Structure-Aware Correspondence Learning for Relative Pose Estimation

Yihan Chen[1]   Wenfei Yang[1]   Huan Ren[1]   Shifeng Zhang[3]   Tianzhu Zhang[1,2*]   Feng Wu[1]

[1]University of Science and Technology of China

[2]National Key Laboratory of Deep Space Exploration, Deep Space Exploration Laboratory

[3]Sangfor Technologies

{yihanchen, rh_hr_666}@mail.ustc.edu.cn, zhangshifeng@sangfor.com.cn

{yangwf, tzzhang, fengwu}@ustc.edu.cn

Abstract

Relative pose estimation provides a promising way for achieving object-agnostic pose estimation. Despite the success of existing 3D correspondence-based methods, the reliance on explicit feature matching suffers from small overlaps in visible regions and unreliable feature estimation for invisible regions. Inspired by humans' ability to assemble two object parts that have small or no overlapping regions by considering object structure, we propose a novel Structure-Aware Correspondence Learning method for Relative Pose Estimation, which consists of two key modules. First, a structure-aware keypoint extraction module is designed to locate a set of kepoints that can represent the structure of objects with different shapes and appearance, under the guidance of a keypoint based image reconstruction loss. Second, a structure-aware correspondence estimation module is designed to model the intra-image and inter-image relationships between keypoints to extract structure-aware features for correspondence estimation. By jointly leveraging these two modules, the proposed method can naturally estimate 3D-3D correspondences for unseen objects without explicit feature matching for precise relative pose estimation. Experimental results on the CO3D, Objaverse and LineMOD datasets demonstrate that the proposed method significantly outperforms prior methods, i.e., with $5.7°$ reduction in mean angular error on the CO3D dataset.

1. Introduction

Object pose estimation aims to estimate the 3D translation and 3D rotation of an object from a single image. It plays a crucial role in many real-world applications such as augmented reality (AR) [2, 26], robotic manipulation [27, 44, 45] and autonomous driving [4, 9], drawing increasing attention in recent years. Early works [12, 13, 41, 46] mainly focus on instance-level pose estimation, where the model is trained to estimate the pose for a specific object. However, these methods cannot generalize to other objects. Consequently, the category-level pose estimation [42] is intro-

Figure 1. a) **Task depiction**. Predict the relative pose ΔR between the query and reference images. b) **2D correspondence-based methods** extract keypoints to conduct 2D-2D matching for pose estimation. c) **3D correspondence-based methods** lift 2D features into 3D voxel features and conduct 3D-3D matching for pose estimation. d) **Our method** bypasses the feature matching and directly regresses 3D correspondences for pose estimation.

duced, where the model is trained to estimate the pose for different instances within the same category. Nevertheless, they can hardly generalize to unseen objects of other categories, limiting their application potentials.

To further improve the generalization ability, recent works [19, 48, 49] have shifted towards relative pose estimation for unseen objects. It requires only a single reference image of a novel object to estimate the poses of new images, as shown in Figure 1 (a), providing a promising way to achieve object-agnostic pose estimation. Existing relative pose estimation methods can be broadly categorized into three types, 2D correspondence-based methods [33, 36], hypothesis-and-verification-based methods [19, 48, 49] and 3D correspondence-based methods [49, 50]. 2D correspondence-based methods [33, 36] extract keypoints directly from two images and compute the relative pose based on keypoint matching. However, as shown in Figure 1 (b), these methods suffer from small overlapping regions caused by large pose variations, making it difficult to establish reliable correspondence. Hypothesis-and-verification-based methods [19, 48, 49] sample a large number of pose

arXiv:2503.18671v1 [cs.CV] 24 Mar 2025

hypotheses and then evaluate the score of each hypothesis through global feature matching or score regression. Nevertheless, these methods rely on a discrete sampling process and fail to adequately model the continuous pose space, which can only account for coarse pose estimation. Moreover, the verification of numerous pose hypotheses incurs significant computational costs. 3D correspondence-based methods [49, 50] lift 2D features into 3D voxel features and then conduct 3D-3D matching to estimate the pose, as shown in Figure 1 (c). Notably, these methods can establish correspondences even in invisible regions, showing promising potential. However, it's difficult to infer 3D features of invisible regions from 2D surface features without extra information, resulting in unreliable 3D matching for invisible regions. Moreover, dense 3D-3D matching process incurs high computational costs due to the cubic complexity.

Different from existing methods that rely on explicit feature matching, we take inspiration from how humans assemble two object parts that have small or no overlapping regions. As shown in Figure 1 (a), the query and reference images represent the front and back of a suitcase, with only a small overlap at the top. By considering structural details like shape, the position of the handles, and the color pattern, we humans can intuitively infer how these two parts should be assembled to form a complete suitcase. This process of mentally assembling parts is actually akin to determining their relative pose. Inspired by this intuitive assembly process, we design a framework that leverages such object structural details to map keypoints from the query image into the reference 3D coordinate space, thus naturally establishing 3D-3D correspondences without explicit feature matching, as illustrated in Figure 1 (d). Nevertheless, it is non-trivial because of the following challenges: (1) **How to represent the structure of each object part.** To reason how object parts in query image and reference image should be assembled, the model must understand the structure of these object parts. However, different objects or object parts captured in different views often exhibit significant variations in appearance and shape, making it challenge to design a method that can handle these situations well. (2) **How to extract structure-aware features for correspondence estimation.** While humans can naturally assemble two object parts with complementary structures, it's non-trivial for the neural network. To accurately map points in the query image into the reference coordinate space, it is crucial to encode the structure information of parts in query and reference image into the point features.

Based on the above discussion, we propose a Structure-Aware Correspondence Learning method for Relative Pose Estimation, which consists of a structure-aware keypoint extraction module and a structure-aware correspondence estimation module. Our key insight is to represent the structure of different parts through a set of keypoints and extract structure-aware keypoint features for correspondence estimation. **The structure-aware keypoint extraction module** is designed to locate a set of sparse keypoints that can well represent the structure of different object parts. Specifically, to deal with the significant shape and appearance variations, we use a set of learnable queries to interact with image features to produce image-specific keypoint detectors. We first compute similarities between keypoint detectors and image features to generate keypoint heatmaps, from which keypoint coordinates and features are derived. To guide the learning of the keypoint extraction module, we design an image reconstruction loss by constraining keypoint features and coordinates to reconstruct the image. The intuition behind this loss is that if these keypoints represent the object structure well, the model can recover the original image from them. **The structure-aware correspondence estimation module** is proposed to extract structure-aware keypoint features for correspondence estimation. To incorporate intra-image structure information, we use the self-attention mechanism that integrates relative keypoint positions to aggregate features from other keypoints in the same image. To incorporate inter-image structure information, we apply the cross-attention mechanism to aggregate keypoint features from the other image. Consequently, the extracted structure-aware keypoint features enables the network to perceive how the object parts in two images should be assembled, which can facilitate the correspondence estimation. Given these structured-aware keypoint features, we lift 2D keypoints to 3D space within the query coordinate system and regress their corresponding 3D coordinates within the reference coordinate system to establish 3D-3D correspondences. Finally, we estimate the relative pose with 3D-3D correspondence by employing a weighted Singular Value Decomposition, ensuring end-to-end optimization.

In summary, our contributions are as follows:
- We propose a novel structure-aware correspondence learning method for relative pose estimation, which can establish robust 3D-3D correspondence without explicit feature matching.
- We propose two key designs, a structure-aware keypoint extraction module that can well represent the structure of different object parts, and a structure-aware correspondence estimation module that can help the keypoints to aggregate structure-aware features for robust correspondence estimation.
- Experimental results on three challenging datasets, CO3D [30], Objaverse [6] and LineMOD [14], demonstrate the state-of-the-art performance of our method.

2. Related Work

2.1. Instance-Level Object Pose Estimation

Instance-level pose estimation methods [12, 13, 41, 46] predict the 6D pose of specific known objects by lever-

aging CAD models. Recent approaches include correspondence, template, voting, and direct regression methods. Correspondence-based methods [18, 29, 47] establish matches between inputs and CAD models, then estimate pose via PnP [7] or similar algorithms. Template-based methods [17, 37] match inputs to pre-defined templates, treating pose estimation as a matching problem. Voting-based methods [24, 39] aggregate votes from pixels or points, either by keypoints or direct pose prediction. Regression-based methods [8, 22] directly predict 6D poses from images or depth data, simplifying the pipeline. Although effective for known objects, these methods struggle to generalize to unseen objects.

2.2. Category-level Object Pose Estimation

To extend pose estimation beyond specific instances, researchers have developed category-level methods [20, 21, 23, 31, 32, 38, 42] for estimating object poses within predefined categories without CAD models. These methods fall into two categories: shape prior-based and shape prior-free. Shape prior-based approaches [20, 38, 42] utilize CAD-derived priors for alignment or direct regression of poses. In contrast, shape prior-free methods [5, 21, 23, 31, 32] remove reliance on priors, learning features directly from input data to estimate poses. Although category-level methods improve generalization over instance-specific methods, they still struggle to generalize to unseen categories.

2.3. Relative Object Pose Estimation

To enhance the generalization of pose estimation, recent work [19, 48–50] has focused on relative pose estimation for unseen objects. Unlike instance-specific methods, these approaches require only a single reference image of a novel object to estimate the relative pose of image, making them highly suitable for applications where data acquisition is costly. Current methods can be broadly categorized into three types: 2D correspondence-based methods [25, 33, 36], hypothesis-and-verification-based methods [19, 28, 48, 49], and 3D correspondence-based methods [49, 50]. 2D correspondence-based methods [25, 33, 36] establish correspondences between keypoints from both images and use these to compute relative pose. Learned feature-based approaches like SuperGlue [33] and LoFTR [36] have achieved robustness in feature matching under moderate viewpoint changes and lighting variations. However, with only a single reference image and substantial viewpoint differences, these methods struggle due to their sensitivity, significantly affecting accuracy [48, 49]. Hypothesis-and-verification-based methods [19, 28, 48, 49] mitigate these challenges by generating multiple pose hypotheses over the rotation space and evaluating them through similarity networks, as in RelPose [48] and RelPose++ [19]. While effective in handling larger viewpoint differences,

these methods require extensive sampling and verification, resulting in high computational costs that restrict their suitability for real-time applications. As an alternative, 3D correspondence-based methods [49, 50] lift 2D features into 3D voxel features and then conduct 3D-3D matching to estimate the pose. For example, DVMNet [50] uses lifted 3D voxel features to facilitate matching even in invisible regions. However, it's difficult to infer 3D features solely from 2D surface image features, which leads to unreliable 3D matching. Moreover, the dense 3D-3D matching incurs high computational costs due to the cubic complexity.

3. Method

3.1. Overview

We tackle the problem of estimating the relative pose between a query image and a reference image belonging to previously unseen object categories. Specifically, we denote the set of object categories available during training as ω_{train} and the set of categories used during testing as ω_{test}, where $\omega_{\text{train}} \cap \omega_{\text{test}} = \emptyset$. This setup introduces several key challenges, including significant viewpoint variations between the images, minimal overlapping regions, and the generalization to novel object categories. Since the translation component of 6D pose can be reliably estimated through existing 2D detection techniques [11, 16], we focus on addressing the more challenging 3D rotation estimation, which is the same with the mainstream methods [48–50].

The framework of our method is shown in Figure 2 (a). Given the query image I_q and the reference image I_r, we first employ a shared feature extractor to obtain feature maps $\mathbf{F}_q$ and $\mathbf{F}_r$. These feature maps are subsequently passed through symmetric attention blocks [40] to enhance the image features mutually. Afterward, we use the proposed structure-aware keypoint extraction module to extract keypoints independently from each image. Then, by leveraging our proposed structure-aware correspondence estimation module, we facilitate both intra-image and inter-image feature interactions for the keypoints extracted from the query image, enabling structure-aware feature aggregation. With these updated keypoint features, we lift 2D keypoints to 3D space and regress their corresponding 3D coordinates within the reference coordinate system. Finally, based on the established correspondences, the relative rotation $\Delta \mathbf{R}$ is obtained via a weighted Singular Value Decomposition (wSVD) algorithm [3].

3.2. Feature Extraction

We first utilize a pre-trained backbone [43] to extract image features from both the query and reference images, yielding feature maps $\mathbf{F}_q, \mathbf{F}_r \in \mathbb{R}^{H \times W \times C}$. These feature maps are then processed through a two-step attention mechanism with Multi-Head Self-Attention (MHSA) and Multi-Head

Figure 2. a) Overview of the proposed method. b) Illustration of the Structure-Aware Keypoint Extraction module. We initialize a set of learnable queries that interact with image features to extract keypoints representing the object's structure. And we further employ a reconstructor and $\mathcal{L}_{\text{rec}}$ for supervision of keypoints extraction. c) Illustration of the Structure-Aware Correspondence Estimation module. We employ ROPE and an attention mechanism to extract structure-aware features for 3D correspondence estimation.

Cross-Attention (MHCA) layers to update features:

$$\begin{aligned}
\tilde{\mathbf{F}}_q^{(l-1)} &= \text{MHSA}(\mathbf{F}_q^{(l-1)}) + \mathbf{F}_q^{(l-1)}, \\
\mathbf{F}_q^{(l)} &= \text{MHCA}(\tilde{\mathbf{F}}_q^{(l-1)}, \mathbf{F}_r^{(l-1)}) + \tilde{\mathbf{F}}_q^{(l-1)}.
\end{aligned} \quad (1)$$

The same operations are performed for $\mathbf{F}_r$. Unlike previous works [49, 50], the parameters in the MHSA and MHCA modules are shared across the query (q) and reference (r) images, ensuring consistency between the learned features. After L layers of such interactions, we obtain the final feature maps $\mathbf{F}_q^{(L)}$ and $\mathbf{F}_r^{(L)}$.

To suppress the influence of background features, we apply a lightweight mask predictor to obtain an object mask to refine the feature maps. Formally, the mask $\mathbf{M}_q$ and $\mathbf{M}_r$ are derived from $\mathbf{F}_q^{(L)}$ and $\mathbf{F}_r^{(L)}$, respectively:

$$\begin{aligned}
\mathbf{M}_q &= g(\mathbf{F}_q^{(L)}), \\
\mathbf{M}_r &= g(\mathbf{F}_r^{(L)}).
\end{aligned} \quad (2)$$

We use binary cross-entropy (BCE) loss and ground truth masks to compute the mask prediction loss as follows:

$$\mathcal{L}_{\text{mask}} = \text{BCE}(\mathbf{M}_q, \mathbf{M}_q^{\text{gt}}) + \text{BCE}(\mathbf{M}_r, \mathbf{M}_r^{\text{gt}}) \quad (3)$$

The final feature maps are obtained by element-wise multiplication with these object masks, effectively retaining only object-relevant features and reducing interference from the background:

$$\begin{aligned}
\mathbf{F}_q' &= \mathbf{F}_q^{(L)} \odot \mathbf{M}_q, \\
\mathbf{F}_r' &= \mathbf{F}_q^{(L)} \odot \mathbf{M}_r.
\end{aligned} \quad (4)$$

3.3. Structure-Aware Keypoint Extraction

As introduced in Section 1, a key challenge is how to effectively represent the structure of object parts, as they often exhibit significant variations in appearance and shape. To address these issues, we propose the structure-aware keypoint extraction module to adaptively select keypoints with structural significance, as illustrated in Figure 2 (b).

In the following, we use the query image as an example to illustrate the keypoint detection process. Specifically, we initialize a set of learnable queries, denoted as $\mathbf{Q} \in \mathbb{R}^{N_{\text{kpt}} \times C}$, where N_{kpt} is the number of keypoints, and C is the feature dimension. To convert the this queries into image-specific keypoint detectors $\tilde{\mathbf{Q}}_q$, we again use the attention mechanism to update these queries with the image features, so as to adapt to the content of different images.

$$\tilde{\mathbf{Q}}_q = \text{MHCA}(\mathbf{Q}, \mathbf{F}_q') + \mathbf{Q}. \quad (5)$$

Next, we compute the similarity between these image-specific keypoint detectors and the image features, gener-

ating keypoints heatmap $\mathbf{H}_q \in \mathbb{R}^{N_{\text{kpt}} \times H \times W}$:

$$\mathbf{H}_q = \text{softmax}\left(\tilde{\mathbf{Q}}_q \cdot \mathbf{F}'_q{}^{\top}\right). \qquad (6)$$

After that, we derive the spatial coordinates and the corresponding features for all keypoints by performing a weighted averaging based on the heatmap $\mathbf{H}_q$:

$$\mathbf{X}_{kpt,q} = \sum_{h=1}^{H} \sum_{w=1}^{W} \mathbf{H}_q(h, w) \cdot (h, w), \qquad (7)$$

$$\mathbf{F}_{kpt,q} = \sum_{h=1}^{H} \sum_{w=1}^{W} \mathbf{H}_q(h, w) \cdot \mathbf{F}'_q(h, w), \qquad (8)$$

where $\mathbf{X}_{kpt,q} \in \mathbb{R}^{N_{\text{kpt}} \times 2}$ denotes the spatial coordinates of all keypoints, and $\mathbf{F}_{kpt,q} \in \mathbb{R}^{N_{\text{kpt}} \times C}$ denotes their corresponding features. However, without explicit constraints, the extracted keypoints often cluster within limited regions, reducing their effectiveness in capturing comprehensive structural information of the object part. To address this, we introduce an image reconstruction loss that drives keypoints to cover semantically rich regions of the object by reconstructing its foreground solely from the keypoint features and coordinates.

$$\hat{\mathbf{I}}_q = f(\mathbf{X}_{\text{kpt},q}, \mathbf{F}_{\text{kpt},q}), \qquad (9)$$

where $f(\cdot, \cdot)$ is a lightweight decoder, and $\hat{\mathbf{I}}_q$ denotes the reconstructed query image. The reconstruction loss consists of an L_2 loss for pixel-wise similarity and a VGG-based perceptual loss:

$$\mathcal{L}_{\text{rec},q} = \lambda_1 \|\hat{\mathbf{I}}_q - \mathbf{I}_q\|_2^2 + \lambda_2 \sum_l \|\phi_l(\hat{\mathbf{I}}_q) - \phi_l(\mathbf{I}_q)\|_2^2, \qquad (10)$$

where $\phi_l(\cdot)$ denotes the feature map from the l-th layer of the VGG network [34], and λ_1, λ_2 are weighting factors. The L_2 loss ensures pixel accuracy, while the perceptual loss maintains semantic consistency. By reconstructing the image, we can optimize keypoint distribution end-to-end, ensuring these keypoints cover semantically rich regions of the object's surface and enhance structural representation.

By applying the above process to both the query and reference images, we obtain structurally significant keypoints that effectively represent the object's structure.

3.4. Structure-Aware Correspondence Estimation

Given the 2D keypoint coordinates $\mathbf{X}_{kpt,q}$, $\mathbf{X}_{kpt,r}$ and associated features $\mathbf{F}_{kpt,q}$, $\mathbf{F}_{kpt,r}$ from both the query and reference images, we extract structure-aware features to lift 2D keypoints to 3D space within the query coordinate system and regress their corresponding 3D coordinates within the reference coordinate system, establishing a set of 3D correspondences for relative pose estimation.

Specifically, for keypoint features $\mathbf{F}_{\text{kpt},q} \in \mathbb{R}^{N_{\text{kpt}} \times C}$ extracted from the query image, we refine these features using self-attention with rotational positional encoding (ROPE) [35], enabling the keypoint features to perceive the intra-image structure. Here, we denote the ROPE encoding as $R(\cdot)$, and $\circledast$ indicates the ROPE positional fusion operation as used in the original method [35].

$$\tilde{\mathbf{F}}_{\text{kpt},q} = \text{MHSA}(\mathbf{F}_{\text{kpt},q} \circledast R(\mathbf{X}_{\text{kpt},q})). \qquad (11)$$

We then apply cross-attention to aggregate structure information from reference image, where the refined keypoint features from the query image act as queries, and the reference keypoint features are used as keys and values:

$$\hat{\mathbf{F}}_{\text{kpt},q} = \text{MHCA}\left(\tilde{\mathbf{F}}_{\text{kpt},q} \circledast R(\mathbf{X}_{\text{kpt},q}), \, \mathbf{F}_{\text{kpt},r} \circledast R(\mathbf{X}_{\text{kpt},q})\right). \qquad (12)$$

These attention mechanisms help capture intra- and inter-image relationships, enabling the keypoint features for robust 3D correspondence estimation. With the updated keypoint features, we lift the 2D keypoints into 3D space within the query coordinate system by regressing a pseudo-depth value $d_{i,q}$ for each keypoint:

$$d_{i,q} = \text{MLP}_{\text{depth}}(\hat{\mathbf{f}}_{i,q}), \qquad (13)$$

where $d_{i,q}$ denotes the pseudo-depth of the i-th keypoint in the query image, and $\hat{\mathbf{f}}_{i,q}$ denotes the i-th keypoint feature from the updated keypoint feature set $\hat{\mathbf{F}}_{\text{kpt},q}$. By concatenating this depth value with the corresponding 2D coordinates, we obtain the 3D coordinates for each keypoint in the query coordinate system:

$$\mathbf{x}_{i,q}^{(\mathcal{Q})} = [\mathbf{x}_{i,q}, d_{i,q}] \in \mathbb{R}^3. \qquad (14)$$

The superscript $(\mathcal{Q})$ indicates 3D coordinates in the query system and subscript q indicates the keypoint extracted from the query image. The notation $(\mathcal{R})$ and r represent the reference correspondingly. Similarly, we estimate the corresponding 3D coordinates for the keypoints in the reference coordinate system using another MLP, which takes the updated keypoint feature $\hat{\mathbf{f}}_{i,q}$ and corresponding 3D coordinate in the query system as input:

$$\mathbf{x}_{i,q}^{(\mathcal{R})}, c_i = \text{MLP}_{\text{ref}}([\hat{\mathbf{f}}_{i,q}, PE(\mathbf{x}_{i,q}^{(\mathcal{Q})})]), \qquad (15)$$

where $\mathbf{x}_{i,q}^{(\mathcal{R})}$ denotes the estimated 3D coordinates of the i-th keypoint in the reference coordinate system, and $c_i \in [0, 1]$ denotes the confidence score of this keypoint. Based on the 3D coordinates of the same keypoints within both the query and reference systems, we can naturally establish 3D-3D correspondences, which helps determine the relative pose. To ensure the accuracy of these correspondences, we propose a loss function that supervises the predicted 3D coordinates. Specifically, the ground truth 3D coordinates in

the reference system are computed using the ground truth relative rotation matrix $\Delta\mathbf{R}_{\text{gt}}$:

$$\mathbf{x}_{i,q,\text{gt}}^{(\mathcal{R})} = \Delta\mathbf{R}_{\text{gt}} \cdot \mathbf{x}_{i,q}^{(\mathcal{Q})}. \tag{16}$$

To align the predicted 3D coordinates $\mathbf{x}_{i,q}^{(\mathcal{R})}$ with these ground truth values, we define the 3D keypoint loss $\mathcal{L}_{\text{pts}}$ as follows:

$$\mathcal{L}_{\text{pts}} = \frac{1}{N_{\text{kpt}}} \sum_{i=1}^{N_{\text{kpt}}} \left(c_i \cdot e_{i,r} - \alpha \log(c_i) \right), \tag{17}$$

where c_i denotes the confidence score of each keypoint, and α is a hyperparameter controlling the influence of the confidence score. The 3D keypoint error $e_{i,r}$ measures the discrepancy between the estimated and ground truth coordinates:

$$e_{i,r} = \frac{1}{2} \left(\|\mathbf{x}_{i,q}^{(\mathcal{R})} - \text{sg}(\mathbf{x}_{i,q,\text{gt}}^{(\mathcal{R})})\|_2^2 + \|\text{sg}(\mathbf{x}_{i,q}^{(\mathcal{R})}) - \mathbf{x}_{i,q,\text{gt}}^{(\mathcal{R})}\|_2^2 \right), \tag{18}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. This symmetric loss penalizes deviations in both coordinate systems, ensuring the 3D coordinates $\mathbf{x}_{i,q}^{(\mathcal{Q})}$ and $\mathbf{x}_{i,q}^{(\mathcal{R})}$ are accurately estimated, ultimately leading to reliable 3D correspondences.

3.5. Pose Estimation via 3D Correspondences

After obtaining the 3D keypoint correspondences, following previous work [50], we employ a weighted Singular Value Decomposition (wSVD) approach to solve the relative rotation $\Delta\mathbf{R}$:

$$\Delta\mathbf{R} = \arg\min_{\mathbf{R}} \sum_{i=1}^{N_q} c_i \|\mathbf{x}_{i,q}^{(\mathcal{R})} - \mathbf{R}\mathbf{x}_{i,q}^{(\mathcal{Q})}\|_2^2, \tag{19}$$

where $\mathbf{x}_{i,q}^{(\mathcal{R})}$ and $\mathbf{x}_{i,q}^{(\mathcal{Q})}$ are the 3D coordinates of the i-th keypoint in the query and reference coordinate systems, and c_i denotes corresponding confidence score estimated earlier. The optimization seeks the rotation matrix $\Delta\mathbf{R}$ that best aligns these two sets of 3D coordinates.

To solve the minimization problem for estimating the optimal rotation matrix $\Delta\mathbf{R}$, we follow the approach of utilizing SVD method to derive the optimal alignment between two sets of 3D keypoints, as used in prior works on point cloud registration [1]. Specifically, we first compute the covariance matrix $\mathbf{H}$ using the query and reference 3D keypoints along with their confidence scores:

$$\mathbf{H} = \sum_{i=1}^{N_q} c_i \mathbf{x}_{i,q}^{(\mathcal{Q})} (\mathbf{x}_{i,q}^{(\mathcal{R})})^\top. \tag{20}$$

We then perform Singular Value Decomposition (SVD) on the covariance matrix $\mathbf{H}$:

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top. \tag{21}$$

The optimal relative rotation $\Delta\mathbf{R}$ is then obtained as:

$$\Delta\mathbf{R} = \mathbf{V}\mathbf{U}^\top. \tag{22}$$

To align the estimated relative rotation $\Delta\mathbf{R}$ with the ground truth $\Delta\mathbf{R}_{\text{gt}}$, we employ the L_1 loss:

$$\mathcal{L}_{\text{rot}} = \|q(\Delta\mathbf{R}) - q(\Delta\mathbf{R}_{\text{gt}})\|_1, \tag{23}$$

where $q(\Delta\mathbf{R})$ and $q(\Delta\mathbf{R}_{\text{gt}})$ denote the 6D representation for $\Delta\mathbf{R}$ and $\Delta\mathbf{R}_{\text{gt}}$ introduced in [51].

3.6. Training and Inference Details

Our framework optimizes a combined loss function during training, which includes the 3D keypoint loss $\mathcal{L}_{\text{pts}}$, the reconstruction loss $\mathcal{L}_{\text{rec}}$, the rotation loss $\mathcal{L}_{\text{rot}}$, and the mask loss $\mathcal{L}_{\text{mask}}$. The total loss can be expressed as:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{pts}} + \lambda_2 \mathcal{L}_{\text{rec}} + \lambda_3 \mathcal{L}_{\text{rot}} + \lambda_4 \mathcal{L}_{\text{mask}}, \tag{24}$$

where λ_1, λ_2, λ_3, and λ_4 are hyperparameters controlling the influence of each term. Additionally, during training, both query and reference images are used symmetrically. Specifically, the keypoints extracted from the reference image are also used to obtain a set of 3D correspondences, effectively providing additional training data that enhances efficiency and robustness. During inference, the model extracts keypoints from both query and reference images, but only regresses the 3D coordinates from the query image for the relative pose estimation.

4. Experiment

Datasets. Following previous works [49, 50], we evaluate our method on CO3D [30], Objaverse [6] and LineMOD [14], which are widely-used datasets for relative pose estimation. These datasets include diverse synthetic and real data across various object categories. The CO3D dataset contains 18,619 video sequences spanning 51 categories. Following [30], we train on 41 categories and test on 10 unseen categories to evaluate generalization. The Objaverse dataset consists of synthetic images rendered from 3D models across diverse viewpoints. We select 128 objects for testing and reserve the remaining for training. For LineMOD, we use calibrated real images of 13 household objects. The test set includes 5 objects, which are excluded from training to ensure complete separation.

Implementation Details. The Adam optimizer [15] is employed with an initial learning rate of 2×10^{-4}, which decays by a factor of 0.1 every 200 epochs. The model is trained for 400 epochs with a batch size of 80. All experiments are conducted on 4 NVIDIA RTX3090 GPUs, taking roughly 36 hours. Following previous works [19, 48–50], we crop the object from the image by utilizing the ground truth bounding box. Further details can be found in the supplementary material.

Table 1. Performance comparison with state-of-the-art methods on CO3D, Objaverse, and LineMOD datasets. Here we denote 2D correspondence-based methods as 2D, hypothesis-and-verification-based methods as H&V, 3D correspondence-based methods as 3D.

Method	Type	CO3D			Objaverse			LineMOD		
		mAE ↓	Acc@30° ↑	Acc@15° ↑	mAE ↓	Acc@30° ↑	Acc@15° ↑	mAE ↓	Acc@30° ↑	Acc@15° ↑
SuperGlue [33]	2D	67.2	45.2	37.7	102.4	15.1	12.1	64.8	26.2	14.3
LoFTR [36]	2D	77.5	37.9	33.1	134.1	9.6	7.7	84.5	24.2	13.5
ZSP [10]	2D	87.5	25.7	14.6	107.2	4.2	1.5	78.6	10.7	2.7
RelPose [48]	H&V	50.0	64.2	48.6	80.4	20.8	6.7	58.3	26.1	7.0
RelPose++ [19]	H&V	38.5	77.0	69.8	33.5	72.3	42.9	46.6	42.5	15.1
3DAHV [49]	H&V 3D	28.5	83.5	71.0	28.1	78.6	58.4	41.7	61.5	29.9
DVMNet [50]	3D	19.9	85.9	62.3	20.2	81.5	57.2	36.8	55.1	23.8
Ours	3D	**14.2**	**93.6**	**80.2**	**15.3**	**90.3**	**74.0**	**27.2**	**76.2**	**41.8**

Evaluation Metrics. Following previous works [49, 50], we evaluate our model using two metrics: mean angular error (mAE) and accuracy under predefined thresholds. The angular error θ between the predicted rotation ΔR and the ground truth ΔR_{gt} is calculated as:

$$\theta = \arccos\left(\frac{\mathrm{Tr}(\Delta R_{gt}^\top \Delta R) - 1}{2}\right), \qquad (25)$$

where ΔR_{gt} and ΔR are the ground truth and predicted rotation matrices, respectively. We also report accuracy as the percentage of test samples with an angular error below thresholds of $30°$ and $15°$.

4.1. Comparison with State-of-the-Art Methods

Table 1 presents a comprehensive comparison between our method and state-of-the-art methods on the CO3D, Objaverse, and LineMOD datasets. Our method consistently outperforms prior methods across all datasets and metrics.

Our approach significantly outperforms traditional 2D feature-based methods, such as SuperGlue [33] and LoFTR [36], primarily due to their inability to reliably match keypoints under large pose differences and minimal overlap areas. Furthermore, our method achieves superior results compared to hypothesis-and-verification-based methods like RelPose [48] and RelPose++ [19]. These methods rely on global features while ignoring local structural cues, and their use of discrete sampling limits their ability to model the continuous pose space accurately.

Compared to the 3D correspondence-based methods, such as DVMNet [50] and 3DAHV [49], our method demonstrates substantial improvements. On the CO3D dataset, our approach reduces mean angular error (mAE) by nearly $6°$, and improves Acc @ 30° and Acc @ 15° by approximately 8%. These gains are primarily because it is difficult to infer reliable 3D features without extra information, often leading to incorrect matches and unreliable 3D correspondences . In contrast, our approach avoids the matching process and directly regresses accurate 3D correspondences using structural information.

Table 2. Ablation study on the effectiveness of the Structure-Aware Keypoint Extraction module.

Setting	mAE ↓	Acc@30° ↑	Acc@15° ↑	MACs(G)
Dense	15.52	92.65	78.20	55.26
Random	20.15	88.34	68.99	49.59
Keypoint	**14.2**	**93.6**	**80.2**	50.05

Our method also achieves state-of-the-art results on the Objaverse and LineMOD datasets, which include a variety of synthetic and real-world object categories. The observed improvements in mAE and accuracy on these datasets demonstrate that the proposed method can generalize well to diverse conditions, providing a promising solution for real-world applications.

4.2. Ablation Studies

In this section, we conduct ablation studies to demonstrate the effectiveness of each design on the CO3D dataset.

Effects of the Structure-Aware Keypoint Extraction. To evaluate the effectiveness of our Structure-Aware Keypoint Extraction module, we conducted ablation experiments by replacing extracted keypoints with dense pixel features or randomly sampled points. As shown in Table 2, our module consistently outperforms both alternatives across all metrics. The results show that our keypoint extraction module effectively captures object structure, even with significant shape and appearance variations. In contrast, dense pixel features introduce irrelevant background noise or insignificant features and incur high computational costs, while random sampling lacks consistency in representing object structure. In summary, our method yields superior performance with fewer Multiply-Accumulate Operations (MACs), demonstrating efficiency and effectiveness.

Effects of the Structure-Aware Correspondence Estimation. To evaluate the effectiveness of the proposed Structure-Aware Correspondence Estimation module, we conducted ablation studies by removing key components: the self-attention and cross-attention mechanisms. These components are crucial for modeling relationships between

Table 3. Ablation study on the effectiveness of the Structure-Aware Correspondence Estimation module.

Setting	mAE ↓	Acc@30° ↑	Acc@15° ↑
w/o Self	17.79	90.22	72.48
w/o Cross	18.53	89.38	70.99
Dense Reg.	19.37	88.51	65.17
Global Reg.	21.97	86.76	68.16
Ours	**14.2**	**93.6**	**80.2**

Table 4. Ablation studies on the mask loss and confidence score.

Setting	mAE ↓	Acc@30° ↑	Acc@15° ↑
w/o $\mathcal{L}_{\text{mask}}$	16.37	91.68	76.82
w/o confidence	15.58	92.02	77.41
Ours	**14.2**	**93.6**	**80.2**

Table 5. Ablation study on the number of keypoints.

N_{kpt}	mAE ↓	Acc@30° ↑	Acc@15° ↑	MACs(G)
16	18.03	90.38	72.99	48.52
32	15.95	92.30	76.54	49.28
48	**14.2**	93.6	**80.2**	50.05
64	14.35	**93.85**	79.67	50.82

keypoints and extracting structure-aware features. As shown in Table 3, without them, the network struggles to capture intra- and inter-image structure, resulting in degraded correspondence estimation. Furthermore, to validate the effectiveness of directly estimating 3D keypoints, we replaced this process with a relative pose regression. We evaluated two variations: averaging point-wise pose regression and directly regressing the relative pose with global features. As shown in Table 3, directly regressing 3D keypoint coordinates is significantly more effective than regressing the entire rotation matrix. By focusing on keypoint coordinate regression, our method captures the underlying structural relationships between different parts of the object, ultimately resulting in more reliable 3D correspondence and precise pose estimation.

Effects of Mask Loss and Confidence Score. As shown in Table 4, both mask loss and confidence score estimation play important roles in improving performance. The mask loss helps the model focus on extracting foreground features, while the confidence score measures the reliability of the estimated 3D correspondences.

Effects of keypoint numbers. In Table 5, we show the impact of the number of keypoints N_{kpt}. It can be observed that as N_{kpt} increases, the performance improves, which is attributed to the fact that more keypoints help in better modeling the structure of the object. For balancing performance gains against computational efficiency (measured in MACs), we select $N_{\text{kpt}} = 48$ by default.

Figure 3. **Qualitative comparisons among LoFTR, DVMNet and our method.** We visualize the ground truth and predicted arrows. Blue indicates ground truth and green indicates prediction.

4.3. Visualization

Qualitative Results. The qualitative results of LoFTR [36], DVMNet [50] and our method on the LineMOD dataset are shown in Figure 3. Specifically, we visualize the object pose depicted in the query image. The query object pose is determined as $\mathbf{R}_q = (\Delta \mathbf{R})^{-1} \mathbf{R}_r$, where $\mathbf{R}_r$ denotes the object pose in the reference image. The green and blue arrows represent the visualization of $\mathbf{R}_q$ calculated from the prediction and the ground truth, respectively. It can be observed that our method demonstrates more accurate performance compared to previous matching-based methods, especially in cases where there are larger viewpoint differences between the query and reference images.

5. Conclusion

In this paper, we propose a Structure-Aware Correspondence Learning method for Relative Pose Estimation. Specifically, we introduce a structure-aware keypoint extraction module to identify keypoints that can represent the structure of objects with different shapes and appearance. Furthermore, we propose a correspondence estimation module that models relationships between keypoints to extract structure-aware features, enabling robust 3D correspondence regression without explicit feature matching. Comprehensive experiments on three datasets demonstrate the effectiveness of our method.

6. Acknowledgements

References

[1] K Somani Arun, Thomas S Huang, and Steven D Blostein. Least-squares fitting of two 3-d point sets. *IEEE Transactions on pattern analysis and machine intelligence*, PAMI-9 (5):698–700, 1987. 6

[2] Ronald T Azuma. A survey of augmented reality. *Presence: Teleoperators and Virtual Environments/MIT press*, 1997. 1

[3] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, pages 586–606. Spie, 1992. 3

[4] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017. 1

[5] Yamei Chen, Yan Di, Guangyao Zhai, Fabian Manhardt, Chenyangguang Zhang, Ruida Zhang, Federico Tombari, Nassir Navab, and Benjamin Busam. Secondpose: Se (3)-consistent dual-stream feature fusion for category-level pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9959–9969, 2024. 3

[6] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2, 6

[7] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3

[8] Ge Gao, Mikko Lauri, Yulong Wang, Xiaolin Hu, Jianwei Zhang, and Simone Frintrop. 6d object pose regression via supervised learning on point clouds. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3643–3649. IEEE, 2020. 3

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1

[10] Walter Goodwin, Sagar Vaze, Ioannis Havoutis, and Ingmar Posner. Zero-shot category-level object pose estimation. In *European Conference on Computer Vision*, pages 516–532. Springer, 2022. 7

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[12] 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 *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 11632–11641, 2020. 1, 2

[13] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3003–3013, 2021. 1, 2

[14] 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 *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I 11*, pages 548–562. Springer, 2013. 2, 6

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

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3

[17] Hongyang Li, Jiehong Lin, and Kui Jia. Dcl-net: Deep correspondence learning network for 6d pose estimation. In *European Conference on Computer Vision*, pages 369–385. Springer, 2022. 3

[18] Zhigang Li, Gu Wang, and Xiangyang Ji. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7678–7687, 2019. 3

[19] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. In *2024 International Conference on 3D Vision (3DV)*, pages 106–115. IEEE, 2024. 1, 3, 6, 7

[20] Jiehong Lin, Zewei Wei, Changxing Ding, and Kui Jia. Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks. In *European Conference on Computer Vision*, pages 19–34. Springer, 2022. 3

[21] Jiehong Lin, Zewei Wei, Yabin Zhang, and Kui Jia. Vi-net: Boosting category-level 6d object pose estimation via learning decoupled rotations on the spherical representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14001–14011, 2023. 3

[22] Muyuan Lin, Varun Murali, and Sertac Karaman. 6d object pose estimation with pairwise compatible geometric features. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10966–10973. IEEE, 2021. 3

[23] Xiao Lin, Wenfei Yang, Yuan Gao, and Tianzhu Zhang. Instance-adaptive and geometric-aware keypoint learning for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21040–21049, 2024. 3

[24] Xingyu Liu, Shun Iwase, and Kris M Kitani. Kdfnet: Learning keypoint distance field for 6d object pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent*

Robots and Systems (IROS), pages 4631–4638. IEEE, 2021. 3

[25] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 3

[26] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 22(12):2633–2651, 2015. 1

[27] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2901–2910, 2019. 1

[28] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Yinlin Hu, Renaud Marlet, Mathieu Salzmann, and Vincent Lepetit. Nope: Novel object pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17923–17932, 2024. 3

[29] 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 *Proceedings of the IEEE international conference on computer vision*, pages 3828–3836, 2017. 3

[30] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021. 2, 6

[31] Huan Ren, Wenfei Yang, Xiang Liu, Shifeng Zhang, and Tianzhu Zhang. Learning shape-independent transformation via spherical representations for category-level object pose estimation. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[32] Huan Ren, Wenfei Yang, Shifeng Zhang, and Tianzhu Zhang. Rethinking correspondence-based category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[33] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 1, 3, 7

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[35] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 5

[36] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 1, 3, 7, 8

[37] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *Proceedings of the european conference on computer vision (ECCV)*, pages 699–715, 2018. 3

[38] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 530–546. Springer, 2020. 3

[39] Meng Tian, Liang Pan, Marcelo H Ang, and Gim Hee Lee. Robust 6d object pose estimation by learning rgb-d features. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6218–6224. IEEE, 2020. 3

[40] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[41] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3343–3352, 2019. 1, 2

[42] 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*, pages 2642–2651, 2019. 1, 3

[43] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17969–17980, 2023. 3

[44] Bowen Wen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal. You only demonstrate once: Category-level manipulation from single visual demonstration. *arXiv preprint arXiv:2201.12716*, 2022. 1

[45] Chaozheng Wu, Jian Chen, Qiaoyu Cao, Jianchi Zhang, Yunxin Tai, Lin Sun, and Kui Jia. Grasp proposal networks: An end-to-end solution for visual learning of robotic grasps. *Advances in Neural Information Processing Systems*, 33:13174–13184, 2020. 1

[46] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 1, 2

[47] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1941–1950, 2019. 3

[48] Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *European Conference on Computer Vision*, pages 592–611. Springer, 2022. 1, 3, 6, 7

[49] Chen Zhao, Tong Zhang, and Mathieu Salzmann. 3d-aware hypothesis & verification for generalizable relative object

pose estimation. *arXiv preprint arXiv:2310.03534*, 2023. 1, 2, 3, 4, 6, 7

[50] Chen Zhao, Tong Zhang, Zheng Dang, and Mathieu Salzmann. Dvmnet: Computing relative pose for unseen objects beyond hypotheses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20485–20495, 2024. 1, 2, 3, 4, 6, 7, 8

[51] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. 6