

ZeroPose: CAD-Prompted Zero-shot Object 6D Pose Estimation in Cluttered Scenes

Jianqiu Chen, Zikun Zhou, Mingshan Sun, Rui Zhao,
Liwei Wu, Tianpeng Bao, Zhenyu He ^{† Senior Member, IEEE}

Abstract—Many robotics and industry applications have a high demand for the capability to estimate the 6D pose of novel objects from the cluttered scene. However, existing classic pose estimation methods are object-specific, which can only handle the specific objects seen during training. When applied to a novel object, these methods necessitate a cumbersome onboarding process, which involves extensive dataset preparation and model retraining. The extensive duration and resource consumption of onboarding limit their practicality in real-world applications. In this paper, we introduce ZeroPose, a novel zero-shot framework that performs pose estimation following a Discovery-Orientation-Registration (DOR) inference pipeline. This framework generalizes to novel objects without requiring model retraining. Given the CAD model of a novel object, ZeroPose enables in seconds onboarding time to extract visual and geometric embeddings from the CAD model as a prompt. With the prompting of the above embeddings, DOR can discover all related instances and estimate their 6D poses without additional human interaction or presupposing scene conditions. Compared with existing zero-shot methods solved by the render-and-compare paradigm, the DOR pipeline formulates the object pose estimation into a feature-matching problem, which avoids time-consuming online rendering and improves efficiency. Experimental results on the seven datasets show that ZeroPose as a zero-shot method achieves comparable performance with object-specific training methods and outperforms the state-of-the-art zero-shot method with 50x inference speed improvement.

Index Terms—6D object pose estimation, unseen pose estimation, zero-shot learning, three-dimensional displays, CAD model.

I. INTRODUCTION

POSE estimation is a fundamental task for bin-picking or robot-grasping in the manufacturing field. It not only involves detecting objects in 3D space but also accurately estimating the six degrees of freedom pose transformation, including the relative orientation and position, *w.r.t.* the defined CAD model. Based on the estimated pose transformation, we can establish the point-level correspondences between the given CAD model and real-world observations. Such correspondences enable robots to interact with their environment in a more informed and safe manner, especially for applications that require high precision, including collision detection, and grasp point detection in the robotic arm control.

Classic solutions [1]–[7] for pose estimation opt to learn a specific model for each object of interest using corresponding

Jianqiu Chen and Zhenyu He (Corresponding author [†]) are with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China (e-mail: zhenyuhe@hit.edu.cn). Zhenyu He is also with Pengcheng Laboratory, Shenzhen, China. Zikun Zhou is with Pengcheng Laboratory, Shenzhen, China. Mingshan Sun, Tianpeng Bao, Rui Zhao, and Liwei Wu are with SenseTime Research.

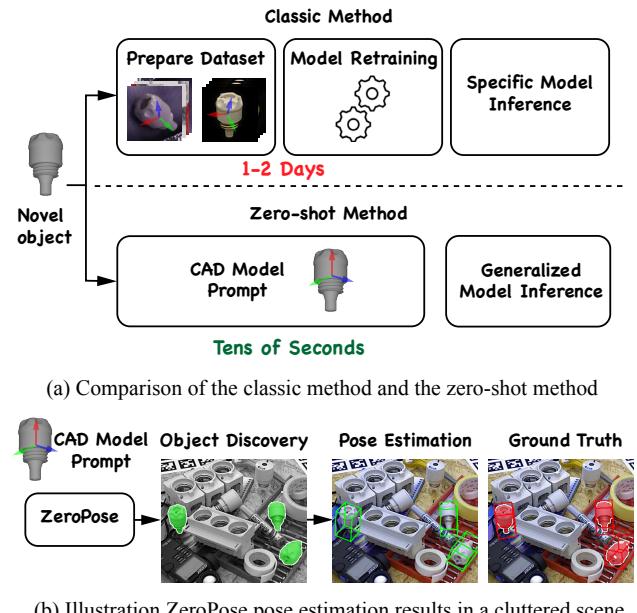


Fig. 1. (a) The classic pose estimation method applied on a novel object needs a cumbersome onboarding process for preparing a training dataset and retraining an object-specific model. The zero-shot pose estimation method adopts a pre-trained generalized model without model retraining for specific objects, reducing the onboarding time from days to tens of seconds. (b) With the CAD model prompting, the proposed ZeroPose solves both object discovery and pose estimation in a zero-shot manner.

data to predict the pose precisely and robustly. However, when encountering an unseen object, they have to collect meticulously annotated data and then train a specific model for the unseen object, as shown in Figure 1(a). Collecting thousands of training samples and retraining a model is time-consuming (usually takes 1 or 2 days) and necessitates the efforts of professional engineers. The high onboarding costs restrict the broad application of 6D object pose estimation in robotics and industrial fields.

To reduce the onboarding costs, several studies [8]–[13] pay attention to zero-shot object 6D pose estimation, which aims to empower the model with the ability to handle the object unseen training dataset, instead of retraining a new model for the novel object. Nevertheless, the zero-shot setting places rigorous demands on the generalization of the model. These algorithms rely on either additional human interactions or presupposing scene conditions, which limit the practicality in real-world applications.

Several algorithms [8], [10]–[12] rely on human interactions

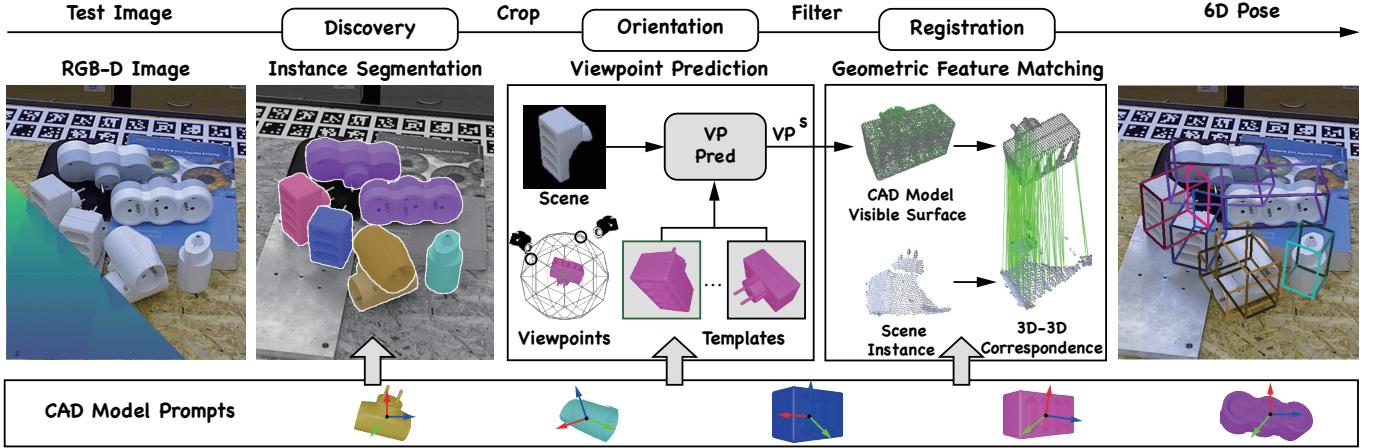


Fig. 2. A high-level overview of ZeroPose. With the prompting from CAD models, ZeroPose enables both object discovery and pose estimation in a zero-shot manner following a Discovery-Orientation-Registration (DOR) inference pipeline.

to manually localize all instances of objects from the scene and estimate their 6D poses by the model, showing limited practicality. In addition, these pose estimation models [8], [10]–[12] have an efficiency and effective issue from the render-and-compare paradigm. The computation complexity in this paradigm scales linearly with the number of candidate templates, leading to high computation costs and low inference speed. Besides, this paradigm is sensitive to illumination inconsistency, damaging the generalization ability of illumination conditions. The recent OSOP method [13] proposes a pipeline enabling automatic object discovery under specific scene conditions. This pipeline employs a zero-shot semantic segmentation module, capable of performing object discovery under the assumption of only one instance for each target object within the scene. However, object discovery by semantic segmentation is insufficient to identify the multiple instances of an object facing the challenge in cluttered scenes.

In this paper, we propose a universal framework called ZeroPose, which solves both object discovery and pose estimation in a zero-shot manner without additional human interaction or presupposing scene conditions. It performs pose estimation following a Discovery-Orientation-Registration (DOR) inference pipeline on the RGB-D image, and these three inference steps are all achieved via feature matching in which the CAD model is used as a reference. Therefore, our ZeroPose can generalize to the unseen object given its CAD model. Given the CAD model of an unseen object, ZeroPose begins with extracting its visual and geometric embeddings, which are further used as prompts. The process of the CAD model is also known as onboarding. ZeroPose then predicts the 6D pose for each instance of the object following the DOR pipeline with the prompting of the above embeddings. Figure 2 showcases the DOR inference process within a cluttered scene, delineating each step along with its corresponding results.

The first step of DOR is to discover all instances of the target object from the cluttered scene. We resort to the Segment Anything Model [14], [15] for discovering all foreground instances and then conducting feature matching between these instances and the CAD model prompt, segment-

ing the instances belonging to every target object. After lifting the instance segmentation results into 3D, we conduct point matching between the instance point cloud in the scene and the CAD model to solve the zero-shot pose estimation task. However, it is a challenging task since the instance point cloud is incomplete due to self-occlusion. Particularly, the points in the CAD model corresponding to the occluded region increase the risk of mismatching. To address this issue, we divide the matching into two steps: Orientation and Registration. Herein the Orientation step is to estimate the camera observation viewpoint to find the points of the CAD model corresponding to the visible points of the instance and filter out the remaining points of the CAD model. Based on the predicted orientation, the Registration step performs point matching between the instance point clouds and the filtered CAD model to estimate the pose transformation.

We evaluate the proposed ZeroPose on the seven datasets of the BOP benchmark [3] where there are over a hundred objects with a variety of shapes and textures and twenty thousand images under different scenes. Experimental results show that ZeroPose as a zero-shot method achieves comparable performance with object-specific training methods and shows 50x running speed improvement compared with the state-of-the-art zero-shot method.

In summary, our paper makes the following contributions:

- We propose a universal zero-shot pose estimation framework, ZeroPose, which employs a three-step Discovery-Orientation-Registration pipeline. This framework can generalize to novel objects without the need for model retraining or prior assumptions about scene conditions.
- we build a CAD model prompted zero-shot instance segmentation module based on SAM, which first leverages SAM to generate all possible proposals and then associates the potential proposals with the CAD models.
- We introduce a lightweight step-wise pose estimation paradigm that simplifies the challenge of 6D object pose estimation into camera viewpoint prediction and point cloud registration tasks.

II. RELATED WORK

A. Zero-Shot Instance Segmentation

The task of instance segmentation is designed to discover all instances of target objects within a cluttered scene. In the field of pose estimation, instance segmentation also serves as a crucial preliminary step. Many pose estimation methods [10], [16], [17] take instance segmentation as input to estimate the 6D poses of each detected instance.

Existing zero-shot instance segmentation methods [18], [19] are typically based on the text prompt as the reference of the target objects. For the dataset with the common target objects such as COCO [20], text-prompt is effective to be understood by the language model and aligned with the visual feature from the scene for segmentation. However, for the pose estimation dataset, objects with highly customizable attributes and minor inter-object variations. These challenges make it difficult to distinguish objects through textual descriptions.

Besides the text-prompt zero-shot instance segmentation, there are some related works for object discovery. Previous research [11], [13], [14], [21], [22] has primarily concentrated on zero-shot semantic segmentation and zero-shot category-agnostic instance segmentation. Some methods [11], [22] leverage the RGB and depth image for unseen object instance segmentation in a semantic category-agnostic manner. Recently, the promptable zero-shot instance segmentation methods [14], [15] are proposed, enabling segmenting instances under various types of prompts, *e.g.*, points, boxes, and anything in the foreground. However, these methods are primarily designed for foreground instance segmentation, unable to associate the predicted instances to the candidate target objects. In summary, the existing text or point prompts zero-shot instance segmentation methods are insufficient to distinguish the target objects in the pose estimation task.

B. Unseen Object 6D Pose Estimation

Category-level pose estimation is proposed to alleviate the expensive dataset preparation and training cost in recent methods [23]–[25]. The target objects in these methods are divided into categories and the model is trained for generalization on these categories. During inference, the model can be generalized to the novel object belonging to the categories seen in the training, eliminating the need for additional training. However, the applicability of this category-level pose estimation approach is limited when encountering objects of categories not present in the training dataset or when there are large intra-category variations in shape and appearance.

CAD-model-free pose estimation methods [26]–[28] leverage **one-shot** or **few-shot** pose-annotated images of the object as a reference to estimate the object pose in query images as the “virtual anchors” of Augmented Reality (AR) effects. Since target objects in AR applications are arbitrary household objects from daily lives, the CAD model is unknown and unnecessary for establishing the point-level correspondence like demands in robotic applications. Therefore, annotating the object poses by the user in a few scene images is an adorable and effective way to perform pose estimation in AR applications. Additionally, the pose of specific targets,

such as humans, can be estimated using prior information like human skeletons [29]. Recent works propose a multi-stage pipeline [30] and decentralized pose representation [31] to successfully estimate multi-person poses in challenging crowded scenes.

However, since the pose-annotated image prompt does not have a high-fidelity shape definition, these methods can not establish the important point-level correspondences from the pose-annotated image prompt for grab points detection or collision detection tasks. Moreover, the user annotation prompt reduces automation and practicality.

CAD-model-based pose estimation is introduced in recent methods [7]–[10], [12], [32]–[34]. Since the CAD model is known as the pose and shape definition of the target object and does not require any annotated samples from the real scene, these methods are usually seen as **zero-shot** learning methods. The existing zero-shot methods employ a rendering-and-compare pipeline for the generalization of different novel objects. Given a CAD model of a novel object, the pipeline online renders numerous pose hypotheses as templates for each instance of the scene and selects the optimal pose with the highest appearance feature similarity. However, the computational consumption and inference runtime increase linearly with the number of potential templates presenting an efficiency challenge. Moreover, the instances are required to be cropped from the scene by human interaction, which limits the practicality and results in a partial zero-shot setting.

Recent method OSOP [13] achieves the zero-shot pipeline under a presupposing scene condition with one and only one instance of the target object in the scene. This pipeline employs a zero-shot semantic segmentation module, capable of performing object discovery under the assumption of only one instance for each target object within the scene. However, for the cluttered scene such as in the T-LESS dataset [35], object discovery by semantic segmentation is insufficient to discover multiple instances of an object.

Besides, we analyze a potential zero-shot pose estimation approach, point cloud registration [36]–[39]. It is typically designed for scene-level registration to estimate the camera pose transformation between two scene images. However, due to the capturing source gap and a wide range of object scales, existing registration methods underperform on real-world object pose estimation datasets [3], as evidenced in the literature [40]. The capturing source gap of the scene and CAD model (RGB-D camera and 3D scanning) leads to the ratio of matchable regions less than 50%. There are only a few points as reliable inliers for matching, and the rest of the points in the CAD model are outliers leading to a mismatch and failure pose transformation. Moreover, the wide range of object scales introduces another challenge to generalization capability. Although the object pose estimation method MatchNorm [40] introduces an additional normalization layer for generalizing to different object scales, the learnable layer in MatchNorm requires training for specific objects limiting its practicability.

The proposed ZeroPose introduces a novel DOR inference pipeline addressing the 6D object pose estimation through feature matching, which avoids the need for time-consuming online rendering and improves computational efficiency. Fol-

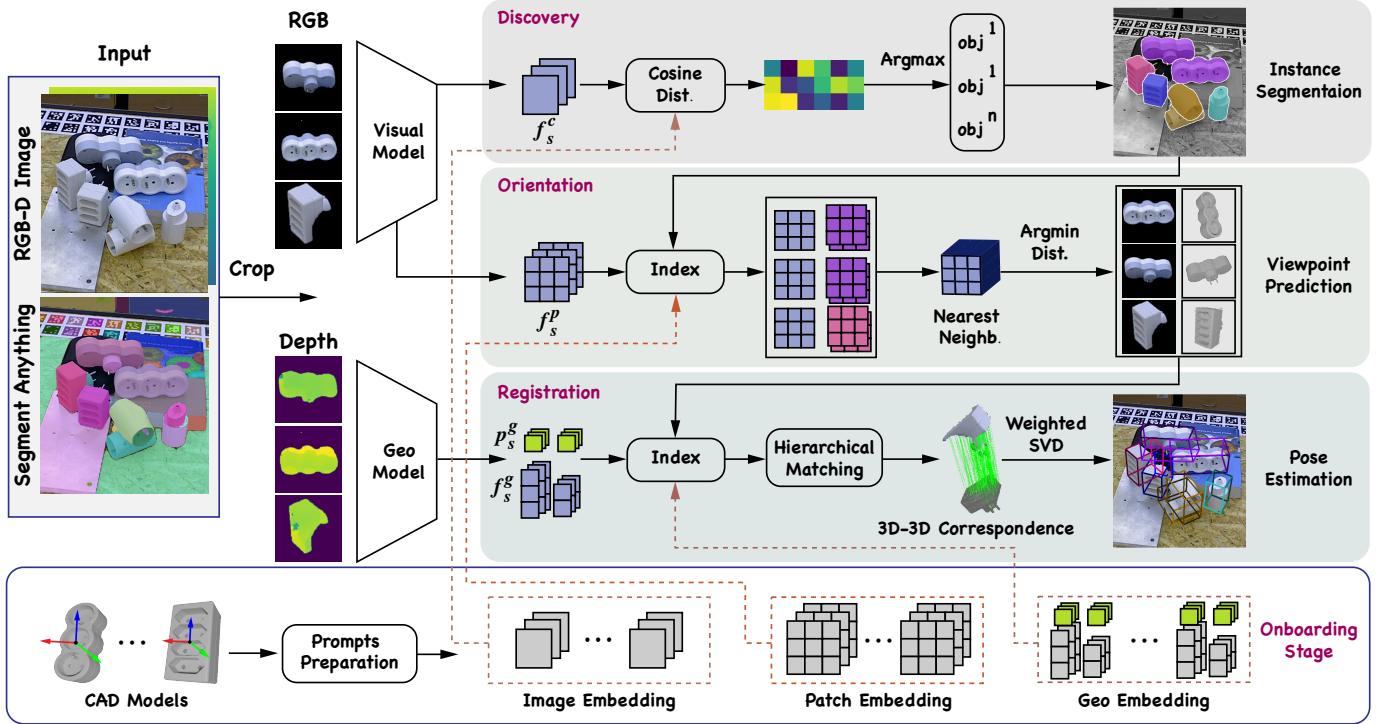


Fig. 3. An illustration of ZeroPose. ZeroPose begins with extracting embeddings from the CAD model as prompt. With the prompting, the Discovery-Orientation-Registration (DOR) inference pipeline achieves pose estimation at three inference steps. The Discovery step aims for object discovery from the cluttered scene. It calculates the image embedding cosine similarity between foreground instances from scene segment anything results and CAD model prompts and associates them based on the cosine similarity score. The Orientation step is to estimate the camera observation viewpoint to find the points of the CAD model corresponding to the visible instance points. It is based on the discovery results to index related CAD model patch embedding and estimates the camera observation viewpoint by the nearest neighbor patch embedding feature distance. The Registration step solves the pose transformation by geometric embedding matching between the instance point clouds and the filtered CAD model point clouds.

Following the release of our preprint, the significance and generalization potential of zero-shot pose estimation has garnered increased attention within the research community [41]–[43], underscoring the impact of ZeroPose.

III. METHOD

Figure 3 illustrates the ZeroPose framework. The goal of ZeroPose is the discovery of all instances of target novel objects and estimate their 6D object poses relative to the CAD model. To achieve that, given the CAD models, ZeroPose begins with preparing visual and geometric embeddings from the CAD model as prompts, at the onboarding stage (Section III-A). At inference, ZeroPose performs pose estimation following a Discovery-Orientation-Registration (DOR) inference pipeline with the prompting of the above embeddings. The discovery step (Section III-B) is to segment and crop all instances of objects from the cluttered scene and associate their related CAD models by image embedding feature matching. The orientation step (Section III-C) leverages the point embedding matching to estimate the camera observation viewpoint to find the points of the CAD model corresponding to the visible points of the scene instance. The registration step (Section III-D) aims to estimate the pose transformation from the scene instance point clouds and filtered CAD model point clouds by geometric embedding feature matching.

A. Onboarding Stage

The CAD model serves as the reference for the orientation and position of the target object, encompassing diverse visual and geometric information. To leverage the information as a prompt, we introduce a strategy to extract both visual and geometric embeddings from the CAD model during the onboarding stage. As depicted in Figure 4, we initially render the CAD model from various camera observation viewpoints, as outlined in Section III-C, resulting in R template RGB-D images and extract embeddings from the template images.

For extracting visual embeddings, a visual foundation model DINOv2 [44] (denoted visual model in Figure 4), is utilized to extract visual embeddings from the template images. The visual embeddings consist of the image embedding and the patch embedding. The image embedding is an image-level visual feature presented in f_t^c with shape (R, C) , where R is the number of rendered template images, and C is the feature dimension for the visual foundation model. The patch embedding is the patch-level visual feature presented in f_t^p with shape (R, P_l, C) , where $P_l = P_h * P_w$, P_h, P_w are the height and width of the patch-level visual feature output of visual model [44]. For extracting geometric embeddings, we employ a pretrained geometric model (referred to as the Geo model) to extract point cloud and geometric features from the CAD model. To achieve that, template RGB-D images are lifted into a color point cloud with normalized representation, and

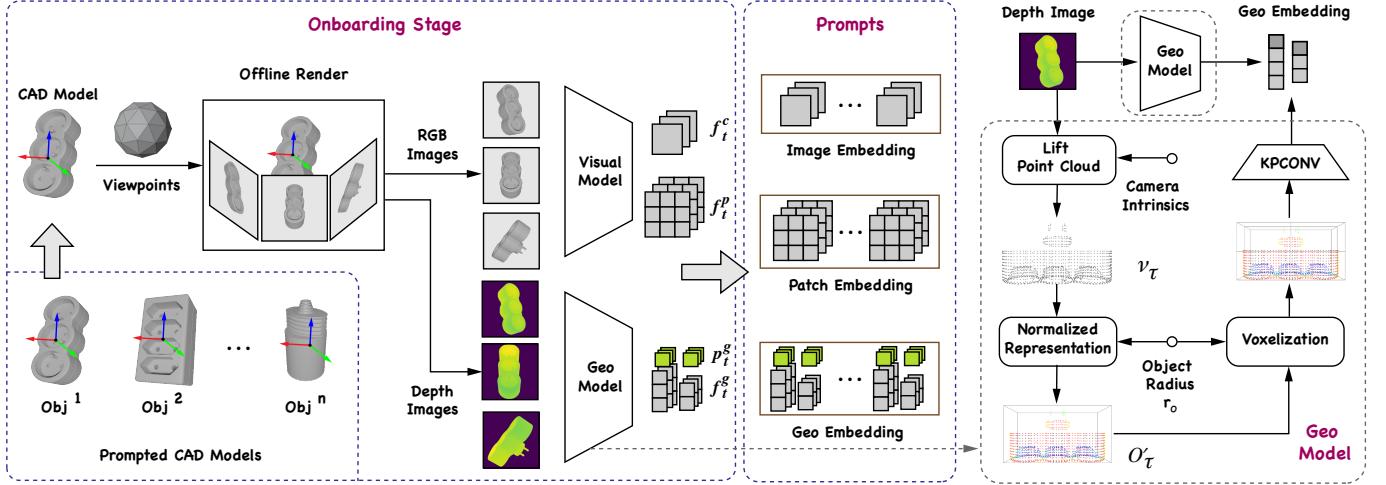


Fig. 4. Left: The onboarding stage aims to extract visual and geometric embeddings from the CAD model as the prompt of the target object, which is offline and only requires running once for each object. Right: Illustration for the Geo Model.

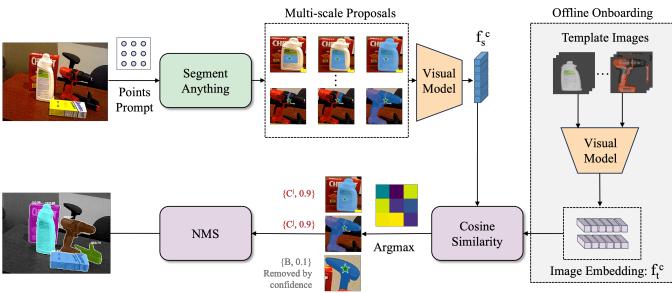


Fig. 5. An illustration of the discovery step of ZeroPose.

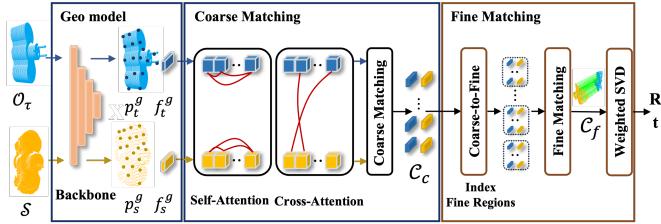


Fig. 6. The architecture of hierarchical matching in the Registration step.

voxelize the point cloud p_t^g to extract hierarchical geometric features f_t^g . The geometric embedding consists of $\{p_t^g, f_t^g\}$, where $p_t^g = \{p_{t,c}^g, p_{t,f}^g\}$ and $f_t^g = \{f_{t,c}^g, f_{t,f}^g\}$. c and f denote the coarse-level and the fine-level.

Both operations in the onboarding stage are conducted offline and performed only once for each novel object before the inference stage.

B. Discovery

Object discovery, as the first step in the DOR inference pipeline, aims to discover all instances of objects from the scene. To achieve zero-shot object discovery, we resort to a Segment Anything Model (SAM) [15] to generate all possible proposals and then associate the potential proposals with the CAD models via feature matching.

For proposal generation, we leverage the SAM with a uniform point set as a prompt to generate proposal instance masks. Since the proposal instance masks are object-agnostic, the proposal instances are needed to associate with the prompted CAD model, and irrelevant instances need to be filtered out. To this end, we propose an object association approach, demonstrated in Figure 5. Specifically, we crop these instances from the image and resize them into a fixed image size for extracting their image embedding through the visual model. Image embedding is the image-level visual feature f_s^c , with dimensions (M, C) . M represents the number of proposal instances. Then, we calculate the global visual feature similarity between the proposal instances and prepared template images. The proposal instance is scored by this highest similarity and assigned with the ID of the CAD model corresponding to the highest feature similarity template images. The association processing is formulated as follows:

$$O_{id} = \arg \max_{n=1, \dots, N} \max_{\mathcal{T}=1, \dots, R} \left(\frac{f_s^c \cdot f_{t,n,\mathcal{T}}^c}{\|f_s^c\| \|f_{t,n,\mathcal{T}}^c\|} \right), \quad (1)$$

where O_{id} is the related CAD model ID of scene instances. The instance mask combined with the predicted CAD model ID results in zero-shot instance segmentation.

To filter irrelevant instances, we remove the proposal instances by score threshold. Besides, for some over/under-segmentation results, we leverage the Non-Maximum Suppression (NMS) to filter out. That is because the scoring mechanism will assign quite a low score for the over/under-segmentation result and assign a higher score for the accurate-segmentation result. Therefore, we filter out the suboptimal segmentation masks by NMS.

Additionally, we introduce a text-prompt zero-shot instance segmentation within pose estimation datasets for comparison. There is a similar pipeline that instead of the image embedding into the text embedding. This pipeline leverages the multi-modal vision model [45] instead of the visual model [44] to extract the multi-modal feature of the scene image feature and the text feature of target objects for matching. However, in the

pose estimation dataset, many objects are manufactured often without text descriptions for embedding. To solve this issue, we leverage the point cloud foundation modal PointLLM [46] to generate a caption of the CAD model for extracting the text embedding. Details of the design and experiment results are shown in Section IV-C.

C. Orientation

The Orientation step aims to estimate the camera observation viewpoint to find the points of the CAD model corresponding to visible points of the predicted instance in the scene and filter out the remaining points of the CAD model. Since the instance in the scene is a partial observation from a camera viewpoint, it is incomplete due to self-occlusion. Directly estimating the 3D correspondence between the occluded scene instance and the CAD model increases the risk of mismatching. To solve this, the orientation step predicts the camera observation viewpoint by patch embedding feature matching with a few discrete templates. Then, based on the pin-hole imaging principle, we index the points of the CAD model corresponding to the visible instance points and filter out the remaining points of the CAD model.

Compared with the 6 degrees of freedom (DoF) camera perspective rendering for pose estimation, the predefined camera viewpoints are tailored for filtering invisible points of the CAD model. To reduce the number of rendering templates and improve efficiency, we only vary the visible region-related 2 DoF optical axis direction in camera perspective for template rendering and fix the rest 3 DoF translation and 1 DoF in-plane rotation. Specifically, by revisiting the camera observation viewpoint, we find that the 3 DoF orientation of camera viewpoint \mathbf{R}_O can be decomposed into a 2 DoF optical axis direction \mathbf{R}_γ of the image plane and a one-degree-of-freedom in-plane rotation \mathbf{R}_θ among this direction. This decomposition is formalized by the equation:

$$\mathbf{R}_O = \mathbf{R}_\gamma \mathbf{R}_\theta, \quad (2)$$

where

$$\mathbf{R}_\gamma = \begin{bmatrix} \cos \varphi \cos \psi & -\sin \varphi & \cos \varphi \sin \psi \\ \sin \varphi \cos \psi & \cos \varphi & \sin \varphi \sin \psi \\ -\sin \psi & 0 & \cos \psi \end{bmatrix}, \quad (3)$$

$$\mathbf{R}_\theta = \begin{bmatrix} \cos \beta & -\sin \beta & 0 \\ \sin \beta & \cos \beta & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (4)$$

(φ, ψ, β) are Euler angles. Since in-plane rotation does not affect the visible region, the \mathbf{R}_O prediction task can be simplified into a two-degree-of-freedom viewpoint direction \mathbf{R}_γ prediction task. The candidate camera viewpoints can be uniformly sampled from a sphere surface [47] and adopt a few discrete templates to approximate its distribution because of the narrowed sampling space.

To predict the closest viewpoint from templates associated with the detected object, we introduce patch-level feature matching in the viewpoint prediction step. Since closed viewpoint template images only have minor and local appearance

variations, the image embedding for global semantic information in the discovery step is insufficient to distinguish these similar viewpoint template images of an object. To solve that, we extract patch-level features from visual model [44] as patch embedding to calculate candidate viewpoints by patch embedding matching between the instance in the scene and template images. Specifically, given viewpoint template images and the scene instance predicted from the discovery step, we adopt a long-side resize to scale them into a fixed image size and hold the optical axis in the center. Given the cropped images, we extract patch-level features from visual model [44] as patch embedding, which is rotate-invariant for in-plane rotation \mathbf{R}_θ . Specifically, each patch region in the segmented proposal instance searches for the nearest patch in the reference template image and uses the mean of local maximum feature similarity as the score for selecting viewpoints as follows:

$$\mathbf{S}_P^T = \frac{1}{N_s} \sum_{j=1}^{N_s} \max_{i=1, \dots, N_t^T} \frac{\mathbf{f}_{s,j}^p \cdot \mathbf{f}_{t,i}^p}{\|\mathbf{f}_{s,j}^p\|_2 \|\mathbf{f}_{t,i}^p\|_2}. \quad (5)$$

The template viewpoints are sorted by the local patch feature similarity \mathbf{S}_P^T and we select the top k as candidates. Besides, we also leverage the patch-level feature similarity to revise the predicted instance score and use the similarity to estimate the visibility of segmented instances. Then, we update the score of each instance with local similarity and visibility via weighted summation. The visible region of the CAD model under these viewpoints can also be calculated from the depth image, as introduced in the onboarding stage Section III-A.

D. Registration

The registration step aims to estimate rigid pose transformation $\mathbf{R} \in SO(3), \mathbf{t} \in \mathbb{R}^3$ from the scene instance point clouds and filtered CAD model point clouds. However, recent research studies [48], [49] have observed that the image feature matching for establishing the point-to-point correspondence is unreliable in the scenario of object pose estimation. The visual features are sensitive to the illumination inconsistency between the real scene and synthetic template images. To solve that, we introduce the geometric embedding feature matching. As illustrated in Figure 6, the Geo model hierarchically samples the point clouds and extracts their geometric features, which are then combined into a geometric embedding. A hierarchical matching module subsequently utilizes these point clouds and their associated geometric features to facilitate feature interaction, aiming to find their 3D-3D correspondences. Based on the predicted correspondence, we adopt the 3D coordinates of corresponding point pairs and the pose transformation formula to calculate the 6D pose R, t by least-squares minimizing the transformed point cloud distances.

1) **Geo Model:** Geo model is to extract a stable geometric embedding that can be generalized for different scale and shape objects. There are three primary steps illustrated in Figure 4. The lift point cloud step is to lift the point cloud pairs from the scene and template depth images. For each instance in the scene, we crop and mask them from the image and

lift their point cloud from the depth image given the camera intrinsic matrix, calculated as follows:

$$\begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} = \begin{bmatrix} \frac{u_i - c_x}{f_x} \\ \frac{v_i - c_y}{f_y} \\ 1 \end{bmatrix} \times d_i, \quad (6)$$

where c_x, c_y, f_x, f_y are from camera intrinsic parameters and u_i, v_i is the coordinate in the 2D image and d_i is the cooresponding depth value from the depth image. Based on the segmentation result, we separately calculate each point cloud of instance $\mathcal{S} = \{\mathbf{s}_i \in \mathbb{R}^3 \mid i = 1, \dots, m\}$ from the depth image and combine their corresponding color from RGB image as texture feature. For the filtered CAD model point cloud, we initially lift point cloud $\mathcal{V}_{\mathcal{T}} = \{\mathbf{v}_i \in \mathbb{R}^3 \mid i = 1, \dots, n\}$ from the template RGB-D image and undergo an inverse transformation using the template pose $T_{\mathcal{T}}$ to align to the base orientation and position in the CAD model as follows:

$$\begin{bmatrix} a_i \\ b_i \\ c_i \\ 1 \end{bmatrix} = T_{\mathcal{T}}^{-1} \begin{bmatrix} x_i^T \\ y_i^T \\ z_i^T \\ 1 \end{bmatrix}, \quad (7)$$

where the (x_i^T, y_i^T, z_i^T) represents the value of p_i and (a_i, b_i, c_i) is the surface point \mathbf{o}_i of the viewpoint-filtered CAD model $\mathcal{O}_{\mathcal{T}} = \{\mathbf{o}_i \in \mathbb{R}^3 \mid i = 1, \dots, n\}$.

To address scale variability in unseen objects, we propose a normalized representation in the following step. Given the prompted CAD model, we calculate the radius $r_{\mathcal{O}}$ of the circumscribed sphere of the CAD model. For a point cloud pairs $(\mathcal{S}, \mathcal{O}_{\mathcal{T}})$, we scale both of them by the $\frac{1}{r_{\mathcal{O}}}$ into a normalized space, achieving $(\mathcal{S}', \mathcal{O}'_{\mathcal{T}})$. This normalized representation allows for the maintenance of a receptive field for geometric features, which can dynamically scale the point cloud into a shared normalized space. Additionally, the normalized representation combines with centroid clustering enabling filtering of the noisy points from the sensor noise or inaccuracy segmentation. In the centroid clustering algorithm, the critical bandwidth parameters can be selected to 1 at any normalized representation instance point clouds, eliminating the need for manual hyperparameter tuning.

After normalization, the point clouds are inputted into the geometric backbone network to extract geometric features. We utilize KPConv [50] as the backbone, a method that voxelizes the point cloud across various local receptive fields, thereby extracting hierarchical point clouds and geometric features as geometric embedding $\{p_s^g, f_s^g\}$, where $p_s^g = \{p_{s,c}^g, p_{s,f}^g\}$ and $f_s^g = \{f_{s,c}^g, f_{s,f}^g\}$. c and f denote the coarse-level and the fine-level. The Geo model is trained using correspondence losses together with the following geometric matching model.

2) Hierarchical Matching for Registration: In this step, the goal is to estimate 3D correspondences at geometric embeddings and estimate the final pose transformation parameters from the correspondences. Given input geometric embeddings from scene instances and predicted template instances, the hierarchical matching module conducts coarse-to-fine feature interaction, and through the similarity of these features establishes the correspondence. Then, we utilize the

point cloud coordinates of correspondence features to fit a pose transformation that minimizes the distance between the transformed point clouds.

As shown in Figure 6, given the input geometric embeddings from Geo model, we use the coarse-level point cloud coordinates and features from the scene instances $\{p_{s,c}^g, f_{s,c}^g\}$ and template instances $\{p_{t,c}^g, f_{t,c}^g\}$ as input to perform the feature fusion and interaction by the attention modules and select the features with high similarity as coarse-level correspondence \mathcal{C}_c .

Then, we adopt the receptive field of coarse-level features in correspondence \mathcal{C}_c to index related fine-level points and features. By refining correspondence in the fine-level feature matching, we estimate the dense point-to-point correspondence \mathcal{C}_f . Based on the dense correspondence \mathcal{C}_f and point cloud coordinates, we can formulate the object pose estimation into a weighted least-squares problem. We adopt the coordinates of the corresponding points to minimize the point cloud distance, as follows:

$$\min_{\mathbf{R}, \mathbf{t}} \sum_{(\mathbf{o}_{x_i}, \mathbf{s}_{y_i}) \in \mathcal{C}_f} \|\mathbf{R} \cdot \mathbf{o}_{x_i} + \mathbf{t} - \mathbf{s}_{y_i}\|_2^2, \quad (8)$$

where \mathbf{s}_{y_i} and \mathbf{o}_{x_i} are matched corresponding points in the fine-level matching. The closed-form solution of the least-squares problem is the pose parameters \mathbf{R}, \mathbf{t} with minimal point cloud distance, which can be solved by the weighted singular value decomposition (weighted SVD) [51] algorithm.

To enhance the robustness of viewpoint prediction, we introduce a multi-hypothesis registration approach that supports multiple candidate viewpoint-filtered point cloud input. Given k different viewpoint-filtered CAD model point clouds, we establish the 3D correspondences between the scene instance and them and result in k candidate poses. Subsequently, these filtered CAD model point clouds are transformed into the scene according to their respective candidate poses. We calculate their Chamfer distances [52] and select the final pose estimation result with minimal point cloud distances.

IV. EXPERIMENTS

A. Benchmark

Datasets. For the training dataset, we take the GSO dataset [58] to pretrain the Geo model and the geometric feature matching module, which contains 1000 3D objects under household scenes and 1 million synthetic images provided by Megapose [10]. For the test datasets, we follow the BOP challenge [3] to select the seven core datasets where all test images are captured from the real-world environment, including LineMod Occlusion (LMO), T-LESS, TUD-L, IC-BIN, YCB-Video (YCB-V), ITODD, and HomebrewedDB (HB). The first 5 datasets are open for offline access and typically adopted as the ablation study benchmark, named BOP 5. The 7 core datasets are denoted BOP 7. BOP datasets include over a hundred objects with a variety of shapes and textures and twenty thousand images from challenging real scenes.

Metrics. For the instance segmentation metric, we adopt the Mean Average Precision (mAP) as defined in the BOP

TABLE I

INSTANCE SEGMENTATION RESULTS ON THE BOP CHALLENGE DATASETS. WE REPORT THE MAP AS A METRIC. THE HEADER “REAL” INDICATES TRAINING WITH REAL SCENE IMAGES. †DENOTES THAT THE DATASET IS NO REAL IMAGES PROVIDED FOR TRAINING. * DENOTES THAT THE OBJECT NAME IS NOT PROVIDED AND WE USE THE PREDICTED CAPTION FROM POINTLLM [46] AS AN ALTERNATIVE.

Method	Settings		BOP7 Datasets									Mean	Time (s)
	Zero-shot	Real	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V				
1 Mask RCNN [53]	X	X	37.5	51.7	30.6	31.6	12.2	47.1	42.9	36.2	0.054		
2 Mask RCNN [53]	X	✓	37.5†	54.4	48.9	31.6†	12.2†	47.1†	52.0	40.5	0.055		
3 ZebraPoseSAT [16]	X	X	50.6	62.9	51.4	37.9	36.1	64.4	62.6	52.3	0.080		
4 ZebraPoseSAT [16]	X	✓	50.6†	70.9	70.7	37.9†	36.1†	64.4†	74.0	57.8	0.080		
5 Text-prompt (ImageBind)	✓	X	10.7	1.2*	17.3*	15.1	6.3	16.3*	38.3	15.0	0.427		
6 CAD-prompt (ImageBind)	✓	X	30.4	25.1	28.6	18.8	19.7	41.5	48.5	30.4	0.429		
7 CAD-prompt (DINOv2)	✓	X	37.7	34.7	46.0	20.6	20.2	47.8	57.4	37.8	0.220		
8 CAD-prompt (DINOv2/Pyrender)	✓	X	34.4	31.3	51.5	21.7	14.6	47.6	60.0	37.1	0.220		

TABLE II

POSE ESTIMATION RESULTS ON THE BOP CHALLENGE DATASETS. WE REPORT THE AR SCORE ON EACH OF THE 7 CORE DATASETS IN THE BOP CHALLENGE AND THE MEAN SCORE ACROSS DATASETS. ZERO-SHOT STANDS FOR THE MODEL IN THE OBJECT DISCOVERY OR POSE ESTIMATION STEP ENABLING GENERALIZATION TO THE NOVEL OBJECT WITHOUT RETRAINING. FOR EACH COLUMN, WE DENOTE THE BEST OVER RESULT IN *italics* AND THE BEST ZERO-SHOT POSE ESTIMATION METHOD FOR EACH SETTING BLOCK IN **BOLD**. THE UNIT OF TIME COLUMN IS SECONDS (S).

Method	Object Discovery		Pose Estimation		BOP7 Datasets									Mean	Time (s)
	Zero-shot	Inst. level	Zero-shot	Refinement	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V				
1 CosyPose [54]	X	✓	X	✓	63.3	64.0	68.5	58.3	21.6	65.6	57.4	57.0	0.5		
2 CDPNv2 [55]	X	✓	X	✓	63.0	43.5	79.1	45.0	18.6	71.2	53.2	53.4	1.5		
3 SurfEmb [17]	X	✓	X	✓	76.0	82.8	85.4	65.9	53.8	86.6	79.9	75.8	9.0		
4 Coupled [56]	X	✓	X	✓	73.2	82.0	85.8	60.6	47.2	87.3	82.9	74.1	-		
5 MegaPose [10]	X	✓	✓	-	18.7	19.7	20.5	15.3	8.00	18.6	13.9	16.2	25.6		
6 OVE6D [12]	X	✓	✓	✓	49.6	52.3	-	-	-	-	57.5	-	-		
7 Ours	X	✓	✓	-	58.3	55.9	86.9	53.2	33.8	69.3	70.1	61.1	1.8		
8 OVE6D [12]	X	✓	✓	✓	62.7	54.6	-	-	-	-	58.7	-	-		
9 GCPose [57]	X	✓	✓	✓	65.2	67.9	92.6	-	-	-	75.2	-	-		
10 MegaPose [10]	X	✓	✓	✓	58.3	54.3	71.2	37.1	40.4	75.7	63.3	57.2	93.3		
11 Ours	X	✓	✓	✓	66.3	63.0	94.9	52.0	44.2	82.0	84.1	69.5	48.3		
12 DrostPPF [36]	✓	✓	✓	✓	52.7	-	-	-	-	-	34.4	-	-		
13 PPF + Zephyr [8]	✓	✓	✓	✓	59.8	-	-	-	-	-	51.6	-	-		
14 OSOP [13]	✓	X	✓	✓	48.2	-	-	-	-	-	60.5	57.2	-		
15 Ours + MegaPose [10]	✓	✓	✓	✓	60.1	46.8	84.3	32.7	47.9	68.6	75.7	59.4	234.1		
16 Ours (Pyrender)	✓	✓	✓	-	49.0	39.6	74.7	30.3	37.3	56.1	66.6	50.5	4.95		
17 Ours	✓	✓	✓	-	50.9	42.1	60.4	42.0	46.3	53.8	64.8	51.5	4.81		
18 Ours	✓	✓	✓	✓	58.3	49.6	72.5	44.9	51.5	64.0	79.0	60.0	85.9		

Challenge [3], [20], [59]. For the pose estimation metric, we follow [3], [59] to measure the Average Recall (AR) of the mean of VSD (Visible Surface Discrepancy), MSSD (Maximum Symmetry-Aware Surface Distance), and MSPD (Maximum Symmetry-Aware Projection Distance) to measure the pose estimation performance of pose estimation. More detailed definitions of evaluation metrics can refer to the BOP Challenge [3], [59].

B. Implementation Details

ZeroPose is trained on the SLURM cluster with 8 NVIDIA V100 and inferring on the PC with NVIDIA RTX 3090.

Visual model. We adopt the pretrained visual foundation models DINOv2 [44] as our visual model, which is based on the ViT architecture [60]. The visual features of foreground instances are cropped and resized with shape (224, 224, 3) as input to extract the scene images template images embeddings, and patch embeddings. For the proposed text-prompt zero-shot instance segmentation pipeline, we substitute the visual

model with a multi-modal vision model [45] and replace the template image embedding with text embedding derived from the given object name. For datasets without object names or text descriptions, we generate a caption from the 3D-text multi-modal model PointLLM [46] as the text description.

Geo model. To improve the robustness of geometric extraction and matching at the proposed normalized point cloud pairs, we pretraining the Geo model and geometric matching model on the synthetic GSO [10], [58] dataset. To enhance robustness against variations in illumination and occlusions, our approach incorporates data augmentation techniques, including color and position jittering. For the training of high-level feature extraction and matching, we adopt the overlap-aware circle loss [37], to train the model to locate the high overlap region. We calculate the correspondence regions by the ground truth pose and select the overlap of correspondence regions more than 10% as positive samples. The others that have an overlap of less than a threshold are viewed as negative samples. For low-level, we match each low-level point pair by

the distance within a matching radius and adopt negative log-likelihood loss to train the network to find the correct matched pairs. High-level and low-level features can be simultaneously trained in a network architecture with an equal weight based on their respective loss functions.

Segmentation anything. We adopt the recent lightweight version interactive segmentation method FastSAM [15], with default CNN backbone to generate the mask of foreground instances and filter out the tiny regions with less than 128 pixels. Moreover, since there are possible multiple point prompts for one instance, we remove the duplicating regions in predicted foreground instances by the Non-Maximum Suppression algorithm to filter the Intersection over Union (IOU) of masks over 0.25 for each object.

C. Evaluation of Zero-Shot Instance Segmentation

Since there are no other related methods following the CAD-model-prompt zero-shot instance segmentation setting, we select two classic supervised object-specific instance segmentation methods and implement a text-prompt zero-shot instance segmentation method for comparison. Compared with the object-specific methods, although without the training for target objects, the proposed object discovery module achieves comparable performance with the classic Mask RCNN method [53]. However, compared with the state-of-the-art object-specific method ZebraPoseSAT [16], there is still room for improvement. Moreover, for a fair comparison of different prompts, we implement a text-prompt zero-shot instance segmentation pipeline following the same pipeline as in our discovery step. We substitute the visual model with a multi-modal vision model [45] and replace the template image embedding with text embedding derived from the given object name. As depicted in Table I, our CAD-model-prompt zero-shot instance segmentation method shows obvious performance improvements compared with the text-prompt method from 15.0% to 30.4%. When converting the multi-modal vision model Imagebind [45] into the visual foundation modal DINOv2 [44], the performance is increased to 37.8%. These demonstrate the proposed CAD model prompt is more reliable than the text prompt in the object pose estimation scenario.

D. Evaluation of Zero-Shot Pose Estimation

To evaluate our pose estimation performance, we compared classic object-specific pose estimation methods in Table II rows 1-4 and the latest partial (rows 5-11) or fully (rows 12-17) zero-shot pose estimation methods, where the partial zero-shot methods adopt an object-specific object discovery model training on the prompted objects. Compared with classic object-specific methods, the proposed ZeroPose without model retraining on specific test objects outperforms the CosyPose [54] and CDPNv2 [55]. For the ADD-S 2CM metric in the DenseFusion [1], under the same segmentation results, our method achieves 96.6 in the YCB-V [2] dataset, which is comparable with the 96.8 in DenseFusion [1] training on the specific test objects. However, as a zero-shot method, it still has room for improvement compared with the recent object-specific methods.

Compared with zero-shot methods, ZeroPose demonstrates better accuracy and generalization capability under both the partial and fully zero-shot settings. The Discovery-Orientation-Registration (DOR) pipeline enables to perform on a variety and challenging scenes such as cluttered scenes and generalize to unseen objects. OVE6D [12], GCPose [57], and Megapose [10] are capable of performing unseen object pose estimation without retraining. However, these methods require human intervention to segment the target objects from the scene image, preventing them from operating in a fully zero-shot manner. These methods evaluate the performance of the BOP datasets requiring a customized Mask-RCNN model [53] to segment the objects of interest and then perform pose estimation with these accurate masks. For fairly compared with them, we adopt the same object-specific training Mask RCNN model [53] for object discovery and compare the pose estimation performance. As demonstrated in rows 5 and 7, ZeroPose outperforms the state-of-the-art method MegaPose and achieves 50 times running speed improvement and 98% running time reduction. Since the existing pose refinement strategies [10], [51], [56] are typically generalized to novel objects, pose initialization is the biggest challenge in zero-shot object pose estimation. Compared to MegaPose [10] for pose initialization under the same instance segmentation results provided by the MaskRCNN [53], our method achieves a 44.9% performance gain and is 14.2 times faster, as shown in rows 5 and 7 of Table II. Moreover, the MegaPose has a limitation in that it requires a detection result from the user as input (supervised Mask RCNN provided in the BOP evaluation) to locate the candidate object but our method can estimate the pose in a fully zero-shot manner with comparable performance of 51.5%.

In addition, we also integrate our zero-shot object discovery into MegaPose. As shown in row 15, the MegaPose equipped with our object discovery surpasses its original version (59.4% vs 57.2%), demonstrating the effectiveness of our object discovery. To evaluate the optimal performance of the zero-shot method, we integrated the refiner from MegaPose [10] with the proposed ZeroPose framework. As illustrated in row 18, it achieves 60% for AR outperforming the object-specific methods, demonstrating the effectiveness of the zero-shot method.

Compared to the non-learning-based method DrostPPF [36] and its refinement method Zephyr [8], our ZeroPose achieves better average results and robustness, which can perform at more challenging datasets. Compared to OSOP [13], although its object discovery module is zero-shot, it introduces a presupposing scene condition of one instance per object, limiting its applicability to the cluttered scene with multiple instances of the same object. Even without the presupposing scene condition, ZeroPose surpasses OSOP on the mean of their provided three datasets (55.3% and 56.5%) and shows comprehensive performance surpassing after adopting the refiner [10]. As shown in rows 9 and 11, for datasets (T-LESS, IC-BIN, ITODD) with cluttered scenes, OSOP is unavailable, but the proposed ZeroPose achieves both zero-shot object discovery and pose estimation. These experiment results and comparisons demonstrate the effectiveness and efficiency of

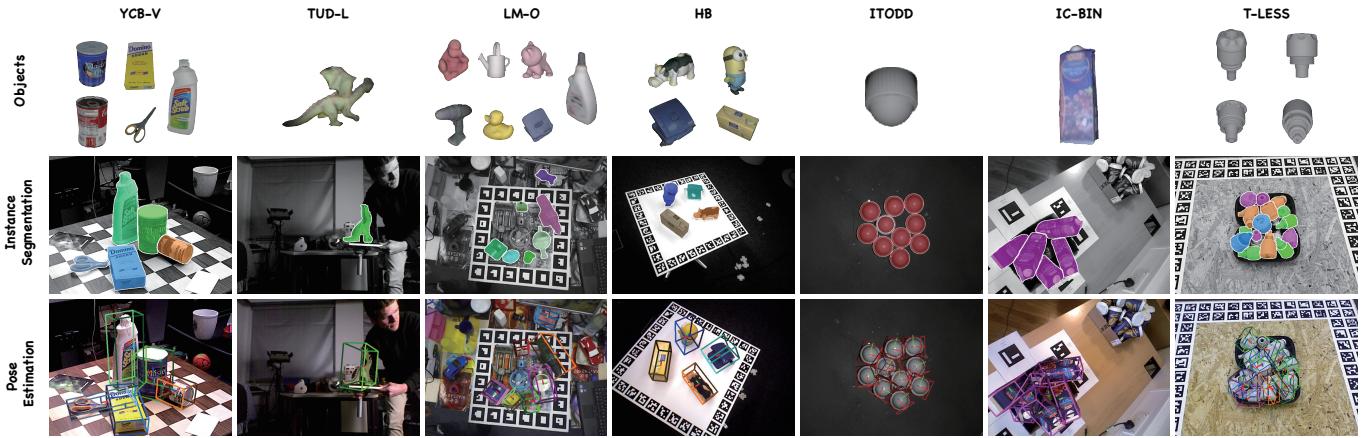


Fig. 7. Visualization instance segmentation and pose estimation results on the 7 BOP core datasets. The first row represents the prompted objects.

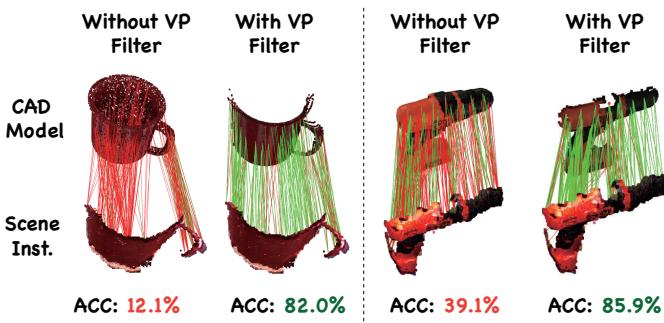


Fig. 8. Visualization of the geometric matching. The first and third columns are complete CAD model point clouds and the second and fourth columns are viewpoint-filtered point clouds. ACC stands for feature matching accuracy.

the proposed zero-shot object pose estimation method.

E. Visualization

As illustrated in Figure 7, ZeroPose performs well on a variety of objects, from simple geometric shapes (as seen in LM-O and TUD-L) to more complex and textured objects (as seen in YCB-V). Furthermore, as demonstrated by the object discovery (instance segmentation) and pose estimation results in the ITODD, IC-BIN, and ITODD datasets, ZeroPose exhibits reasonably good performance in cluttered scenes. For objects with less distinctive features or similar shapes (YCB-V and T-LESS datasets), ZeroPose can distinguish objects based on texture information and local geometric structure differences. The visualization analysis indicates that ZeroPose can handle a diverse set of novel objects and scene conditions.

F. Ablation Study

Comparison of the viewpoint densities. To evaluate the effect of different viewpoint densities, we render the template images from different viewpoint densities and viewpoint sampling strategies. For six template images, the camera positions encompass the front, back, left, right, top, and bottom orientations relative to the object. The others are uniformly sampled from the surface of a unit sphere with different densities following [47]. As shown in Table III, we compare

the onboarding time, cache memory consumption, inference running time, and accuracy (mAP) at different viewpoint densities. Since the 6 template images are too sparse, they are unable to fully cover possible object appearances. Additionally, experimental results indicate that excessively dense template images do not yield additional performance gains in our setting. This is because the viewpoint angular difference between the template images and the scene images is already minimal, which is no more than 15.9 degrees in the 42-viewpoints setting. Consequently, the use of excessively dense template images not only fails to enhance the discriminability of positive samples but also increases the confidence scores of confounding negative samples within the scene, thereby impeding object discovery. Hence, we choose 42 viewpoints template images for a CAD model in this paper.

Comparison of the template images rendering engines. To evaluate the effect on the template images render engine, we compare a high-fidelity engine Blender [62], and a high-speed engine, Pyrender [63]. The rendered template images from Blender have better illumination consistency with the scene images compared with the high-speed engine. Compared with Blender, Pyrender only leads to marginal performance drops of 1% in average AR for pose estimation in line 8 of Table II and 0.7% mAP in line 16 of Table I for instance segmentation. Namely, Pyrender saves 95% of preprocessing time without significantly compromising pose estimation performance. For large-scale applications, we can use Pyrender as the rendering engine. In this way, the proposed method not only performs online pose estimation efficiently but also conducts offline preprocessing at a low time cost.

Effect of occluded object discovery. To analyze the robustness of the object discovery under occlusion, we calculate the segmentation recall of generated proposals having an IoU over 0.5 with Ground Truth and the association accuracy of proposals on 5 core datasets of BOP (BOP5) under different occlusion fractions, as shown in Figure 9. When the object is half occluded (0.4–0.6 occlusion fraction), the segmentation recall is over 70% and the association accuracy is over 80%. Heavy occlusion (e.g., 0.8–1.0 occlusion fraction) will affect the segmentation recall and association accuracy, as too few pixels are visible to recognize the target object.

TABLE III
COMPARISON OF THE VIEWPOINT DENSITIES. THE METRIC IS MAP [3] FOR INSTANCE SEGMENTATION TASK.

Template Num.	Onboarding Time (s)			Template Feat.	BOP5 Datasets					Time (s)
	Render	Feat Extraction	Total		LM-O	T-LESS	TUD-L	IC-BIN	YCB-V	
6	0.3	0.1	0.4	0.02	29.7	11.6	45.2	19.8	32.0	27.6 0.240
42	26.5	0.5	27.0	0.16	37.7	34.7	46.0	20.6	57.4	39.3 0.242
162	102.2	2.3	104.5	0.64	37.8	35.7	45.4	20.8	56.0	39.2 0.251
642	405.1	7.2	412.3	2.51	37.4	36.4	44.7	20.5	56.0	38.9 0.289

TABLE IV

COMPARISON OF VIEWPOINT PREDICTION STRATEGIES. THE METRIC IS THE DEGREE FOR THE PREDICTED VIEWPOINT DIRECTION WITH THE GROUND TRUTH DIRECTION. THE DEGREE ERROR IS FROM 0 TO 180, WHERE THE LOWER VALUE INDICATES A HIGHER ACCURACY IN PREDICTING THE VIEWPOINT DIRECTION. D STEP INDICATES THE RESULT FROM THE FIRST DISCOVERY STEP.

Setting	LM-O	T-LESS	TUD-L	IC-BIN	YCB-V	Mean
Image-Level (D step)	57.3	45.6	51.7	74.9	55.9	57.1
Patch-Level (top1)	41.2	34.7	34.3	63.6	36.1	42.0
Patch-Level (top5)	19.5	21.5	15.5	26.4	18.9	20.4

TABLE V

COMPARISON OF DIFFERENT POSE ESTIMATION STRATEGIES. HGM IS THE HIERARCHICAL GEOMETRIC MATCHING MODEL. IP ROT. DENOTES THE TEMPLATE VIEWPOINTS WITH ADDITIONAL IN-PLANE ROTATIONS

Model	Paradigm	AR (%)
1 Visual	Feature Matching	16.3
2 Visual	Template Matching	24.4
3 Visual	Template Matching + ICP [61]	22.9
4 Visual	Template Matching + IP Rot.	29.9
5 Geo	Feature Matching	49.1
6 Geo + HGM	Feature Matching	70.4

Effect of over/under-segmentation filtering. We also conducted experiments to analyze the effectiveness of the scoring and filtering strategies to filter over/under-segmentation. We found that 88% of the under-segmentation masks and 74% of the over-segmentation masks can be correctly filtered out.

Comparison of viewpoint prediction strategies. To evaluate the effect of viewpoint prediction strategies, we utilize the metric of cosine degree (range from 0 to 180) between the predicted viewpoint direction and the ground truth viewpoint direction. Compared with the image embedding matching in the discovery stage, the proposed viewpoint prediction from the patch embedding matching strategy reduces the degree of viewpoint direction error from 57.1 to 42 degrees, as shown in Table IV. Furthermore, when selecting the top 5 candidate viewpoints, the error decreases to 20.4 degrees, approaching the minimal angle error of viewpoint direction in the templates (15.9 degrees). This demonstrates the effectiveness of the proposed camera observation viewpoint prediction.

Considering that the global feature may be sensitive to occlusions, we leverage the patch-level feature to revise the score of predicted instances. As shown in Table VII, the score revision method leads to a mean performance gain of 3.1% in mAP over 5 BOP datasets, demonstrating its effectiveness.

Comparison of pose estimation strategies. To compare

TABLE VI

COMPARISON OF POSE ESTIMATION AT DIFFERENT VIEWPOINT-FILTERED CAD MODEL POINTS. W/O VP IS THE CAD MODEL WITHOUT FILTERING FROM THE PREDICTED VIEWPOINT. TOP N REPRESENTS ADOPTING THE HIGHEST N SCORE OF PREDICTED VIEWPOINTS TO FILTER POINT CLOUDS.

Setting	LM-O	T-LESS	TUD-L	IC-BIN	YCB-V	Mean	Time (s)
W/O VP	50.2	37.9	86.8	51.0	65.3	58.2	0.058
TOP1	53.0	44.3	81.4	44.0	63.1	57.2	0.053
TOP2	61.7	55.1	89.3	54.2	72.3	66.5	0.080
TOP3	64.6	59.3	90.8	56.3	73.9	69.0	0.108
TOP4	66.1	61.5	91.9	56.9	73.4	70.0	0.138
TOP5	67.1	62.6	92.7	56.8	73.0	70.4	0.165
TOP6	67.5	63.7	93.2	56.6	72.8	70.8	0.191

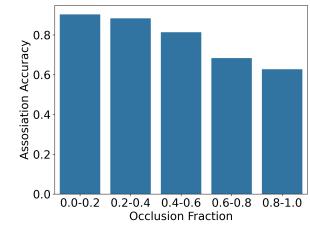
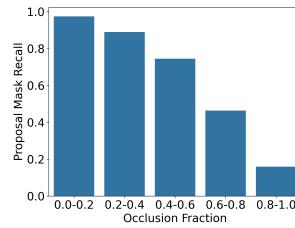


Fig. 9. Proposal object-agnostic in- Fig. 10. Accuracy of association pro- stance segmentation masks recall at posal masks and templates at different instance visibility fractions.

the effect of the different pose estimation strategies, we implement four distinct pose estimation strategies. The first strategy utilizes a visual model to extract patch embeddings for feature matching, subsequently estimating the pose through the RANSAC-Perspective-n-Points algorithm [64]. As delineated in Section III-D, this strategy is unreliable in the object pose estimation scenario, showing a mere 16.3% average AR across the BOP 5 datasets. The second strategy adopts the template pose from the Orientation step as rotation and estimates the translation component, as described in [47], for the final pose estimation. Based on the second strategy, the third and fourth strategies further introduce an ICP algorithm and templates across different in-plane rotations (ten samples with a density of 36 degrees), thereby generating a candidate rotation for each template. Since the difference between the predefined 2 DoF viewpoint and 6 DoF camera perspective, the viewpoint prediction is unable to perniciously predict the object pose.

When instead of the visual model in the proposed Geo model, the feature matching by geometric embedding results in an obvious improvement in performance relative to the visual patch embedding feature matching paradigm. Furthermore,

TABLE VII
INSTANCE SEGMENTATION MAP USING THE SCORE FROM DIFFERENT STEPS OVER FIVE BOP DATASETS.

Step	LM-O	T-LESS	TUD-L	IC-BIN	YCB-V	Mean
Discovery	37.7	34.7	46.0	20.6	57.4	37.8
Orientation	40.0	37.1	47.2	22.8	57.4	40.9

TABLE VIII
COMPARISON OF POINT CLOUD INPUT AND REPRESENTATION. VIEWPOINT FILTERED REPRESENTS THE CAD MODEL POINT CLOUD FILTERED BY PREDICTED CAMERA OBSERVATION VIEWPOINT. NORMALIZED REP. REPRESENTS THE POINT CLOUD THAT IS SCALED INTO THE NORMALIZED REPRESENTATION. AR IS THE AVERAGE ON THE BOP5 DATASETS.

Viewpoint	Filtered	Normalized Rep.	AR (%)
1	-	-	48.3
2	-	✓	58.2
3	✓	-	65.1
4	✓	✓	70.4

the integration of a hierarchical geometric matching model enhances the reliability of the matching process, leading to an additional performance gain.

Effect of viewpoint-filtered CAD model point cloud and normalized representation. We compare the CAD model point cloud with or without filtering the self-occlusion region by viewpoint prediction. As illustrated in Figure 8, there are many ambiguous regions in the CAD model leading to mismatching between the points in the scene instances and the ambiguous self-occlusion regions in the CAD model. The estimated camera viewpoint effectively finds the points of the CAD model corresponding to the visible points of the scene instance and filters out the remaining points of the CAD model, improving the matching accuracy.

Moreover, we quantitatively analyze the impact of the viewpoint-filtered CAD model point cloud on pose estimation in Table VI. The AR performance markedly increases when adopting more than two candidate viewpoint-filtered point clouds. For the number of viewpoint candidates, although fewer viewpoint-filtered point clouds result in faster running time, it is sensitive to the viewpoint prediction result. More candidate point clouds can improve the robustness of inaccurate viewpoint prediction and advanced performance gain. To balance speed and accuracy, we choose the top 5 candidates for viewpoint-filtered point cloud in this paper.

To validate the effectiveness of normalized representation point clouds, we compared scene and object point clouds with or without the proposed normalization. As demonstrated in Table VIII, the normalized representation point clouds can enhance the performance whether under the completed point clouds (+ 9.9%) or viewpoint-filtered point clouds (+ 5.1%), verifying the effectiveness of normalized representation.

V. CONCLUSION

In this paper, we propose a universal framework ZeroPose, which solves both object discovery and pose estimation in a zero-shot manner without additional human interaction or presupposing scene conditions. It performs pose estimation in

a novel Discovery-Orientation-Registration (DOR) inference pipeline through stepwise feature matching across three distinct inference steps. For limitation, the three steps in the DOR pipeline are independent and can not be trained end-to-end. An end-to-end pipeline can learn the mapping from input data to pose estimation output directly, simplifying the inference pipeline. However, since the end-to-end pipeline is fully data-driven, there is a lack of large volumes of labeled real data for training. For further work, a large-scale real-world pose estimation dataset with large volumes of labeled data is valuable for improving the performance and generalization of the model. Additionally, exploring multi-task learning frameworks that integrate related object discovery tasks including object detection, tracking, instance segmentation, and pose estimation is a promising research topic.

REFERENCES

- [1] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, “Densefusion: 6d object pose estimation by iterative dense fusion,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Jun. 2019, p. 3343–3352.
- [2] 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 (RSS)*, Jun. 2018, p. 110.
- [3] T. Hodaň, M. Sundermeyer, B. Drost, Y. Labb  , E. Brachmann, F. Michel, C. Rother, and J. Matas, “Bop challenge 2020 on 6d object localization,” in *Computer Vision-ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II* 16, Aug. 2020, pp. 577–594.
- [4] J. Chen, M. Sun, Y. Zheng, T. Bao, Z. He, D. Li, G. Jin, R. Zhao, L. Wu, and X. Jiang, “Geo6d: Geometric constraints learning for 6d pose estimation,” *arXiv preprint arXiv:2210.10959*, pp. 1–8, Oct. 2022.
- [5] G. Wang, F. Manhardt, X. Liu, X. Ji, and F. Tombari, “Occlusion-aware self-supervised monocular 6d object pose estimation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 3, pp. 1788–1803, Mar. 2024.
- [6] G. Feng, T.-B. Xu, F. Liu, M. Liu, and Z. Wei, “Nvr-net: Normal vector guided regression network for disentangled 6d pose estimation,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 2, pp. 1098–1113, Feb. 2024.
- [7] T. Cao, W. Zhang, Y. Fu, S. Zheng, F. Luo, and C. Xiao, “Dgecn++: A depth-guided edge convolutional network for end-to-end 6d pose estimation via attention mechanism,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, Oct. 2023.
- [8] B. Okorn, Q. Gu, M. Hebert, and D. Held, “Zephyr: Zero-shot pose hypothesis rating,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, Jun. 2021, pp. 14 141–14 148.
- [9] K. Park, A. Mousavian, Y. Xiang, and D. Fox, “Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Aug. 2020, pp. 10 707–10 716.
- [10] Y. Labb  , L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpenter, M. Aubry, D. Fox, and J. Sivic, “MegaPose: 6D Pose Estimation of Novel Objects via Render & Compare,” in *CoRL*, Dec. 2022.
- [11] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, “Unseen object instance segmentation for robotic environments,” *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1343–1359, Mar. 2021.
- [12] D. Cai, J. Heikkil  , and E. Rahtu, “Ove6d: Object viewpoint encoding for depth-based 6d object pose estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp. 6803–6813.
- [13] I. Shugurov, F. Li, B. Busam, and S. Ilic, “Osop: A multi-stage one shot object pose estimation framework,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 6825–6834.
- [14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, “Segment anything,” *arXiv preprint arXiv:2304.02643*, pp. 1–30, Apr. 2023.
- [15] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang, “Fast segment anything,” *arXiv preprint arXiv:2306.12156*, pp. 1–11, Jun. 2023.

- [16] Y. Su, M. Saleh, T. Fetzer, J. Rambach, N. Navab, B. Busam, D. Stricker, and F. Tombari, "Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp. 6738–6748.
- [17] R. L. Hauggaard and A. G. Buch, "Surfemb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp. 6749–6758.
- [18] Y. Zheng, J. Wu, Y. Qin, F. Zhang, and L. Cui, "Zero-shot instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Jun. 2021, pp. 2593–2602.
- [19] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, pp. 1–11, Jan. 2024.
- [20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, Sep. 2014, pp. 740–755.
- [21] Y. Xiang, C. Xie, A. Mousavian, and D. Fox, "Learning rgb-d feature embeddings for unseen object instance segmentation," in *Conference on Robot Learning*, Nov. 2021, pp. 461–470.
- [22] E. P. Örnek, A. K. Krishnan, S. Gayaka, C.-H. Kuo, A. Sen, N. Navab, and F. Tombari, "Supergbd: Zero-shot instance segmentation in cluttered indoor environments," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3709–3716, Apr. 2023.
- [23] Y. Lin, J. Tremblay, S. Tyree, P. A. Vela, and S. Birchfield, "Single-stage keypoint-based category-level object pose estimation from an rgb image," in *2022 International Conference on Robotics and Automation (ICRA)*, May 2022, pp. 1547–1553.
- [24] 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*, Jun. 2019, pp. 2642–2651.
- [25] J. Liu, Z. Cao, Y. Tang, X. Liu, and M. Tan, "Category-level 6d object pose estimation with structure encoder and reasoning attention," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6728–6740, Apr. 2022.
- [26] 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 international conference on computer vision*, Jun. 2022.
- [27] 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*, Nov. 2022, pp. 35 103–35 115.
- [28] 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 international conference on computer vision*, Jun. 2022, pp. 6814–6824.
- [29] L. Jin, X. Wang, X. Nie, L. Liu, Y. Guo, and J. Zhao, "Grouping by center: Predicting centripetal offsets for the bottom-up human pose estimation," *IEEE Transactions on Multimedia*, vol. 25, pp. 3364–3374, 2023.
- [30] J. Zhao, J. Li, Y. Cheng, T. Sim, S. Yan, and J. Feng, "Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing," in *Proceedings of the 26th ACM International Conference on Multimedia*, Oct. 2018, p. 792–800.
- [31] T. Wang, L. Jin, Z. Wang, X. Fan, Y. Cheng, Y. Teng, J. Xing, and J. Zhao, "Decenternet: Bottom-up human pose estimation via decentralized pose representation," in *Proceedings of the 31st ACM International Conference on Multimedia*, Oct. 2023, p. 1798–1808.
- [32] G. Zhou, D. Wang, Y. Yan, H. Chen, and Q. Chen, "Semi-supervised 6d object pose estimation without using real annotations," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5163–5174, Dec. 2022.
- [33] M. Sun, Y. Zheng, T. Bao, J. Chen, G. Jin, L. Wu, R. Zhao, and X. Jiang, "Unid6v2: Noise elimination for 6d pose estimation," in *International Conference on Artificial Intelligence and Statistics*, Apr. 2023, pp. 1832–1844.
- [34] J. Liu, W. Sun, C. Liu, X. Zhang, S. Fan, and W. Wu, "Hff6d: Hierarchical feature fusion network for robust 6d object pose tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7719–7731, Jun. 2022.
- [35] T. Hodan, P. Halužík, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis, "T-less: An rgb-d dataset for 6d pose estimation of texture-less objects," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2017, pp. 880–888.
- [36] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3d object recognition," in *2010 IEEE computer society conference on computer vision and pattern recognition*, Jun. 2010, pp. 998–1005.
- [37] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, S. Ilic, D. Hu, and K. Xu, "Geotransformer: Fast and robust point cloud registration with geometric transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9806–9821, Mar. 2023.
- [38] Y. Wu, X. Hu, Y. Zhang, M. Gong, W. Ma, and Q. Miao, "Sacf-net: Skip-attention based correspondence filtering network for point cloud registration," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 3585–3595, Jan. 2023.
- [39] X. Huang, J. Zhang, Q. Wu, L. Fan, and C. Yuan, "A coarse-to-fine algorithm for matching and registration in 3d cross-source point clouds," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2965–2977, Jul. 2018.
- [40] Z. Dang, L. Wang, Y. Guo, and M. Salzmann, "Learning-based point cloud registration for 6d object pose estimation in the real world," in *European conference on computer vision*, Oct. 2022, pp. 19–37.
- [41] V. N. Nguyen, T. Groueix, G. Poniatkin, 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*, Sep. 2023, pp. 2134–2140.
- [42] J. Lin, L. Liu, D. Lu, and K. Jia, "Sam-6d: Segment anything model meets zero-shot 6d object pose estimation," *arXiv preprint arXiv:2311.15707*, pp. 1–20, Nov. 2023.
- [43] A. Caraffa, D. Boscaini, A. Hamza, and F. Poiesi, "Object 6d pose estimation meets zero-shot learning," *arXiv preprint arXiv:2312.00947*, pp. 1–22, Dec. 2023.
- [44] M. Oquab, T. Darcret, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," *arXiv:2304.07193*, p. 32, Apr. 2023.
- [45] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2023, pp. 15 180–15 190.
- [46] R. Xu, X. Wang, T. Wang, Y. Chen, J. Pang, and D. Lin, "Pointilm: Empowering large language models to understand point clouds," *arXiv preprint arXiv:2308.16911*, pp. 1–14, Aug. 2023.
- [47] 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 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022.
- [48] A. Lin, J. Y. Zhang, D. Ramanan, and S. Tulsiani, "Relpose++: Recovering 6d poses from sparse-view observations," *arXiv preprint arXiv:2305.04926*, pp. 1–17, Mar. 2023.
- [49] C. Zhao, T. Zhang, Z. Dang, and M. Salzmann, "Dvmnet: Computing relative pose for unseen objects beyond hypotheses," *arXiv preprint arXiv:2403.13683*, pp. 1–13, Mar. 2024.
- [50] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," in *Proceedings of the IEEE/CVF international conference on computer vision*, Jun. 2019, pp. 6411–6420.
- [51] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611, Feb. 1992, pp. 586–606.
- [52] M. A. Butt and P. Maragos, "Optimum design of chamfer distance transforms," *IEEE Transactions on Image Processing*, vol. 7, no. 10, pp. 1477–1484, Oct. 1998.
- [53] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, Oct. 2017, pp. 2961–2969.
- [54] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, "Cosopose: Consistent multi-view multi-object 6d pose estimation," in *European Conference on Computer Vision*, Aug. 2020, pp. 574–591.
- [55] Z. Li, G. Wang, and X. Ji, "Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Jun. 2019, pp. 7678–7687.
- [56] L. Lipson, Z. Teed, A. Goyal, and J. Deng, "Coupled iterative refinement for 6d multi-object pose estimation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 6718–6727.

- [57] 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*, Jun. 2023, pp. 14 045–14 054.
- [58] 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 *2022 International Conference on Robotics and Automation (ICRA)*, May 2022, pp. 2553–2560.
- [59] 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," *arXiv preprint arXiv:2403.09799*, pp. 1–10, Mar. 2024.
- [60] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, pp. 1–22, Oct. 2020.
- [61] P. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.
- [62] M. Denninger, M. Sundermeyer, D. Winkelbauer, Y. Zidan, D. Olefir, M. Elbadrawy, A. Lodhi, and H. Katam, "Blenderproc," *arXiv preprint arXiv:1911.01911*, pp. 1–7, Nov. 2019.
- [63] Pyrender, "<https://github.com/mmatl/pyrender>," Jan. 2019. [Online]. Available: <https://github.com/mmatl/pyrender>
- [64] M. A. Fischler and R. C. Bolles, "Random sample consensus," *Communications of the ACM*, p. 381–395, Jun. 1981.



Rui Zhao received the B.S. degree from the University of Science and Technology of China in 2010 and the Ph.D. degree in electronic engineering from The Chinese University of Hong Kong in 2015. He joined a startup venture called SenseNets as the CTO right after postgraduate study. In 2018, he joined SenseTime Research, Shenzhen, Guangdong, China, as a Research Director, and has been the Head of the Smart City Group, Research and Development Department, SenseTime, since 2021. While in SenseTime Research, he led a team designing

and developing competitive deep learning models and techniques, which are applied to products for smart city applications including public services, transportation, epidemic control, and smart manufacturing. He is currently an Adjunct Researcher at the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, the Tsinghua Shenzhen International Graduate School, and the Qing Yuan Research Institute, Shanghai Jiao Tong University. His research interests span a range of topics in computer vision and deep learning, including face recognition, person re-identification, large-scale clustering, unsupervised/self-supervised learning, few-shot/zero-shot learning, and visual-language foundation models. He has published more than 60 technical papers and book chapters on these topics.



Jianqiu Chen received the M.S. degree from University of New South Wales in 2021. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen. His research interests include object pose estimation and zero-shot learning.



Liwei Wu received the B.S. degree in Automation from Nanjing University, Nanjing, China, in 2013 and the M.S. degree in Automation from Tsinghua University, Beijing, China, in 2016. He is currently working as a researcher in SenseTime Research. His research interests include computer vision and machine learning.



Zikun Zhou received his Ph.D. and Master's degree from Harbin Institute of Technology in 2022 and 2018, respectively. He is currently an assistant research fellow with Pengcheng Laboratory. His research interests include computer vision and machine learning.



Tianpeng Bao received the B.S. degree in Automation from Tsinghua University, Beijing, China, in 2010 and the M.S. degree in Control Science and Engineering from Tsinghua University, Beijing, China, in 2014. He is currently working as a researcher in SenseTime Research. His research interests include computer vision and LLM-based agents.



Mingshan Sun received the B.S. degree in Electronic Commerce from Guangxi University, Nanning, China, in 2017, and the M.S. degree in computer technology from the Harbin Institute of Technology, Shenzhen, China, in 2020. She is currently working at SenseTime Research. Her research interests include computer vision and machine learning.



Zhenyu He received his Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 2007. From 2007 to 2009, he worked as a postdoctoral researcher in the department of Computer Science and Engineering, Hong Kong University of Science and Technology. He is currently a full professor in the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. His research interests include machine learning, computer vision, image processing and pattern recognition.