Fang. Reference grid-assisted network for 3D point signature learning from point clouds. In *WACV (2020)*. 2

[73] Menglong Zhu, Konstantinos G Derpanis, Yinfei Yang, Samarth Brahmbhatt, Mabel Zhang, Cody Phillips, Matthieu Lecce, and Kostas Daniilidis. Single image 3d object detection and pose estimation for grasping. In *ICRA*, 2014. 1