Scalable Unseen Objects 6-DoF Absolute Pose Estimation with Robotic Integration

Jian Liu, Wei Sun, Kai Zeng, Jin Zheng, Hui Yang, Hossein Rahmani, Ajmal Mian, and Lin Wang

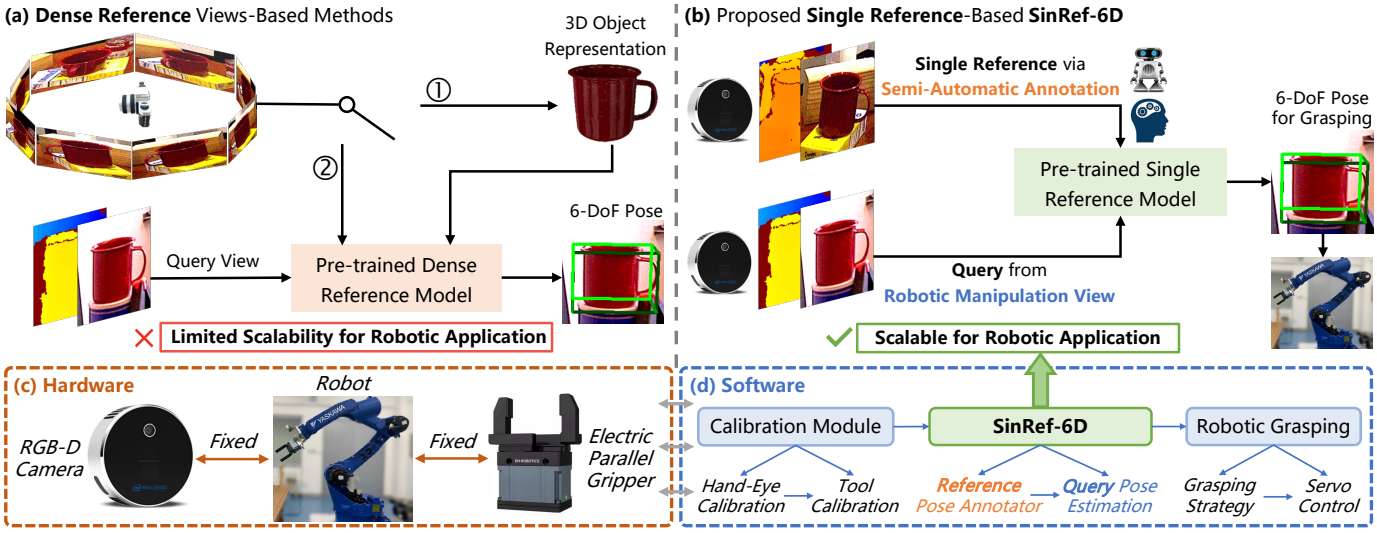

Fig. 1. Overview of the proposed task setup and robotic integration for unseen object 6-DoF absolute pose estimation tailored for practical robotic applications. **(a)** and **(b)** compare two types of manual reference view-based methods. **(a)** Dense reference views-based methods typically rely on ①: 3D object reconstruction or ②: template matching, which is time- and memory-consuming (*not suitable for robotic applications*). **(b)** The proposed method estimates unseen object pose from only a single reference view, providing enhanced efficiency and scalability (*suitable for robotic applications*). **(c)** and **(d)** are the detailed hardware and software architectures (see Sec. IV for further description) of our integrated robotic system. We also develop an efficient semi-automatic annotator based on the proposed task setup, enabling single reference annotation for unseen object within one minute.

***Abstract*—Pose estimation-guided unseen object 6-DoF robotic manipulation is a key task in robotics. However, the scalability of current pose estimation methods to unseen objects remains a fundamental challenge, as they generally rely on CAD models or dense reference views of unseen objects, which are difficult to acquire, ultimately limit their scalability. In this paper, we introduce a novel task setup, referred to as SinRef-6D, which addresses 6-DoF absolute pose estimation for unseen objects using only a single pose-labeled reference RGB-D image captured during robotic manipulation. This setup is more scalable yet technically nontrivial due to large pose discrepancies and the limited geometric and spatial information contained in a single view. To address these issues, our key idea is to iteratively establish point-wise alignment in a common coordinate system with state space models (SSMs) as backbones. Specifically, to handle large pose discrepancies, we introduce an iterative object-space point-wise alignment strategy. Then, Point and RGB SSMs are proposed to capture long-range spatial dependencies from a single view, offering superior spatial modeling capability with linear complexity. Once pre-trained on synthetic data, SinRef-6D can estimate the 6-DoF absolute pose of an unseen object using only a single reference view. With the estimated pose, we further develop a hardware-software robotic system and integrate the proposed SinRef-6D into it in real-world settings. Extensive experiments on six benchmarks and in diverse real-world scenarios demonstrate that our SinRef-6D offers superior scalability. Additional robotic grasping experiments further validate the effectiveness of the developed robotic system. The code and robotic demos are available at our homepage.**

***Index Terms*—Pose estimation, 6-DoF robotic manipulation, unseen object, single reference, state space model.**

This work is supported by the National Natural Science Foundation of China under Grants 62473141 and U22A2059, National Science and Technology Major Project of China under Grants 2026ZD1610900, Natural Science Foundation of Hunan Province under Grant 2024JJ5098, Open Foundation of the State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body, and Open Foundation of the Engineering Research Center of Multi-Mode Control Technology and Application for Intelligent System of the Ministry of Education. Ajmal Mian was supported by the Australian Research Council Future Fellowship Award funded by the Australian Government under Project FT210100268. (*Corresponding authors: Wei Sun; Hui Yang.*)

Jian Liu, Wei Sun, Kai Zeng, and Hui Yang are with the National Engineering Research Center for Robot Visual Perception and Control Technology, School of Artificial Intelligence and Robotics, Hunan University, Changsha, 410082, China. (e-mail: {jianliu, wei_sun, huiyang}@hnu.edu.cn).

Jin Zheng is with the School of Architecture and Art, Central South University, Changsha, 410082, China.

Hossein Rahmani is with the School of Computing and Communications, Lancaster University, LA1 4YW, United Kingdom.

Ajmal Mian is with the Department of Computer Science and Software Engineering, The University of Western Australia, WA 6009, Australia.

Jian Liu and Lin Wang are with the School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore.

I. Introduction

UNSEEN object 6-DoF robotic manipulation is a fundamental task underlying scalable robotic applications across diverse domains [1]–[4]. At its core lies the challenge of estimating the 6-DoF absolute pose of objects not encountered

during training [5]–[10]. Typically, 6-DoF pose comprises of 3-DoF rotation and 3-DoF translation of an object coordinate system relative to the camera coordinate system [11]–[17].

Object pose estimation methods can be divided into three main categories. *Instance-level methods* [18]–[26] have attained high precision but are limited to objects encountered during training. In contrast, *category-level methods* [27]–[34] can generalize to objects within the same category but still necessitate retraining for novel object categories. Furthermore, some *unseen object pose estimation methods* [35]–[41] have been proposed recently that *do not require retraining* for novel object categories, thereby exhibiting enhanced scalability.

Unseen object pose estimation methods can be further divided into two categories: *CAD model-based* [42]–[45], where a textured CAD model of the unseen object is required during training and inference; *manual reference view-based* [46]–[48], where a set of manually labeled reference views of the unseen object are required. Accurate textured CAD models can only be obtained with specialized equipment and expert knowledge, which hinders scalability in mobile devices [49], [50]. Since manual reference views are relatively easy to acquire, methods in this category offer greater scalability. Manual reference view-based methods typically solve pose through 3D object reconstruction or directly obtain coarse pose via template matching, as shown in Fig. 1 (a), where the switch indicates whether 3D reconstruction is required based on dense reference views. Dense reference views consume time for acquisition and memory for storage. Furthermore, template matching-based methods require the use of novel template generation techniques or an additional pose refinement overhead which further increases the computational complexity.

To address the aforementioned challenges, our motivation is to explore a CAD model-free, sparse reference view-based unseen object 6-DoF absolute pose estimation framework, eliminating the need for either 3D object reconstruction or template-based retrieval. Specifically, we formulate the task as the extreme case of sparse reference view, where *only a single reference is available*. Motivated by robotic manipulation scenarios, we design a *scalable label collection pipeline* where each unseen object is annotated with a single RGB-D reference view in a semi-automatic manner, while absolute pose recovery is obtained through the annotated reference view. The overview of our task setup is shown in Fig. 1 (b). The robot first captures the object from its default manipulation viewpoint, and a custom-developed annotator provides the corresponding 6-DoF pose label. Given only a single annotated view as the sparse reference prior of an unseen object, our goal is to accurately estimate its 6-DoF absolute pose from arbitrary novel viewpoints in different scenes. However, using a single reference introduces several unique challenges, including large pose discrepancies and limited spatial information.

With the task setup, we propose a scalable SinRef-6D framework. *Our key idea is to iteratively establish point-wise alignment between the single reference view and a query view in a common coordinate system to solve the 6-DoF pose of unseen objects.* SinRef-6D introduces two key components: 1) Iterative object-space point-wise alignment, which addresses large pose discrepancies by leveraging geometric and spatial consistency to refine pose estimation; 2) State Space Models (SSMs), which efficiently capture long-range spatial dependencies from single-view data, offering linear computational complexity and strong spatial modeling capability. Specifically, we propose to align the reference and query point clouds within the object coordinate system (Sec. III-C). Given the importance of spatial information for point-wise alignment and the need for a lightweight model for mobile deployment, we introduce Point and RGB SSMs (Sec. III-D) to establish point-wise alignment for pose solving (Sec. III-E). To handle the potentially large pose discrepancies between the reference and query views, we propose to iteratively refine the alignment in the object coordinate system, which gives more accurate and robust pose estimation (Sec. III-F). Furthermore, we develop a complete hardware-software robotic system that integrates the proposed SinRef-6D to evaluate its scalability in real-world scenarios (Sec. IV), as shown in Fig. 1 (c) and (d). Our main contributions are summarized as follows:

- We introduce an efficient and scalable task setup for unseen object 6-DoF absolute pose estimation using only a single reference view captured during robotic manipulation, eliminating the need for computation-intensive template matching and multi-view reconstruction. We further develop an integrated hardware-software robotic system tailored to the proposed task setup and framework, validating their efficacy in real-world scenarios.
- We propose an object-space point-wise alignment strategy with iterative refinement, facilitating direct alignment of query and reference views while effectively handling large pose discrepancies. This enhances geometric consistency and spatial awareness, enabling unseen object pose estimation without category-specific retraining.
- We propose Point and RGB SSMs to capture rich spatial information for establishing point-wise alignment, enabling efficient long-range spatial modeling with linear computational complexity.
- Extensive experiments demonstrate that our task setup and framework enable highly scalable 6-DoF robotic grasping of unseen objects in diverse environments.

The remainder of this paper is structured as follows. Sec. II reviews recent advances in unseen object pose estimation. Sec. III introduces the proposed task setup and corresponding framework. Sec. IV describes the developed 6-DoF robotic grasping system that integrates both hardware and software. Sec. V presents comprehensive experimental results that validate the scalability of SinRef-6D and the effectiveness of the robotic system. Finally, Sec. VI summarizes the paper.

II. Related Work

This section provides an overview of state-of-the-art methods in unseen object absolute (Sec. II-A and Sec. II-B) and relative (Sec. II-C) pose estimation, followed by a discussion on how our work differs from existing approaches.

A. CAD Model-based Methods

Research in the domain of CAD model-based methods first require obtaining the precise CAD model of the unseen object,

which is then used as prior knowledge for pose estimation. These methods can be further categorized into 1) feature matching-based and 2) template matching-based.

Feature matching-based methods [51]–[55] learn a model to match features between the observed image and CAD model, establishing 2D-3D or 3D-3D correspondences to estimate object pose. Specifically, GCPose [52] proposes a geometry correspondence-based approach that leverages generic, object-agnostic geometric features to establish clear and robust 3D-3D correspondences. SAM-6D [53] introduces a novel matching score based on semantics, appearance, and geometry to improve segmentation. For pose estimation, it employs a two-stage point matching model to establish dense 3D-3D correspondences. FreeZe [54] develops a method that combines visual and geometric features from various pre-trained models to improve pose prediction stability and accuracy. MatchU [55] proposes a technique for predicting object pose from RGB-D images by integrating 2D texture with 3D geometric cues.

Template matching-based methods [56]–[60] render multiple template views of the object with different poses from the CAD model. Then, they retrieve the template that best matches the observed image to obtain a coarse pose, followed by a refinement process to achieve accurate pose estimation. For example, MegaPose [58] proposes a render-and-compare-based method and a coarse-to-fine pose estimation strategy. GenFlow [59] introduces a shape-constrained recurrent flow framework that predicts optical flow between the query and template images while iteratively refining the pose. GigaPose [60] achieves fast and robust pose estimation by striking an effective balance between template matching and patch correspondences. FoundationPose [61] increases the quantity and diversity of synthetic data based on diffusion model and achieves superior performance through render-and-compare.

B. Manual Reference View-based Methods

To eliminate the need for a precise CAD model, manual reference view-based methods employ manual reference views as the prior knowledge for unseen objects. These methods can also be categorized into 1) feature matching-based and 2) template matching-based.

Feature matching-based methods [62]–[66] aim to establish 3D-3D correspondences between the query view and reference views, or 2D-3D correspondences between the query view and the 3D object representation reconstructed from reference views. Specifically, FS6D [62] proposes a dense prototype matching method to explore geometric and semantic relations between the query view and reference views, estimating the pose of unseen objects using only a few reference views. OnePose [63] first utilizes Structure from Motion (SfM) to reconstruct the 3D representation of the unseen object using all reference views, and then establishes 2D-3D correspondences between the query view and the reconstructed 3D representation using a graph attention network. OnePose++ [64] introduces a keypoint-free SfM method to reconstruct a semi-dense 3D representation of textureless objects by leveraging the detector-free feature matching approach LoFTR [67], enhancing robustness against textureless objects.

Template matching-based methods [68]–[72] primarily utilize a retrieval and refinement strategy. They directly use labeled reference views as templates to retrieve a coarse pose, followed by a refinement process to enhance accuracy. Specifically, LatentFusion [68] reconstructs 3D object representation and estimates translation using bounding boxes and depth values. Then, the initial rotation is determined by angle sampling and further refined through gradient updates using render and compare. Gen6D [69] first detects object bounding boxes, then compares the query and reference images via similarity scores to obtain an initial pose. Next, the pose is refined via a proposed refiner. FoundationPose [61] introduces an object-centric neural field to enable accurate 3D object modeling and RGB-D rendering, achieving performance comparable to instance-level methods. GS-Pose [70] joints segmentation and introduces a 3D gaussian splatting-based refiner, which simultaneously enhances the accuracy of object localization and pose estimation.

C. Unseen Object Relative Pose Estimation Methods

Relative object pose estimation [73]–[78] refers to computing the pose transformation of an object between two different views. 3DAHV [73] proposes a 3D-aware hypothesis-and-verification framework for relative pose estimation of unseen objects from a reference image, achieving robust generalization under large pose variations without relying on dense multi-view supervision. Building on this idea, DVMNet [74] introduces an end-to-end voxel-based framework that bypasses discrete hypothesis generation by directly aligning voxelized 3D features from two RGB images, resulting in improved accuracy and reduced computational cost. In contrast, NOPE [75] presents a fast, training-free method that estimates relative pose by predicting pose-conditioned viewpoint embeddings using an attention-enhanced U-Net, without requiring 3D models. While these methods demonstrate strong scalability, the absence of depth information limits their ability to estimate the full 3-DoF relative translation.

More recently, some works [76]–[78] have explored pose estimation using a single RGB-D reference view to reduce onboarding cost for unseen objects. UNOPose [76] incorporates depth data and proposes a one-reference-based pose estimation framework that constructs an SE(3)-invariant reference representation and adaptively weights correspondences to handle low viewpoint overlap. One2Any [77] further introduces a category-agnostic method for 6-DoF object pose estimation that leverages a reference-query RGB-D pair to generate pose embeddings and decode object coordinates. Any6D [78] estimates both object pose and size from an RGB-D anchor image by leveraging joint object alignment and a render-and-compare strategy. Despite their effectiveness, these methods primarily focus on relative pose estimation between the reference and query views, which is insufficient for robotic manipulation scenarios where absolute object poses in a common coordinate system are required for action execution. In contrast, our work targets single-reference 6-DoF absolute pose estimation under robotic manipulation settings. To this end, we introduce a semi-automated reference acquisition and annotation pipeline,

a single reference view-based point cloud focalization strategy to establish a common coordinate system, and SSMs-based feature extraction networks tailored for the limited geometric and spatial information available from a single view. This problem-driven design enables direct deployment in manipulation pipelines while maintaining scalability to unseen objects.

Discussions: Overall, CAD model-based methods depend on textured CAD models, and manual reference view-based methods require dense reference views, both adding manual effort in real-world applications. Related works such as FoundationPose [61] also employ transformer-based architectures for iterative pose refinement; however, our SSM-based backbone is explicitly designed to model long-range spatial dependencies under severely limited geometric information, which is particularly critical in our single-reference setting. Additionally, relative pose estimation methods are not well-suited for robotic manipulation tasks that require absolute poses for action execution. Hence, this paper seeks to enable unseen object 6-DoF absolute pose estimation with a single reference view, reducing manual overhead and enhancing scalability for robotic applications. Most recently, 3D foundation models such as SAM 3D [79] and VGGT [80] suggest a clear trend toward large-scale, data-driven geometric perception. However, these advances do not diminish the importance of reliable pose estimation; instead, they increase the demand for scalable modules that can provide accurate geometric initialization for annotation bootstrapping and downstream reasoning. In this broader context, our method can also be viewed as a complementary component: a practical and scalable solution for unseen object 6-DoF pose estimation that remains valuable even as 3D foundation models continue to evolve.

III. Methodology

We begin with an overview of the overall task setup and framework, including its input and output (Sec. III-A). We then describe the initialization process, which involves unseen object segmentation from the input RGB-D image (Sec. III-B). Next, we present the proposed point focalization strategy (Sec. III-C), Point and RGB SSMs (Sec. III-D), and point-wise alignment for pose solving (Sec. III-E). Finally, we explain the training procedure and supervision scheme (Sec. III-F).

A. Task Setup and Framework Overview

Overall, our work is problem-driven, aiming to enable scalable 6-DoF absolute pose estimation of unseen objects for robotic manipulation with minimal prior information. To this end, we present a unified system that operates with only a single reference view, where each component is explicitly designed to address the challenges arising from single-reference, manipulation-oriented absolute pose estimation. Specifically, for unseen objects which are not encountered during training, SinRef-6D takes a pair of RGB-D images captured from robotic manipulation viewpoints as *input*: a single reference image and a query image. The reference image is selected only once and annotated with a 6-DoF pose through a semi-automatic manner using our custom-developed pose annotator. The *output* of SinRef-6D is the estimated 6-DoF absolute pose of the unseen object in the arbitrary query image. Algorithm 1 shows the overall pipeline of the proposed framework.

Algorithm 1 Overall Pipeline of SinRef-6D

Input: Reference RGB-D image; Query RGB-D image
Output: Estimated 6-DoF object pose $[R_{final} \mid t_{final}]$
Segment reference and query images to obtain object RGB masks I_r, I_q and corresponding depth masks via SAM-driven similarity matching
Back-project depth masks to get object reference and query point clouds P_r, P_q
Transform P_r into the object coordinate system using known reference pose (obtained via a semi-automated annotator) $[R_r \mid t_r]$: $P_r^o = R_r^\top (P_r - t_r)$
Initialize query pose: $[R_1 \mid t_1] \leftarrow$ identity matrix, average coordinate(P_q)
for $i = 1$ to T **do**
  Transform P_q into the object coordinate system: $P_q^i = R_i^\top (P_q - t_i)$
  Extract point-wise and RGB features:
    $F_r \leftarrow Point\ SSM(P_r^o) \oplus RGB\ SSM(I_r)$
    $F_q^i \leftarrow Point\ SSM(P_q^i) \oplus RGB\ SSM(I_q)$
  Perform point-wise feature alignment:
    $\bar{F}_r, \bar{F}_q^i \leftarrow$ GeoTransformer(F_r, F_q^i)
  Compute point-wise affinity: $A^i = \bar{F}_q^i \otimes \bar{F}_r^\top$
  Estimate 6-DoF object pose via weighted SVD:
    $[R_{i+1}, t_{i+1}] = WSVD(A^i, P_r^o, P_q)$
end for
Return: $[R_{final}, t_{final}] \leftarrow [R_{T+1}, t_{T+1}]$

Figure 2 shows the overall workflow that comprises four main components: (A) *Initialization* segments the unseen object in the input reference and query views. (B) *Points Focalization* focalizes the unseen object in the reference and query views into the object coordinate system using their corresponding poses. (C) *Point & RGB SSMs* employ state space models to extract point-wise reference and query features. (D) *Point-wise Alignment & Pose Solving* derives point-wise alignment relationships using the features extracted in (C) to solve the object pose in the query view. In addition, iterating the process from (B) to (D) allows for further obtaining more accurate point-wise alignment and object pose.

B. Initialization

Notably, randomly sampling arbitrary rendering viewpoints may introduce extreme perspectives (e.g., near top-down views) that deviate significantly from real-world reference acquisition, whereas manually selecting viewpoints for all objects would be labor-intensive and may reduce robustness to viewpoint variations. From a practical real-world application perspective, during training, the synthetic reference view is sampled from a viewpoint range that approximates the robotic manipulation viewpoint while introducing natural perturbations. Importantly, this reference view is *not carefully selected*. To support this claim, we additionally generate reference views using the same rendering protocol as GigaPose [60].

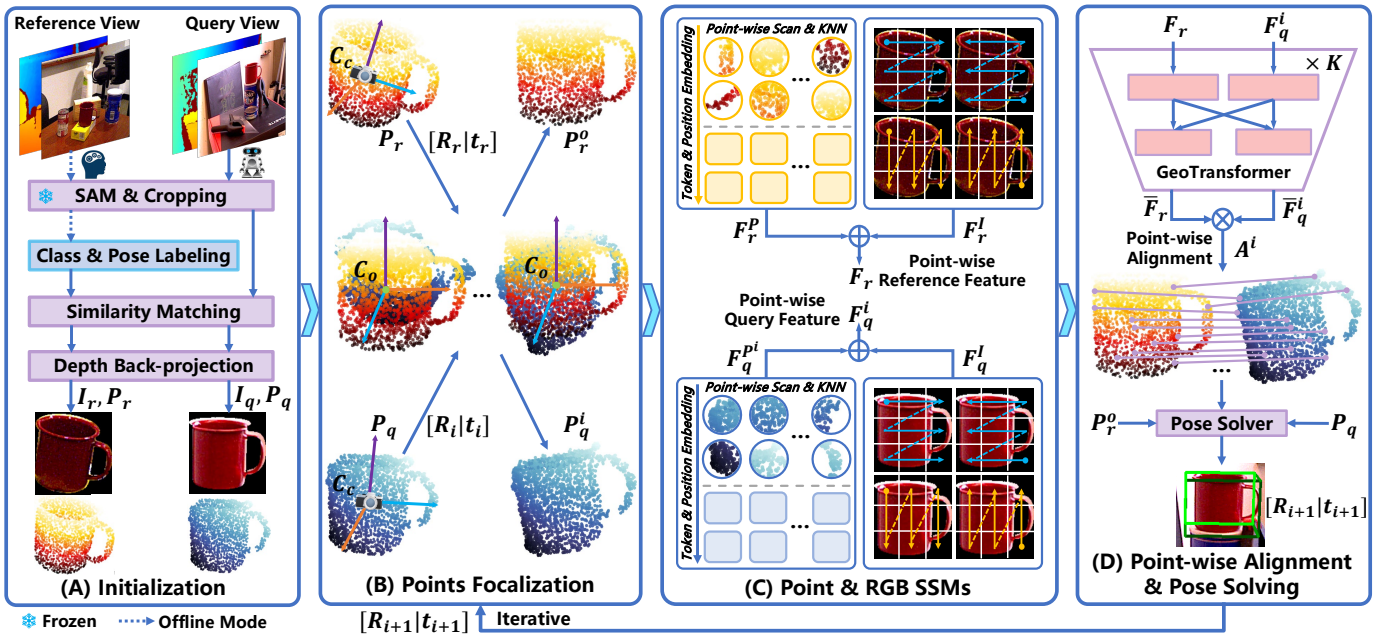

Fig. 2. Our proposed SinRef-6D framework. Given a normal RGB-D reference view of an unseen object, we aim to predict its 6-DoF absolute pose from any query view. SinRef-6D comprises four modules: (A) The reference view is labeled via a semi-automatic annotator, then the RGB-D images of the reference and query views are segmented, and the segmented depth maps are back-projected into point clouds. (B) The corresponding point clouds of the reference and query views are focalized from the camera coordinate system to the object coordinate system. (C) Leveraging the proposed Point and RGB SSMs (details are shown in Fig. 3 and Fig. 4), features are extracted from the focalized point clouds and RGB images, forming point-wise reference and query features. (D) These features are then used to establish point-wise alignment to solve the object pose. Finally, the computed pose is fed back into module (B) to iteratively improve the accuracy of the point-wise alignment, yielding a more precise object pose.

Specifically, we randomly sample one viewpoint from the *50th* to the *120th* in its rendering sequence, which is designed to approximate the robotic manipulation viewpoint. This simulates manual reference view acquisition while introducing natural pose perturbations. During the evaluation in real-world robotic scenarios, we adopt a semi-automatic manner. The reference view for each unseen object is captured by the robot from an occlusion-free manipulation viewpoint and annotated using our custom-developed annotator. The rotation is determined using a calibration board, while the translation and size are manually adjusted through keyboard control (some visualizations are shown in the first row of Fig. 8). For testing on public benchmarks, we adopt both reference view acquisition strategies to align with those used in training.

The pipeline of the initialization process is shown in part (A) of Fig. 2. Since both the reference and query views often contain cluttered backgrounds, we first segment the background. For a fair comparison, we employ Mask R-CNN [81] or zero-shot CNOS [82] with FastSAM to segment the input images, and then back-project the segmented depth maps into point clouds. This results in the segmented RGB images and point clouds for both reference ($I_r, P_r \in \mathbb{R}^{N_r \times 3}$) and query ($I_q, P_q \in \mathbb{R}^{N_q \times 3}$) views, where N_r and N_q denote the number of points in the reference and query point clouds, respectively. Notably, CNOS relies on object CAD models for rendering template images, which contrasts with our CAD model-free setup. Based on this, we also use only our single reference view as the template image for similarity matching in CNOS segmentation (see the first two rows of Tab. IV for details) [82].

C. Points Focalization

Since SinRef-6D aims to iteratively align point clouds for precise object pose solving, our first step is to focalize the reference and query point clouds within a common coordinate system. This focalization facilitates point-wise alignment, ensures geometric consistency during iterative refinement, and inherently decouples pose estimation from category priors, enhancing robustness to unseen objects. Specifically, as the reference point cloud P_r has a pose annotation $[R_r|t_r]$, we can transform it from the camera coordinate system C_c to the object coordinate system C_o as follows:

$$P_r^o = R_r^\top \left(P_r - t_r\right), \tag{1}$$

where t_r and R_r denote the annotated translation and rotation, respectively. $\top$ denotes matrix transpose, P_r^o denotes the reference point cloud in the object coordinate system.

For the query point cloud, we apply the same method to transform it into the object coordinate system as follows:

$$P_q^i = R_i^\top \left(P_q - t_i\right), \tag{2}$$

where t_i and R_i represent the translation and rotation of the object in the i-*th* iteration. P_q^i represents the query point cloud in the object coordinate system after the i-*th* iteration. Since the object pose in the query view is initially unknown, we do not perform rotation transformation during the first points focalization and instead set the translation t_1 to the average coordinate of the object. In subsequent iterations, we use the object pose $[R_{i+1}|t_{i+1}]$ solved in the previous round for coordinate transformation. The overall process is shown in part (B) of Fig. 2.

D. Point & RGB SSMs

Since point-wise alignment relies on rich spatial features, sequential modeling of point clouds and RGB images enables effective long-range spatial encoding, enhancing feature discrimination and geometric consistency for more precise alignment. To handle limited spatial cues and real-time demands, we adopt a simple-yet-efficient design, incorporating lightweight Point and multiscale RGB SSMs for efficient long-range modeling from sparse single-view data with linear complexity. The selective scan structured state space sequence (S6) models [83] represent a class of sequence models that excel in sequence handling. These models extend the earlier S4 model [84], mapping an input sequence $x(t) \in \mathbb{R} \rightarrow y(t) \in \mathbb{R}$ through a latent state $h(t) \in \mathbb{R}^M$ according to the ordinary linear differential equations:

$$\begin{aligned} h'(t) &= \boldsymbol{A}h(t) + \boldsymbol{B}x(t), \\ y(t) &= \boldsymbol{C}h(t) + \boldsymbol{D}x(t), \end{aligned} \tag{3}$$

where $\boldsymbol{A} \in \mathbb{R}^{M\times M}, \boldsymbol{B} \in \mathbb{R}^{M\times 1}, \boldsymbol{C} \in \mathbb{R}^{1\times M}$, and $\boldsymbol{D} \in \mathbb{R}^1$ are weighting parameters. Specifically, the continuous dynamical systems are discretized using the following zero-order hold discretization method in practical computations:

$$\begin{aligned} &h_t = \bar{\boldsymbol{A}}h_{t-1} + \bar{\boldsymbol{B}}x_t, \quad y_t = \boldsymbol{C}h(t), \\ &\bar{\boldsymbol{A}} = \exp(\Delta\boldsymbol{A}), \bar{\boldsymbol{B}} = (\Delta\boldsymbol{A})^{-1}(\exp(\Delta\boldsymbol{A}) - \boldsymbol{I}) \cdot \Delta\boldsymbol{B}, \end{aligned} \tag{4}$$

where Δ denotes the discrete step size. Given that both the weighting parameters and discretization method remain constant over time, S4 models can be considered linear time-invariant systems. S6 model [83] further extends the projection matrices of S4 models to enable a selective scan of the entire input sequence. Specifically, we model the sequences of point clouds and RGB images by designing a Point SSM (see Fig. 3) and an RGB SSM (see Fig. 4) as follows:

$$\begin{aligned} F_r &= Point\ SSM(P_r^o) \oplus RGB\ SSM(I_r), \\ F_q^i &= Point\ SSM(P_q^i) \oplus RGB\ SSM(I_q), \end{aligned} \tag{5}$$

where $\oplus$ represents matrix addition. $F_r \in \mathbb{R}^{N_r\times C}$ and $F_q^i \in \mathbb{R}^{N_q\times C}$ represent the point-wise reference features and the point-wise query features at the *i-th* iteration, respectively. C is the dimension of feature channels.

The sequence modeled by the SSMs refers to an ordered collection of spatial tokens rather than physical time steps. For the RGB branch, the input image is partitioned into patches and flattened into sequences following two fixed raster-scan orders (as shown in part (C) of Fig. 2), which are consistently used during training and inference. This ordering enables the SSM to capture long-range spatial dependencies across image regions without encoding temporal information. For the point cloud branch, due to the intrinsic unordered nature of point sets, the input sequence is constructed by iterating over all points, and the resulting order is neither fixed nor semantically meaningful. For the Point SSM, as shown in Fig. 3, we first perform a point-wise scan and use K-Nearest Neighbor (KNN) to sample a set of points for each scanned point to form a token. Then, we compute all token embeddings and add a position embedding to them. Subsequently, the token embeddings are concatenated and passed into the points state space (PSS) blocks to obtain the point-wise feature

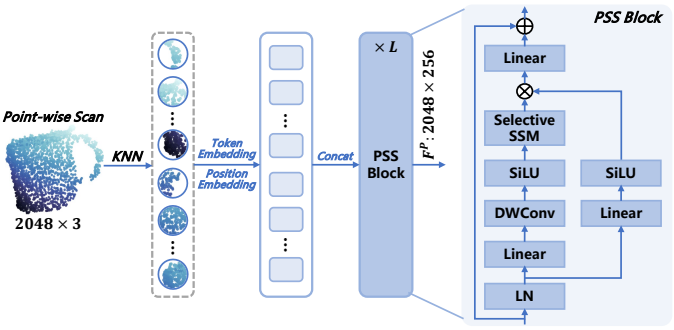

Fig. 3. Architecture of the proposed Point SSM. It takes object point clouds as input and captures point-wise features with fine-grained spatial semantics.

$F^P \in {}^{2048\times 256}$. The details of the selective SSM in PSS blocks can be found in the S6 model [83]. For RGB image feature extraction, we propose an RGB SSM based on the cross-scan manner and multi-scale feature fusion, as depicted in Fig. 4. The architecture consists of four stages, where each stage employs visual state space (VSS) blocks [85] to extract image features at different scales. These multi-scale features are then fused, reshaped, and chosen by using the image mask to obtain the final image feature representation $F^I \in {}^{2048\times 256}$.

Specifically, F^P and F^I denote the generic point-wise and image feature representations, which are instantiated as reference and query features. The complete process is illustrated in part (C) of Fig. 2, where F_r^P and $F_q^{P^i}$ represent the extracted point-wise reference and query features (at the *i-th* iteration), while F_r^I and F_q^I denote the extracted features from the reference and query RGB images.

E. Point-wise Alignment & Pose Solving

Upon acquiring the point-wise reference and query features, our objective is to develop a model with the scalability to handle unseen objects to establish point-wise alignment. This model provides enhanced learnability compared to the direct pose regression model for unseen objects. Specifically, we input F_r and F_q^i into the GeoTransformer [86], where they undergo geometric-aware self-attention and cross-attention, yielding the final point-wise reference and query features $\bar{F}_r$ and $\bar{F}_q^i$. Then, we obtain the point-wise affinity matrix A^i as follows and select the point pairs with the highest similarity for alignment:

$$A^i = \bar{F}_q^i \otimes \bar{F}_r^\top, \tag{6}$$

where $\otimes$ represents matrix multiplication.

The point-wise alignment of the reference and query view point clouds in the object coordinate system is visualized in part (D) of Fig. 2. Once the point-wise alignment relationship is established, we can directly solve the 6-DoF object pose using the weighted singular value decomposition (WSVD) algorithm as follows:

$$[R_{i+1}|t_{i+1}] = WSVD(A^i, P_r^o, P_q), \tag{7}$$

where rotation R_{i+1} and translation t_{i+1} denote the 6-DoF object pose solved from the *i-th* iteration.

Given the considerable pose discrepancies between the initial query and reference point clouds (especially in terms of

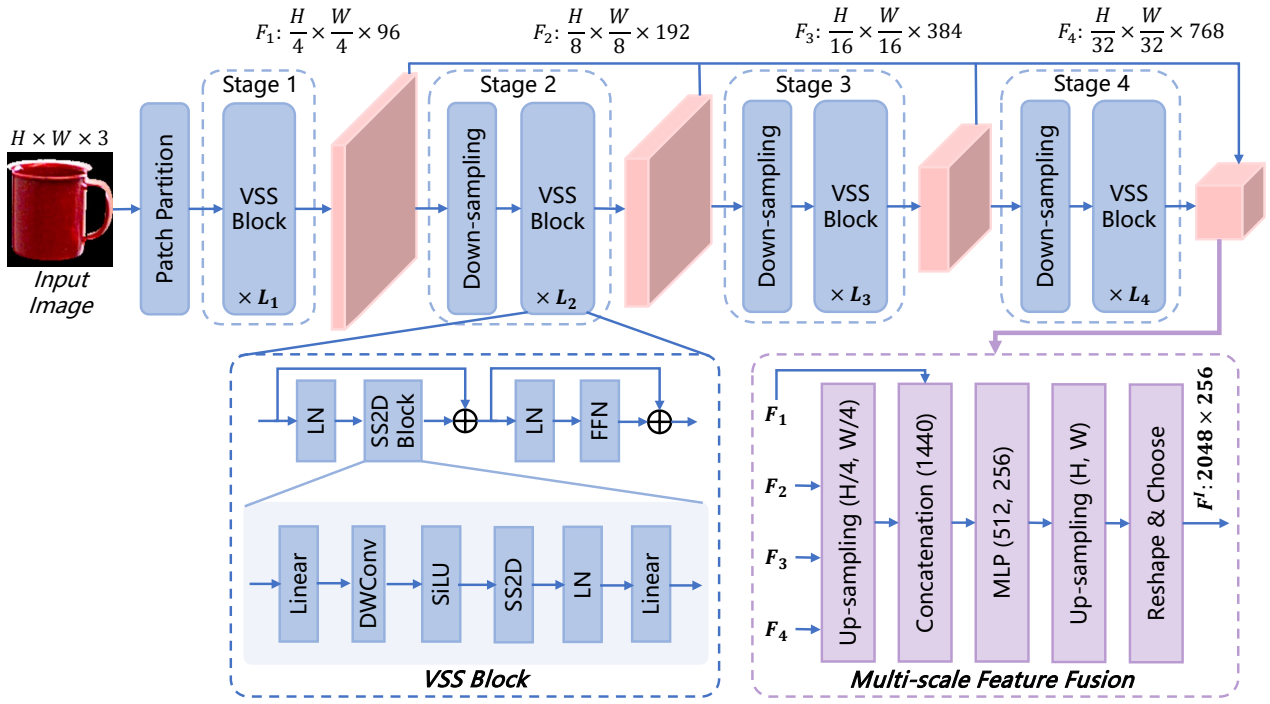

Fig. 4. Detailed architecture of the proposed RGB SSM. It takes segmented RGB mask images as input, where H and W denote the height and width of the image, and outputs rich visual features for subsequent reasoning.

rotation), misaligned point pairs may occur during the alignment process, which will result in inaccurate pose estimation. To mitigate this issue, we introduce an iterative alignment strategy that iteratively performs steps (B) to (D) outlined in Fig. 2. Specifically, the estimated object pose from (D) is fed back into (B) to iterate the focalization of the query point cloud. This iterative strategy facilitates gradual convergence of the query and reference point clouds, ultimately yielding more precise object pose estimation.

F. Training Mode

During practical deployment, we observe that the initial query and reference point clouds have significant pose discrepancies. Subsequently, their pose differences become smaller after each iteration. Using a single GeoTransformer with shared weights during the iterative process (D) in Fig. 2 could adversely affect the accuracy of object pose estimation. Even multi-view transformers have shown that transformer decoder layers can implicitly implement an iterative refinement process for geometric estimation. For example, Stary *et al.* [87] analyzed multi-view transformer architectures and demonstrated that the internal feature state evolves across decoder blocks, progressively refining correspondences and relative camera pose estimates. In such settings, a single network can effectively perform both coarse pose estimation and fine refinement within one unified architecture. However, unseen object pose estimation presents a different regime. Unlike multi-view camera pose estimation, which benefits from strong multi-view geometric constraints and global scene consistency, our setting must handle novel object geometries and limited observations from a single reference view. These factors introduce substantial ambiguity during the initial alignment stage, which we found difficult to reconcile with the objectives of fine-grained pose refinement when using a single backbone.

Motivated by the above observation, we employ two GeoTransformer models with unshared weights to explicitly handle different alignment regimes. The first model is responsible for the initial point-wise alignment, where the pose discrepancy between the reference and query views is typically large and corresponds to a coarse alignment problem. After this stage, the pose discrepancy is significantly reduced, and a second GeoTransformer is used for subsequent iterative refinement under a local alignment regime. This separation allows each model to specialize in a distinct input distribution, improving stability and accuracy during refinement. Detailed rationale for this choice is given in Sec. V-F and Tab. VII.

We employ two GeoTransformer models with unshared weights to explicitly handle different alignment regimes. The first model is responsible for the initial point-wise alignment, where the pose discrepancy between the reference and query views is typically large and corresponds to a coarse alignment problem. After this stage, the pose discrepancy is significantly reduced, and a second GeoTransformer is used for subsequent iterative refinement under a local alignment regime. This separation allows each model to specialize in a distinct input distribution, improving stability and accuracy during refinement.

Since the object pose is solved from the point-wise alignment relationships, we supervise this alignment using the following cross-entropy loss during training:

$$Loss = \sum_{i=1,\dots,K} \Big(CE(A^i, \bar{p}_q) + CE(A^{i^\top}, \bar{p}_r) \Big), \qquad (8)$$

where K denotes the number of point-wise alignment iterations used during training. $CE(\cdot,\cdot)$ represents the cross-entropy loss function. $\bar{p}_q \in \mathbb{R}^{N_q}$ and $\bar{p}_r \in \mathbb{R}^{N_r}$ represent the ground truth for P_q^i and P_r^o. Each element p_q in $\bar{p}_q$, corresponding to the point p_q^i in P_q^i, can be simply obtained using the index of the closest point in P_r^o to p_q^i from the given ground-truth rotation R_{gt} and translation t_{gt}. Note that if the nearest point distance exceeds a specified threshold (set at 15 centimeters in this paper), we discard that point pair. In addition, the elements in $\bar{p}_r$ are obtained in the same manner.

IV. Developed Robotic Grasping System

With the estimated pose, this section presents our developed hardware-software robotic grasping system, which integrates the proposed SinRef-6D into it in real-world settings.

A. Developed Hardware and Software System

The overall system consists of hardware and software components, with their interconnections illustrated in Fig. 1. The hardware includes an Intel RealSense L515 RGB-D camera, a Yaskawa MOTOMAN-MH12 robotic arm, and a DH-PGI-140-80 electric parallel gripper. The RGB-D camera and gripper communicate with the software part via USB, while the robot is connected through Gigabit Ethernet. The camera is mounted in an eye-in-hand configuration on the robotic arm.

The software component of our integrated robotic system consists of three main modules: a calibration module, an unseen object pose estimation module, and a robotic grasping module. These modules are integrated through a unified graphical user interface. The calibration module performs both hand-eye and tool calibration, using the HALCON-based method and a five-point approach, respectively. The pose estimation module supports pose annotation for the reference view as well as pose inference for novel query views. We design a semi-automatic CAD-free pose annotator for unseen objects, where object rotation is solved via a calibration board (as shown in the first two rows of Fig. 8), while translation and size are manually aligned via keyboard control. Overall, our annotator enables 6-DoF pose labeling of a single reference view in *less than one minute* per unseen object. Lastly, the grasping module includes a grasp strategy planner and a servo control interface for executing 6-DoF grasps.

B. 6-DoF Robotic Grasping Workflow

1) Overall Workflow: The complete pipeline for 6-DoF robotic grasping in 3D space can be formulated as:

$$T_{o2t} = T_{e2t} \otimes T_{c2e} \otimes T_{o2c}, \tag{9}$$

where $\otimes$ denotes matrix multiplication. T_{o2t} represents the transformation from the object coordinate to the tool coordinate systems, which is the target transformation required for executing a robotic grasp. The term T_{e2t} denotes the transformation from the robot end coordinate to the end-effector (tool) coordinate systems, obtained via tool calibration. T_{c2e} is the transformation from the camera coordinate to the robot end coordinate systems, computed through hand-eye calibration. Finally, T_{o2c} refers to the transformation from the object coordinate to the camera coordinate systems, which is *the most critical component* and is estimated by the proposed unseen object pose estimation method.

2) Multi-object Grasping Strategy: With the estimated 6-DoF object pose, we employ a lightweight grasping strategy to enable continuous multi-object robotic grasping [88]. Specifically, we adopt a depth-based sequential grasping strategy, where objects are ordered by the depth (relative to the camera) of their center points. For each object, the estimated 3-DoF translation determines the target grasp point, *i.e.*, the position to which the end-effector center is moved, the grasping direction is defined by a vector from its closest visible point projected onto the z-axis to the center point of the object coordinate system, while the gripper closes along the x-axis. For safe execution, the grasp point is translated upward by 2 cm relative to the tabletop. When the angle between the estimated object z-axis and the inward normal of the tabletop falls outside the range of $[20°, 60°]$, the grasp orientation is clamped to $30°$ to ensure stable and collision-free grasping. This simple grasping strategy proves effective for both symmetric and asymmetric objects, despite existing axis ambiguities.

V. Experiments

We first introduce the benchmarks and evaluation metrics (Sec. V-A), followed by the implementation details (Sec. V-B). We then compare SinRef-6D with both manual reference view-based and CAD model-based methods on these real-world benchmarks to validate its superior performance (Sec. V-C and Sec. V-D). Next, we evaluate the effectiveness of our approach in real-world robotic grasping scenarios by deploying it on our integrated hardware-software robotic system to perform grasping tasks (Sec. V-E). Finally, we present comprehensive ablation studies to analyze the contributions of key components, the influence of point cloud alignment iterations, and the effect of random reference view selection (Sec. V-F).

A. Datasets and Evaluation Metrics

Datasets: We conduct extensive experiments on six benchmark datasets (LineMod [89], LM-O [90], TUD-L [91], IC-BIN [92], HB [93], and YCB-V [94]) and real-world robotic scenes. For a fair comparison, we follow the BOP Challenge setting [91] to train on the synthetic dataset generated by MegaPose [58] using the ShapeNet-Objects [95] and Google-Scanned-Objects [96] datasets. This training dataset comprises $\sim$2 million images from $\sim$50K objects.

Evaluation Metrics: 1) Recall of the average point distance (ADD) that is less than 10% of the object diameter (ADD-0.1d) [97]. 2) Area under the curve (AUC) of ADD [94]; 3) BOP metric: Average Recall (AR) of the visible surface discrepancy (VSD), maximum symmetry-aware surface distance (MSSD), and maximum symmetry-aware projection distance (MSPD) metrics [91]. Specifically, we first perform a quantitative comparison using the ADD-0.1d and AUC of ADD metrics for each instance in the LineMod [89] and YCB-V [94] datasets, respectively, aligning with manual reference view-based methods [62]–[64], [67]–[69], [98]. Subsequently,

TABLE I
COMPARISON OF SINREF-6D WITH OTHER MANUAL REFERENCE VIEW-BASED METHODS ON THE LINEMOD DATASET [89], EVALUATED USING THE ADD-0.1D METRIC. "REF." AND "RECON." MEAN "REFERENCE" AND "RECONSTRUCTION". † REPRESENTS GEN6D [69] WITHOUT FINE-TUNING. ^ INDICATES THAT THE REFERENCE VIEW IS MANUALLY SELECTED FROM THE CORRESPONDING DATASET TO APPROXIMATE THE ROBOTIC MANIPULATION VIEWPOINT DURING BOTH TRAINING AND TESTING.

Method	Input	Ref. view	Recon. -free	Object													Mean (%)↑
				ape	benchwise	cam	can	cat	driller	duck	eggbox	glue	holepuncher	iron	lamp	phone	
Gen6D [69]	RGB	200	✓	-	77.0	66.1	-	60.7	67.4	40.5	95.7	87.2	-	-	-	-	-
Gen6D† [69]	RGB	200	✓	-	62.1	45.6	-	40.9	48.8	16.2	-	-	-	-	-	-	-
OnePose [63]	RGB	200	×	11.8	92.6	88.1	77.2	47.9	74.5	34.2	71.3	37.5	54.9	89.2	87.6	60.6	63.6
OnePose++ [64]	RGB	200	×	31.2	97.3	88.0	89.8	70.4	92.5	42.3	99.7	48.0	69.7	**97.4**	**97.8**	76.0	76.9
LatentFusion [68]	RGB-D	16	×	**88.0**	92.4	74.4	88.8	94.5	91.7	68.1	96.3	94.9	82.1	74.6	94.7	91.5	87.1
FS6D [62]	RGB-D	16	✓	74.0	86.0	**88.5**	86.0	**98.5**	81.0	**68.5**	**100.0**	99.5	**97.0**	92.5	85.0	**99.0**	88.9
Oryon [47]	RGB-D	1	✓	1.2	1.3	3.9	0.8	12.7	8.5	0.8	63.2	18.4	1.6	0.6	2.9	11.7	9.8
SinRef-6D (Ours)	RGB-D	1	✓	85.7	**99.3**	73.2	98.3	93.0	**98.7**	66.6	98.5	99.1	74.6	90.9	97.6	97.4	90.3
SinRef-6D^ (Ours)	RGB-D	1	✓	86.3	99.1	74.7	**98.5**	94.5	**98.7**	68.1	98.7	99.5	75.5	92.5	97.0	97.8	**90.8**

TABLE II
COMPARISON OF SINREF-6D WITH OTHER MANUAL REFERENCE VIEW-BASED METHODS ON THE COMPLETE YCB-V DATASET [94], EVALUATED USING THE AUC OF ADD METRIC.

Method	PREDATOR [98]	LoFTR [67]	FS6D-DPM [62]	Ours
Reference view	16	16	16	1
002_master_chef_can	17.4	**50.6**	36.8	44.3
003_cracker_box	8.3	25.5	24.5	**34.4**
004_sugar_box	15.3	13.4	43.9	**83.9**
005_tomato_soup_can	44.4	52.9	**54.2**	53.7
006_mustard_bottle	5.0	59.0	71.1	**79.9**
007_tuna_fish_can	34.2	**55.7**	53.9	53.8
008_pudding_box	24.2	68.1	**79.6**	44.3
009_gelatin_box	37.5	45.2	32.1	**94.6**
010_potted_meat_can	20.9	45.1	**54.9**	25.5
011_banana	9.9	1.6	**69.1**	65.0
019_pitcher_base	18.1	22.3	40.4	**88.2**
021_bleach_cleanser	48.1	16.7	44.1	**72.9**
024_bowl	17.4	1.4	0.9	**31.7**
025_mug	29.5	23.6	39.2	**77.7**
035_power_drill	12.3	1.3	19.8	**53.7**
036_wood_block	10.0	1.4	**27.9**	0.7
037_scissors	25.0	14.6	27.7	**51.2**
040_large_marker	38.9	8.4	74.2	**76.2**
051_large_clamp	34.4	11.2	**34.7**	21.4
052_extra_large_clamp	**24.1**	1.8	10.1	0.4
061_foam_brick	35.5	31.4	45.8	**56.3**
MEAN	24.3	26.2	42.1	**52.8**

we evaluate our results against CAD model-based methods [51], [53], [58], [60] on five BOP datasets [91], utilizing the BOP metric for a comprehensive comparison.

B. Implementation Details

The initial resolution of the input RGB images is 640×480, which are resized to 224×224 after detection and segmentation. Both the reference and query point clouds contain 2048 points (N_r and N_q). The point-wise feature dimension C is set to 256. We use the Adam optimizer for model training

TABLE III
COMPARISON WITH FS6D [62] AND FOUNDATIONPOSE [61] UNDER SINGLE-REFERENCE SETTING ON LINEMOD AND YCB-V DATASETS, EVALUATED ON ADD-0.1D AND AUC OF ADD METRICS, RESPECTIVELY.

Method	Ref. view	LineMod [89]	YCB-V [94]
FS6D [62]	1	77.5	34.7
FoundationPose [61]	1	87.9	47.5
SinRef-6D (Ours)	1	**90.3**	**52.8**

with a batch size of 6, over a total of 2.4 million batches. The learning rate is adjusted using the WarmupCosineLR scheduler, starting from 0 and rapidly increasing to 0.001 during the first 1000 batches, then gradually decreasing until the end of training. The training takes ∼1 week on our workstation. All experiments are conducted on a single GeForce RTX 4090 GPU with an Intel Xeon Gold 6138 CPU.

C. Quantitative Comparisons with SOTA Methods

1) Comparison with Manual Reference View-based Methods: Table I presents a detailed performance comparison of SinRef-6D with other manual reference view-based methods [47], [62]–[64], [68], [69] on the LineMod dataset [89] using the ADD-0.1d metric. For a fair comparison with Oryon [47], we adopt the results reported in One2Any [77], which directly evaluate Oryon using its pretrained model under the original setting including language input.

We evaluate SinRef-6D using two reference view selection strategies, as described in Sec. III-B. *In the first setting*, we follow the same rendering protocol as GigaPose [60], where a single reference view is randomly sampled from the *50th* to the *120th* viewpoint in its rendering sequence. Under this setting, SinRef-6D achieves a mean accuracy of 90.3%, outperforming OnePose [63], OnePose++ [64], and Oryon [47], and achieving comparable accuracy to LatentFusion [68] and FS6D [62]. It is worth noting that all the comparison methods rely on dense reference views. Specifically, Gen6D [69], OnePose [63], and OnePose++ [64] require 200 reference views, while LatentFusion [68] and FS6D [62], which utilize RGB-D inputs like SinRef-6D, still need 16 reference views. Moreover,

TABLE IV
COMPARISON OF SINREF-6D WITH CAD MODEL-BASED METHODS ON THE LM-O [90], TUD-L [91], IC-BIN [92], HB [93], AND YCB-V [94] DATASETS. WE LEVERAGE THE BOP METRIC AND THE MEAN TIME ACROSS ALL DATASETS FOR EVALUATION. ^ DENOTES THAT THE REFERENCE VIEW IS MANUALLY SELECTED FROM THE CORRESPONDING DATASET TO APPROXIMATE THE ROBOTIC MANIPULATION VIEWPOINT DURING TRAINING AND TESTING. * DENOTES USING THE POSE REFINEMENT METHOD OF MEGAPOSE [58]. † MEANS THAT DIRECTLY TEST SAM-6D [53] USES A SINGLE REFERENCE VIEW. ‡ MEANS THAT RETRAIN AND THEN TEST SAM-6D [53] USES A SINGLE REFERENCE VIEW, WITH OTHER SETTINGS UNCHANGED.

Method	Input	CAD model -free	Detection / Segmentation	Dataset					Mean (%)↑	Time (s)↓
				LM-O	TUD-L	IC-BIN	HB	YCB-V		
SinRef-6D (Ours)	RGB-D	✓	Single Ref.-based CNOS	48.4	62.5	31.6	50.7	56.9	50.0	0.7
SinRef-6D^ (Ours)	RGB-D	✓		51.2	65.3	32.7	52.9	58.4	52.1	0.7
MegaPose [58]	RGB	×	Mask R-CNN [81]	18.7	20.5	15.3	18.6	13.9	17.4	25.6
MegaPose* [58]	RGB	×		53.7	58.4	43.6	72.9	60.4	57.8	-
MegaPose* [58]	RGB-D	×		58.3	71.2	37.1	**75.7**	63.3	61.1	93.3
ZeroPose [51]	RGB-D	×		26.1	61.1	24.7	38.2	29.5	35.9	-
ZeroPose* [51]	RGB-D	×		56.2	87.2	41.8	68.2	58.4	62.4	-
SAM-6D† [53]	RGB-D	×		12.9	37.9	11.2	25.2	22.4	21.9	**0.3**
SAM-6D‡ [53]	RGB-D	×		53.7	38.4	26.3	53.2	60.1	46.3	**0.3**
SinRef-6D (Ours)	RGB-D	✓		61.8	88.9	44.0	63.3	65.1	64.6	0.4
SinRef-6D^ (Ours)	RGB-D	✓		**62.0**	**90.4**	**44.5**	64.0	**65.9**	**65.4**	0.4
MegaPose [58]	RGB	×	CNOS (FastSAM) [82]	22.9	25.8	15.2	25.1	28.1	23.4	16.6
MegaPose* [58]	RGB	×		49.9	65.3	36.7	65.4	60.1	55.5	33.9
ZeroPose* [51]	RGB-D	×		53.8	**83.5**	39.2	65.3	65.3	61.4	17.6
GigaPose [60]	RGB	×		29.6	30.0	22.3	34.1	27.8	28.8	**0.4**
GigaPose* [60]	RGB	×		**59.8**	63.1	**47.3**	**72.2**	**66.1**	**61.7**	8.5
SAM-6D† [53]	RGB-D	×		10.4	30.1	9.4	29.0	21.8	20.1	1.2
SAM-6D‡ [53]	RGB-D	×		53.9	32.2	25.0	55.4	59.1	45.1	1.1
SinRef-6D (Ours)	RGB-D	✓		56.5	77.4	35.9	61.0	62.2	58.6	1.5
SinRef-6D^ (Ours)	RGB-D	✓		56.2	78.6	37.1	62.1	62.9	59.4	1.5

Gen6D [69] depends on a template matching process, while OnePose [63], OnePose++ [64], and LatentFusion [68] require reconstructing the 3D object representation prior to pose estimation. These additional requirements will increase model complexity and reduce overall efficiency. *In the second setting*, we manually select an occlusion-free reference view from the training/testing sets that closely approximates the robotic manipulation viewpoint. This strategy leads to a slightly higher accuracy of 90.8%, which we attribute to the improved viewpoint alignment between the reference and query images. The experimental results of these two reference view selection settings collectively demonstrate that SinRef-6D does *not rely on a carefully selected* reference view and *remains robust to variations* in the selection of the reference view.

Additionally, we evaluate SinRef-6D on the complete YCB-V [94] dataset using the AUC of ADD metric, with per-object accuracy results summarized in Tab. II. These quantitative results further validate the advantages of SinRef-6D. Also, we note that objects with flat or geometrically complex structures, such as the extra large clamp, tend to yield lower accuracy, as they are difficult to perceive accurately from a single reference view. Furthermore, we set the number of reference views to 1 for both FS6D [62] and FoundationPose [61] and experiment with the LineMod and YCB-V datasets, *using their pre-trained models and keeping other settings unchanged*. In particular, for FoundationPose, we adopt its ablation setting by reconstructing the unseen object 3D model from a single reference, while keeping the downstream hypothesis generation and pose selection components fixed. Notably, FoundationPose utilizes a higher-quality synthetic dataset generated with diffusion model and requires time-consuming 3D reconstruction (making it ∼10 times slower than SinRef-6D). The results, as shown in Tab. III, further reinforce the advantages of SinRef-6D in the single-reference setting.

Overall, the above experimental results demonstrate the effectiveness of the proposed SinRef-6D task setup and framework, achieving competitive 6-DoF pose estimation accuracy for unseen objects using only a single reference view, without relying on the reconstruction of a 3D object representation.

2) Comparison with CAD Model-based Methods: Table IV presents performance comparisons of SinRef-6D with CAD model-based methods [51], [53], [58], [60] on five popular datasets using the BOP metric (AR metric of VSD, MSSD, and MSPD). We also conduct experiments under two reference view selection strategies, as outlined in Sec. III-B. The first strategy adopts the same rendering manner as GigaPose [60], and the second strategy involves manually choosing an occlusion-free view from the training/testing set that closely aligns with the robotic manipulation view. In addition, we further evaluate SinRef-6D under three different

TABLE V
COMPARISON WITH OTHER SINGLE-REFERENCE METHODS.

Method	LM	LM-O	TUD-L
	ADD-0.1d	AR	
One2Any [77]	52.6	-	-
Any6D [78]	-	28.6	-
UNOPose [76]	-	56.0	67.1
Ours	**90.3**	**56.5**	**77.4**

detection/segmentation methods to assess its robustness across perception inputs. Specifically, we first show the experimental results using only our single reference view as the template image for CNOS segmentation [82] in the first two rows.

When Mask R-CNN [81] is used to segment the input images, SinRef-6D achieves 64.6% AR across five evaluation datasets, outperforming MegaPose [58] and ZeroPose [51]. Remarkably, this accuracy even exceeds the performance of these two methods after they leverage the refinement method introduced by MegaPose [58]. Moreover, we evaluate SAM-6D [53] also with a single reference view under two training settings while keeping other settings (e.g., coarse-to-fine refine) unchanged. Next, when we employ zero-shot CNOS [82] for segmentation, SinRef-6D also achieves a competitive AR of 58.6% across the five datasets. We also note an increase in inference time. This is primarily because CNOS often segments multiple instances for the same object, resulting in repeated pose estimations. Note that the comparison methods all rely on textured CAD models, requiring specialized equipment for acquisition. Additionally, the refinement process of MegaPose [58] is slow and also relies on object CAD models (that is why we do not use it for refinement). In contrast, SinRef-6D is CAD model- and refinement-free, offering enhanced scalability and efficiency. In general, SinRef-6D demonstrates performance on par with CAD model-based methods, while operating in a CAD model-free setup, showcasing its effectiveness and scalability.

Similarly, across all three detection/segmentation methods, we observe that manually selected reference views consistently yield slightly higher accuracy than randomly rendered ones. We argue that this is due to their closer alignment with the robotic manipulation viewpoint and reduced ambiguity. Overall, these results not only highlight the robustness of SinRef-6D to different detection/segmentation pipelines, but also further confirm its resilience to variations in reference view selection.

3) Comparison with Single-Reference Methods: Table V shows explicit quantitative comparisons with the three most closely related single-reference methods discussed in Sec. II-C (One2Any [77], Any6D [78], and UNOPose [76]). To ensure a fair comparison, all methods use the same segmentation and evaluation protocols: ground-truth segmentation on the LM dataset, and CNOS segmentation [82] on the LM-O and TUD-L datasets; ADD-0.1d is used for LM, while AR is used for LM-O and TUD-L, consistent with prior work. The results of One2Any and Any6D are taken directly from their original papers, while UNOPose is evaluated using its pretrained model under CNOS segmentation (same as SinRef-6D). These results show that SinRef-6D achieves more accurate pose estimation performance.

Fig. 5. The qualitative comparison results on the LineMod dataset [89] are presented, visualizing the outputs of Gen6D [69], our SinRef-6D, and ground truth from top to bottom.

D. Qualitative Analysis

1) Comparison with Manual Reference View-based Methods: The comparison between SinRef-6D and Gen6D [69] on the LineMod dataset [89] is presented in Fig. 5. These experimental results highlight the superior performance of our method, which can perform unseen object pose estimation in cluttered scenes. Specifically, Gen6D [69] relies on dense reference views and template matching to estimate object poses, which requires full coverage of the reference view angles. When the number of reference views is limited or of low quality, pose estimation errors will significantly increase. In contrast, SinRef-6D abandons template matching and instead leverages iterative alignment between the reference and query views, enabling effective pose estimation of unseen objects using only a single reference view.

2) Comparison with CAD Model-based Methods: Figure 6 compares SinRef-6D with MegaPose [58] and ZeroPose [51] across five evaluation datasets, all of which use RGB-D input and CNOS [82] for segmentation. These datasets cover a diverse range of scenes and unseen objects. The experimental results further demonstrate that our method outperforms the comparison methods in terms of robustness, effectively estimating the 6-DoF pose of unseen objects even in challenging scenes with occlusions and clutter. Notably, both MegaPose [58] and ZeroPose [51] require textured CAD models of these unseen objects, with MegaPose [58] also depending on the time-consuming render-and-compare process. In contrast, SinRef-6D is a simple-yet-effective CAD-free method, offering greater scalability.

3) Failure Cases Analysis: Since SinRef-6D only uses a single reference view captured from an oblique angle, pose estimation accuracy may decrease when the query view does not adequately capture the object's geometric features, such as in top-down views. Furthermore, for objects with incomplete depth information, like reflective metals or transparent materials, establishing accurate point-wise alignments becomes

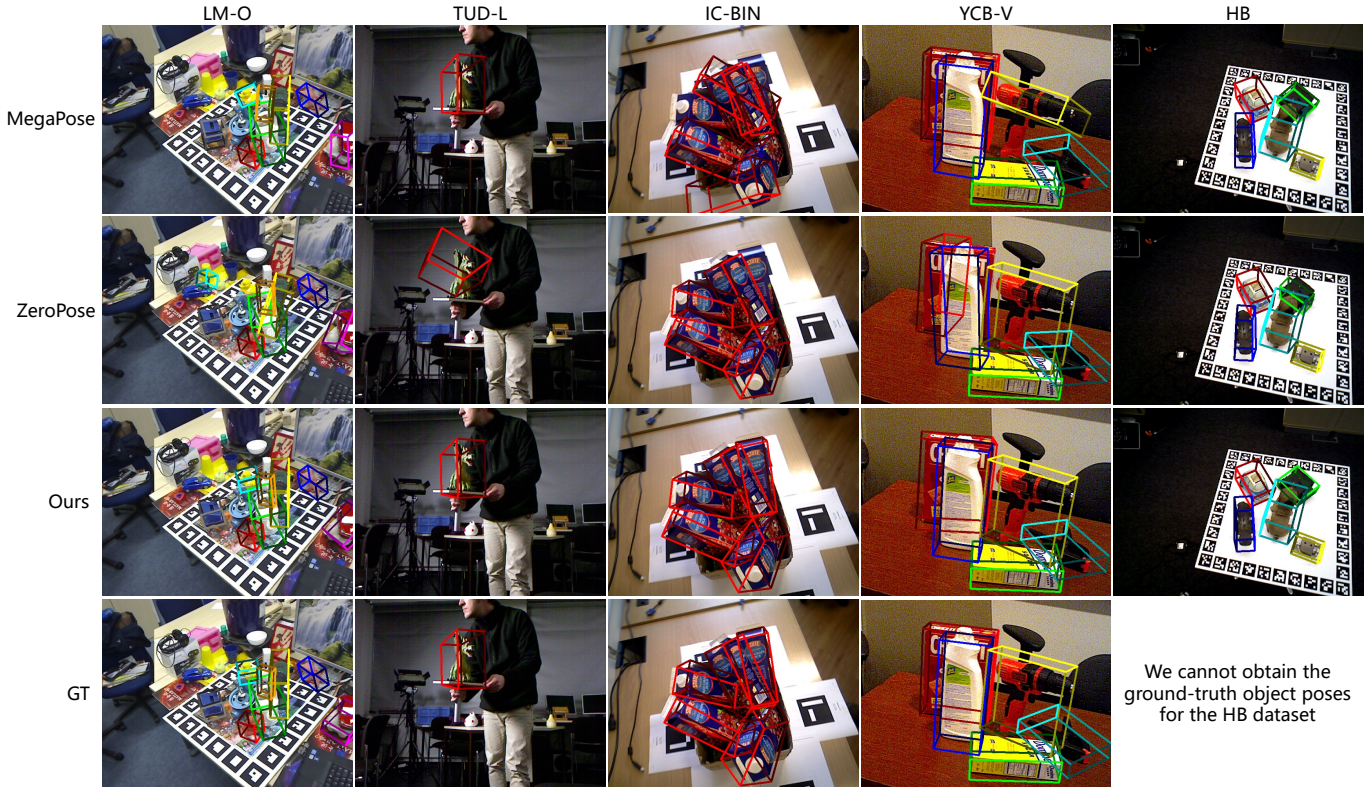

Fig. 6. The qualitative comparison results on the LM-O [90], TUD-L [91], IC-BIN [92], and YCB-V [94] datasets. We visualize the results of MegaPose [58], ZeroPose [51], our SinRef-6D, and ground truth from top to bottom. Note that we cannot obtain the ground-truth poses for the HB dataset [93], as its evaluation is conducted on the official BOP Challenge [91]. Additional qualitative results are presented at our project homepage.

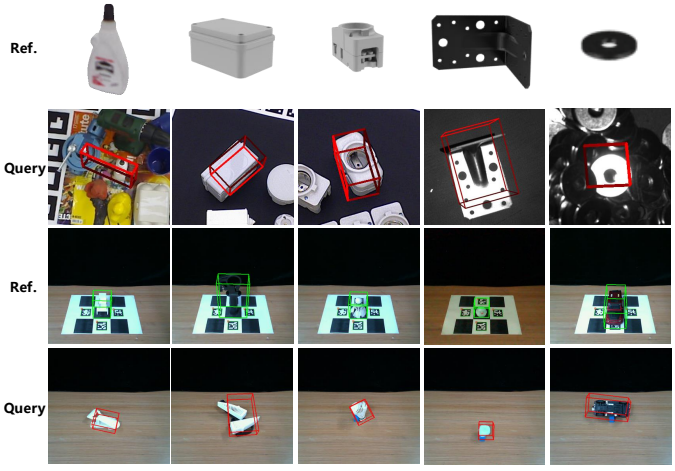

Fig. 7. Failure cases on public datasets (top two rows) and real world (bottom two rows). The top two rows show single reference views (randomly selected RGB-D images as described in Sec. III-B) and the estimated object pose in query views. As observed, the accuracy of SinRef-6D decreases when the query view is a top-down view or the object is a reflective metal. The bottom two rows illustrate the labeled reference views and representative failure cases arising from particularly challenging scenarios, such as severe occlusion/self-occlusion and large viewpoint gaps between the reference and query views.

challenging, leading to a decrease in pose estimation accuracy. The visualization of some failure cases is shown in Fig. 7. As a result, we do not evaluate the T-LESS [99] (includes many top-down views) and ITODD [100] (features top-down views and metallic objects) datasets. Moreover, we provide an explicit analysis of real-world failure cases and observe a performance degradation under non-planar object placements, large reference-query viewpoint discrepancies, and severe occlusions, as illustrated in the last two rows of Fig. 7. Our future work will focus on enhancing the robustness of SinRef-6D in such challenging scenes and objects.

E. Real-world Robotic Grasping

1) Qualitative validation on real-world robotic grasping scenarios: To evaluate the effectiveness of SinRef-6D in estimating the 6-DoF poses of unseen objects in real-world robotic grasping scenarios, we mount an Intel RealSense L515 RGB-D camera on a robotic arm and conduct experiments across four representative scenes: normal, clutter, occlusion, and low light. Qualitative results are shown in Fig. 8. Specifically, the robot first captures a reference image of an unseen object from its manipulation viewpoint, which is then annotated using our custom-developed semi-automatic annotator within one minute (top two rows). This reference is used as prior knowledge for zero-shot unseen object 6-DoF pose estimation from other viewpoints. We place multiple unseen objects in some challenging scenes and evaluate the robustness of SinRef-6D under realistic grasping conditions (middle two rows). Since the robustness to non-planar object placements is critical for real-world robotic deployments, we substantially introduce some objects placed in inverted or non-planar configurations (bottom two rows). In general, these experiments validate the

Fig. 8. Qualitative results on real-world robotic grasping scenarios. *Top two rows*: Single reference view capture and annotation from the robotic manipulation viewpoint for several unseen objects. *Middle two rows*: Unseen object 6-DoF pose estimation in planar aligned query views. *Bottom two rows*: Unseen object 6-DoF pose estimation in non-planar aligned query views. These include some challenging scenes commonly encountered in robotic grasping, including clutter, occlusion, low light, and dark conditions.

effectiveness of SinRef-6D and demonstrate its potential for downstream robotic grasping tasks.

2) Real-world robotic grasping: To evaluate the applicability of SinRef-6D to unseen object 6-DoF robotic grasping and to validate the effectiveness of the developed hardware-software robotic system (described in Sec. IV), we integrate SinRef-6D on the system to perform real-world grasping experiments. Due to the mechanical limitations of the gripper, we select some graspable objects placed in randomly cluttered scenes for testing. Some representative qualitative results are shown in Fig. 9. Specifically, we first estimate the 6-DoF poses of these unseen objects (top-left) under both normal and low light scenarios, and then execute sequential robotic grasping based on the estimated poses. To further assess robustness under more challenging conditions, we additionally include grasping demonstrations involving geometrically irregular unseen objects placed in non-planar configurations and under large reference-query viewpoint discrepancies, as shown in the bottom two rows of Fig. 9. In these experiments, object poses are randomly placed, while scene clutter is generated by randomly moving non-target, non-rigid items such as blankets and cables. Quantitatively, we conduct a total of 200 real-world grasping trials, evenly divided between planar placements and challenging non-planar object configurations (100 trials each). Each scene contains two or three randomly placed unseen objects, achieving overall success rates of 85%

Fig. 9. Real-world robotic grasping visualizations. The top two rows and the bottom two rows represent planar aligned and non-planar aligned object grasping, respectively. These scenes contain clutter and low light. The top-left corner in each image shows the estimated 6-DoF poses of the unseen objects from the robotic manipulation viewpoint. The complete grasping video can be seen through our project homepage.

and 74%, respectively, where a trial is considered successful only if all objects in the scene are successfully grasped. In summary, these real-world experiments not only demonstrate the effectiveness of the proposed task formulation and method for unseen object 6-DoF robotic grasping, but also validate the robustness and practicality of our integrated robotic system.

F. Ablation Study

We investigate the effectiveness of several main components within SinRef-6D, as well as the impact of the number of iterations for point-wise alignment during both training and inference. Additionally, we further conduct multiple experiments with randomly selected reference views to demonstrate the robustness of our method to such variations.

1) Effectiveness of Main Components: We conduct a thorough ablation study on the main components in SinRef-6D to verify their effectiveness. Specifically, we first remove the RGB image component from SinRef-6D, meaning that RGB images are not used during either training or inference. The experimental results (row A of Tab. VI) show that RGB images play a crucial role in enhancing the accuracy of point-wise alignment. Next, we eliminate the point cloud focalization process, directly feeding the reference and query point clouds P_r and P_q from part (B) of Fig. 2 into the model. Therefore, the iterative training of the GeoTransformer is also removed. The experimental results (row B of Tab. VI) reveal a substantial drop in performance, underscoring the significance of the focalization step. We attribute this decline to the larger numerical discrepancies that may arise between the reference and query point clouds without the focalization process. In addition, we only train a single GeoTransformer for point cloud iterative alignment (row C of Tab. VI). Although

TABLE VI
ABLATION OF SINREF-6D COMPONENTS ON THE YCB-V DATASET [94]. PARAM. MEANS TOTAL MODEL PARAMETERS.

Row	Method	AR ↑	Param. (M) ↓
A	w/o RGB	39.5	138.8
B	w/o Points Focalization	0.0	643.6
C	only one GeoTransformer	36.5	643.6
D	RGB SSM → DINOv2 [101]	56.9	1238.8
E	RGB SSM → ViT [102]	52.8	976.7
F	Point SSM → PT [103]	60.9	708.6
G	Full Model	62.2	691.8

the parameter count is slightly reduced, the performance degrades significantly. This is because a single alignment model struggles to simultaneously specialize in handling large initial pose discrepancies and accurately refining small residual errors. Further, we replace the proposed RGB SSM with DINOv2 [101] (a large vision foundation model) and Vision Transformer [102], and also replace Point SSM with Point Transformer [103]. The corresponding results (rows D, E, and F of Tab. VI) confirm the effectiveness of SSMs in improving the accuracy of point-wise alignment while reducing the model parameter count.

2) Impact of the Number of Iterations for Point-wise Alignment: To achieve precise object pose estimation, it is essential to perform point-wise iterative alignment of the reference and query point clouds. Identifying an optimal balance between accuracy and computational efficiency requires careful selection of iteration counts during training and inference. The training iterations refer to the number of times the GeoTransformer weights are updated during training (*i.e.*, K in Eq. (8)), while the inference iterations indicate the number of GeoTransformer iterations during inference. If the inference iterations exceed the training iterations, the difference represents how many times the last GeoTransformer weights from training are repeatedly applied. The results, summarized in Tab. VII, reveal a trade-off among accuracy, speed, and model complexity. As shown in the first row, using a single GeoTransformer for all alignment iterations leads to degraded performance in later refinement stages. We attribute this to a distribution mismatch between the training data, which is dominated by large pose discrepancies, and the inputs encountered during refinement, where the alignment error is much smaller. This observation motivates the use of a separate refinement model specialized for local alignment. Alternative designs, such as conditioning a single alignment model on the iteration index or estimated pose discrepancy, or adopting mixture-of-experts or curriculum learning strategies to handle different alignment regimes, could potentially address this dynamic range challenge and be explored in future work. Based on these ablations, we select 2 training and 3 inference iterations as the optimal configuration to achieve the best balance.

3) Impact of Random Reference View Selection: We investigate the effect of our random reference view selection method (see Sec. III-B for details) on the performance of the object 6-DoF pose estimation. Unlike the reference view

TABLE VII
ABLATION ON THE NUMBER OF ITERATIONS FOR POINT-WISE ALIGNMENT. NOTE THAT THE TIME IS ONLY ON THE YCB-V DATASET [94].

Training iterations	Inference iterations, AR (%) ↑			
	1	2	3	4
1	37.8	35.0	36.5	35.9
2	-	61.7	62.2	62.5
3	-	-	62.2	62.6
Time (s) ↓	0.70	0.99	1.26	1.51

TABLE VIII
ABLATION OF RANDOM REFERENCE VIEW SELECTION, WE REPORT THE MEAN AND VARIANCE OF 20 TIMES EXPERIMENTS ON THE BOP METRIC.

Dataset	Times	Mean	Variance
YCB-V [94]	20	62.10	0.31
LM-O [90]	20	56.58	0.43

variations evaluated in Tabs. I and IV, which compare different reference acquisition strategies, the objective here is to assess the robustness of the proposed model under stochastic reference view selection. Specifically, we conduct 20 tests on both the YCB-V and LM-O datasets, where in each test the reference view is randomly sampled and rendered from the viewpoints *50th* to *120th* defined in GigaPose [60], following the same reference view sampling protocol used during training. Experimental results in Tab. VIII show that the variance of multiple experiments is small, which means that SinRef-6D is robust to random sampled reference views.

VI. CONCLUSION

We proposed SinRef-6D, a simple-yet-effective task setup and framework to tackle the challenges of unseen object 6-DoF pose estimation in existing methods that rely on textured CAD models or dense reference views. SinRef-6D solely requires a single reference view and iteratively establishes point-wise alignment between the reference and query views in the object coordinate system using our proposed SSMs, eliminating the reliance on a CAD model and substantially enhancing the scalability for real-world applications. In addition, we developed a complete hardware-software robotic grasping system tailored to the proposed task setup and framework. This system integrates hand-eye and tool calibration, a semi-automatic single reference annotator, a pre-trained SinRef-6D model, and grasping strategy and control. Extensive experiments on six benchmarks and real-world robotic grasping scenarios demonstrate the superior scalability of SinRef-6D and the effectiveness of our developed robotic system.

Limitation and Future Work: While the proposed framework demonstrates robust real-world pick-and-place performance, the current grasping policy is deliberately simple and does not fully exploit the estimated 6-DoF object pose. Our future work will investigate tighter coupling between pose estimation and downstream manipulation, including pose-conditioned grasp planning and more dexterous manipulation tasks.

REFERENCES

[1] D. Bauer, P. Hönig, J.-B. Weibel, J. García-Rodríguez, M. Vincze *et al.*, "Challenges for monocular 6-d object pose estimation in robotics," *IEEE Transactions on Robotics*, vol. 40, pp. 4065–4084, 2024.

[2] B. Pang, D. Zhai, J. Zhen, L. Wang, and X. Liu, "Fast and accurate 6-d object pose refinement via implicit surface optimization," *IEEE Transactions on Robotics*, vol. 41, pp. 3129–3142, 2025.

[3] J. Liu, W. Sun, H. Yang, Z. Zeng, C. Liu, J. Zheng, X. Liu, H. Rahmani, N. Sebe, and A. Mian, "Deep learning-based object pose estimation: A comprehensive survey," *International Journal of Computer Vision*, vol. 134, no. 81, pp. 1–45, 2026.

[4] H. Chen, T. Kiyokawa, Z. Hu, W. Wan, and K. Harada, "A multi-level similarity approach for single-view object grasping: matching, planning, and fine-tuning," *IEEE Transactions on Robotics*, 2025.

[5] X. Liu, G. Wang, Y. Li, and X. Ji, "Catre: Iterative point clouds alignment for category-level object pose refinement," in *European Conference on Computer Vision*, 2022, pp. 499–516.

[6] K. Chen, R. Cao, S. James, Y. Li, Y.-H. Liu, P. Abbeel, and Q. Dou, "Sim-to-real 6d object pose estimation via iterative self-training for robotic bin picking," in *European Conference on Computer Vision*, 2022, pp. 533–550.

[7] B. Fu, S. K. Leong, X. Lian, and X. Ji, "6d robotic assembly based on rgb-only object pose estimation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2022, pp. 4736–4742.

[8] J. Liu, W. Sun, H. Yang, P. Deng, C. Liu, N. Sebe, H. Rahmani, and A. Mian, "Diff9d: Diffusion-based domain-generalized category-level 9-dof object pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 7, pp. 5520–5537, 2025.

[9] S. Tyree, J. Tremblay, T. To, J. Cheng, T. Mosier, J. Smith, and S. Birchfield, "6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2022, pp. 13 081–13 088.

[10] J. Zhou, Q. Zhu, Y. Wang, M. Feng, J. Liu, J. Huang, and A. Mian, "A state space model for multiobject full 3-d information estimation from rgb-d images," *IEEE Transactions on Cybernetics*, vol. 55, no. 5, pp. 2248–2260, 2025.

[11] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. 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*, 2019, pp. 2642–2651.

[12] H. Li, J. Lin, and K. Jia, "Dcl-net: Deep correspondence learning network for 6d pose estimation," in *European Conference on Computer Vision*, 2022, pp. 369–385.

[13] T. Cao, F. Luo, Y. Fu, W. Zhang, S. Zheng, and C. Xiao, "Dgecn: A depth-guided edge convolutional network for end-to-end 6d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3783–3792.

[14] Y. Wu, M. Zand, A. Etemad, and M. Greenspan, "Vote from the center: 6 dof pose estimation in rgb-d images by radial keypoint voting," in *European Conference on Computer Vision*, 2022, pp. 335–352.

[15] J. Zhou, K. Chen, L. Xu, Q. Dou, and J. Qin, "Deep fusion transformer network with weighted vector-wise keypoints voting for robust 6d object pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 967–13 977.

[16] A. Krishnan, A. Kundu, K.-K. Maninis, J. Hays, and M. Brown, "Ominnocs: A unified nocs dataset and model for 3d lifting of 2d objects," in *European Conference on Computer Vision*, 2024, pp. 127–145.

[17] M. Zhang, T. Wu, T. Wang, T. Wang, Z. Liu, and D. Lin, "Omni6d: Large-vocabulary 3d object dataset for category-level 6d object pose estimation," in *European Conference on Computer Vision*, 2024, pp. 216–232.

[18] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "Pvnet: Pixel-wise voting network for 6dof pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4561–4570.

[19] B. Wen, C. Mitash, B. Ren, and K. E. Bekris, "Se(3)-tracknet: Data-driven 6d pose tracking by calibrating image residuals in synthetic domains," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020, pp. 10 367–10 373.

[20] G. Wang, F. Manhardt, J. Shao, X. Ji, N. Navab, and F. Tombari, "Self6d: Self-supervised monocular 6d object pose estimation," in *European Conference on Computer Vision*, 2020, pp. 108–125.

[21] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 611–16 621.

[22] Y. Di, F. Manhardt, G. Wang, X. Ji, N. Navab, and F. Tombari, "So-pose: Exploiting self-occlusion for direct 6d pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 396–12 405.

[23] L. Xu, H. Qu, Y. Cai, and J. Liu, "6d-diff: A keypoint diffusion framework for 6d object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9676–9686.

[24] Y. He, H. Huang, H. Fan, Q. Chen, and J. 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*, 2021, pp. 3003–3013.

[25] Z. Dang, L. Wang, Y. Guo, and M. Salzmann, "Match normalization: Learning-based point cloud registration for 6d object pose estimation in the real world," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 06, pp. 4489–4503, 2024.

[26] X. Deng, A. Mousavian, Y. Xiang, F. Xia, T. Bretl, and D. Fox, "Poserbpf: A rao–blackwellized particle filter for 6-d object pose tracking," *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1328–1342, 2021.

[27] J. Liu, W. Sun, C. Liu, H. Yang, X. Zhang, and A. Mian, "Mh6d: Multi-hypothesis consistency learning for category-level 6-d object pose estimation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 3, pp. 4820–4833, 2025.

[28] J. Liu, Y. Chen, X. Ye, and X. Qi, "Ist-net: Prior-free category-level pose estimation with implicit space transformation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 978–13 988.

[29] Y. Di, R. Zhang, Z. Lou, F. Manhardt, X. Ji, N. Navab, and F. Tombari, "Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6781–6791.

[30] T. Lee, J. Tremblay, V. Blukis, B. Wen, B.-U. Lee, I. Shin, S. Birchfield, I. S. Kweon, and K.-J. Yoon, "Tta-cope: Test-time adaptation for category-level object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 285–21 295.

[31] L. Zheng, T. H. E. Tse, C. Wang, Y. Sun, H. Chen, A. Leonardis, W. Zhang, and H. J. Chang, "Georef: Geometric alignment across shape variation for category-level object pose refinement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 693–10 703.

[32] H. Jung, S.-C. Wu, P. Ruhkamp *et al.*, "Housecat6d–a large-scale multi-modal category level 6d object pose dataset with household objects in realistic scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 498–22 508.

[33] J. Liu, W. Sun, H. Yang, J. Zheng, Z. Geng, H. Rahmani, and A. Mian, "Monodiff9d: Monocular category-level 9d object pose estimation via diffusion model," in *IEEE International Conference on Robotics and Automation*, 2025.

[34] J. Shi, H. Yang, and L. Carlone, "Optimal and robust category-level perception: Object pose and shape estimation from 2-d and 3-d semantic keypoints," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 4131–4151, 2023.

[35] V. N. Nguyen, Y. Hu, Y. Xiao, M. Salzmann, and V. Lepetit, "Templates for 3d object pose estimation revisited: Generalization to new objects and robustness to occlusions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6771–6780.

[36] I. Shugurov, F. Li, B. Busam, and S. Ilic, "Osop: A multi-stage one shot object pose estimation framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6835–6844.

[37] M. Gou, H. Pan, H.-S. Fang, Z. Liu, C. Lu, and P. Tan, "Unseen object 6d pose estimation: A benchmark and baselines," *arXiv preprint arXiv:2206.11808*, 2022.

[38] F. Hagelskjær and R. L. Haugaard, "Keymatchnet: Zero-shot pose estimation in 3d point clouds by generalized keypoint matching," in *IEEE International Conference on Automation Science and Engineering*, 2024, pp. 870–877.

[39] Z. Fan, P. Pan, P. Wang, Y. Jiang, D. Xu, and Z. Wang, "Pope: 6-dof promptable pose estimation of any object in any scene with one

reference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7771–7781.

[40] C. Zhao, Y. Hu, and M. Salzmann, "Locposenet: Robust location prior for unseen object pose estimation," in *International Conference on 3D Vision*, 2024, pp. 1072–1081.

[41] P. Pan, Z. Fan, B. Y. Feng, P. Wang, C. Li, and Z. Wang, "Learning to estimate 6dof pose from limited data: A few-shot, generalizable approach using rgb images," in *International Conference on 3D Vision*, 2024, pp. 1059–1071.

[42] G. Pitteri, S. Ilic, and V. Lepetit, "Cornet: Generic 3d corners for 6d pose estimation of new objects without retraining," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[43] G. Pitteri, A. Bugeau, S. Ilic, and V. Lepetit, "3d object detection and pose estimation of unseen objects in color images with local surface embeddings," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[44] M. Sundermeyer, M. Durner, E. Y. Puang, Z.-C. Marton, N. Vaskevicius, K. O. Arras, and R. Triebel, "Multi-path learning for object pose estimation across domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 916–13 925.

[45] B. Okorn, Q. Gu, M. Hebert, and D. Held, "Zephyr: Zero-shot pose hypothesis rating," in *IEEE International Conference on Robotics and Automation*, 2021, pp. 14 141–14 148.

[46] J. Wu, Y. Wang, and R. Xiong, "Unseen object pose estimation via registration," in *IEEE International Conference on Real-time Computing and Robotics*, 2021, pp. 974–979.

[47] J. Corsetti, D. Boscaini, C. Oh, A. Cavallaro, and F. Poiesi, "Open-vocabulary object 6d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 071–18 080.

[48] B. Wen, J. Tremblay, V. Blukis, S. Tyree, T. Müller, A. Evans, D. Fox, J. Kautz, and S. Birchfield, "Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 606–617.

[49] Z. He, Q. Li, X. Zhao, J. Wang, H. Shen, S. Zhang, and J. Tan, "Contourpose: Monocular 6-d pose estimation method for reflective textureless metal parts," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 4037–4050, 2023.

[50] R. Talak, L. R. Peng, and L. Carlone, "Certifiable object pose estimation: Foundations, learning models, and self-training," *IEEE Transactions on Robotics*, vol. 39, no. 4, pp. 2805–2824, 2023.

[51] J. Chen, M. Sun, T. Bao, R. Zhao, L. Wu, and Z. He, "Zero-pose: Cad-model-based zero-shot pose estimation," *arXiv preprint arXiv:2305.17934*, 2023.

[52] H. Zhao, S. Wei, D. Shi, W. Tan, Z. Li, Y. Ren, X. Wei, Y. Yang, and S. Pu, "Learning symmetry-aware geometry correspondences for 6d object pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 045–14 054.

[53] J. Lin, L. Liu, D. Lu, and K. Jia, "Sam-6d: Segment anything model meets zero-shot 6d object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 906–27 916.

[54] A. Caraffa, D. Boscaini, A. Hamza, and F. Poiesi, "Freeze: Training-free zero-shot 6d pose estimation with geometric and vision foundation models," in *European Conference on Computer Vision*, 2024.

[55] J. Huang, H. Yu, K.-T. Yu, N. Navab, S. Ilic, and B. Busam, "Matchu: Matching unseen objects for 6d pose estimation from rgb-d images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 095–10 105.

[56] E. P. Örnek, Y. Labbé, B. Tekin, L. Ma, C. Keskin, C. Forster, and T. Hodan, "Foundpose: Unseen object pose estimation with foundation features," in *European Conference on Computer Vision*, 2024, pp. 163–182.

[57] T. Wang, G. Hu, and H. Wang, "Object pose estimation via the aggregation of diffusion features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 238–10 247.

[58] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, and J. Sivic, "Megapose: 6d pose estimation of novel objects via render & compare," in *Proceedings of the 6th Conference on Robot Learning*, 2022.

[59] S. Moon, H. Son, D. Hur, and S. Kim, "Genflow: Generalizable recurrent flow for 6d pose refinement of novel objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 039–10 049.

[60] V. N. Nguyen, T. Groueix, M. Salzmann, and V. Lepetit, "Gigapose: Fast and robust novel object pose estimation via one correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9903–9913.

[61] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 868–17 879.

[62] Y. He, Y. Wang, H. Fan, J. Sun, and Q. Chen, "Fs6d: Few-shot 6d pose estimation of novel objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6814–6824.

[63] J. Sun, Z. Wang, S. Zhang, X. He, H. Zhao, G. Zhang, and X. Zhou, "Onepose: One-shot object pose estimation without cad models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6825–6834.

[64] X. He, J. Sun, Y. Wang, D. Huang, H. Bao, and X. Zhou, "Onepose++: Keypoint-free one-shot object pose estimation without cad models," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 35 103–35 115.

[65] P. Castro and T.-K. Kim, "Posematcher: One-shot 6d object pose estimation by deep feature matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2148–2157.

[66] J. Lee, Y. Cabon, R. Brégier, S. Yoo, and J. Revaud, "Mfos: Model-free & one-shot object pose estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, 2024, pp. 2911–2919.

[67] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8922–8931.

[68] K. Park, A. Mousavian, Y. Xiang, and D. Fox, "Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 710–10 719.

[69] Y. Liu, Y. Wen, S. Peng, C. Lin, X. Long, T. Komura, and W. Wang, "Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images," in *European Conference on Computer Vision*, 2022, pp. 298–315.

[70] D. Cai, J. Heikkilä, and E. Rahtu, "Gs-pose: Cascaded framework for generalizable segmentation-based 6d object pose estimation," in *International Conference on 3D Vision*, 2025.

[71] N. Gao, V. A. Ngo, H. Ziesche, and G. Neumann, "Sa6d: Self-adaptive few-shot 6d pose estimator for novel and occluded objects," in *7th Annual Conference on Robot Learning*, 2023.

[72] Y. Du, Y. Xiao, M. Ramamonjisoa, V. Lepetit *et al.*, "Pizza: A powerful image-only zero-shot zero-cad approach to 6 dof tracking," in *International Conference on 3D Vision*, 2022, pp. 515–525.

[73] C. Zhao, T. Zhang, and M. Salzmann, "3d-aware hypothesis & verification for generalizable relative object pose estimation," in *International Conference on Learning Representations*, 2023.

[74] C. Zhao, T. Zhang, Z. Dang, and M. Salzmann, "Dvmnet: Computing relative pose for unseen objects beyond hypotheses," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 485–20 495.

[75] V. N. Nguyen, T. Groueix, G. Ponimatkin, Y. Hu, R. Marlet, M. Salzmann, and V. Lepetit, "Nope: Novel object pose estimation from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 923–17 932.

[76] X. Liu, G. Wang, R. Zhang, C. Zhang, F. Tombari, and X. Ji, "Unopose: Unseen object pose estimation with an unposed rgb-d reference image," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22 023–22 034.

[77] M. Liu, S. Li, A. Chhatkuli, P. Truong, L. Van Gool, and F. Tombari, "One2any: One-reference 6d pose estimation for any object," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6457–6467.

[78] T. Lee, B. Wen, M. Kang, G. Kang, I. S. Kweon, and K.-J. Yoon, "Any6d: Model-free 6d pose estimation of novel objects," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 11 633–11 643.

[79] X. Chen, F.-J. Chu, P. Gleize, K. J. Liang, A. Sax, H. Tang, W. Wang, M. Guo, T. Hardin, X. Li *et al.*, "Sam 3d: 3dfy anything in images," *arXiv preprint arXiv:2511.16624*, 2025.

[80] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "Vggt: Visual geometry grounded transformer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5294–5306.

[81] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 2961–2969.

[82] V. N. Nguyen, T. Groueix, G. Ponimatkin, V. Lepetit, and T. Hodan, "Cnos: A strong baseline for cad-based novel object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2134–2140.

[83] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *Conference on Language Modeling*, 2024.

[84] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," in *International Conference on Learning Representations*, 2022.

[85] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "Vmamba: Visual state space model," in *Advances in Neural Information Processing Systems*, 2024.

[86] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, and K. Xu, "Geometric transformer for fast and robust point cloud registration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 143–11 152.

[87] M. Stary, J. Gaubil, A. Tewari, and V. Sitzmann, "Understanding multi-view transformers," *arXiv preprint arXiv:2510.24907*, 2025.

[88] J. Liu, W. Sun, C. Liu, X. Zhang, and Q. Fu, "Robotic continuous grasping system by shape transformer-guided multiobject category-level 6-d pose estimation," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 11, pp. 11 171–11 181, 2023.

[89] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit, "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2011, pp. 858–865.

[90] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6d object pose estimation using 3d object coordinates," in *European Conference on Computer Vision*, 2014, pp. 536–551.

[91] T. Hodan, M. Sundermeyer, Y. Labbe, V. N. Nguyen, G. Wang, E. Brachmann, B. Drost, V. Lepetit, C. Rother, and J. Matas, "Bop challenge 2023 on detection segmentation and pose estimation of seen and unseen rigid objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5610–5619.

[92] A. Doumanoglou, R. Kouskouridas, S. Malassiotis, and T.-K. Kim, "Recovering 6d object pose and predicting next-best-view in the crowd," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3583–3592.

[93] R. Kaskman, S. Zakharov, I. Shugurov, and S. Ilic, "Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[94] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," in *Robotics: Science and Systems*, 2018.

[95] A. X. Chang, T. Funkhouser, L. Guibas *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[96] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke, "Google scanned objects: A high-quality dataset of 3d scanned household items," in *IEEE International Conference on Robotics and Automation*, 2022, pp. 2553–2560.

[97] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *Asian Conference on Computer Vision*, 2013, pp. 548–562.

[98] S. Huang, Z. Gojcic, M. Usvyatsov, A. Wieser, and K. Schindler, "Predator: Registration of 3d point clouds with low overlap," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4267–4276.

[99] T. Hodan, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis, "T-less: An rgb-d dataset for 6d pose estimation of texture-less objects," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2017, pp. 880–888.

[100] B. Drost, M. Ulrich, P. Bergmann, P. Hartinger, and C. Steger, "Introducing mvtec itodd-a dataset for 3d object recognition in industry," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 2200–2208.

[101] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024.

[102] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[103] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 259–16 268.