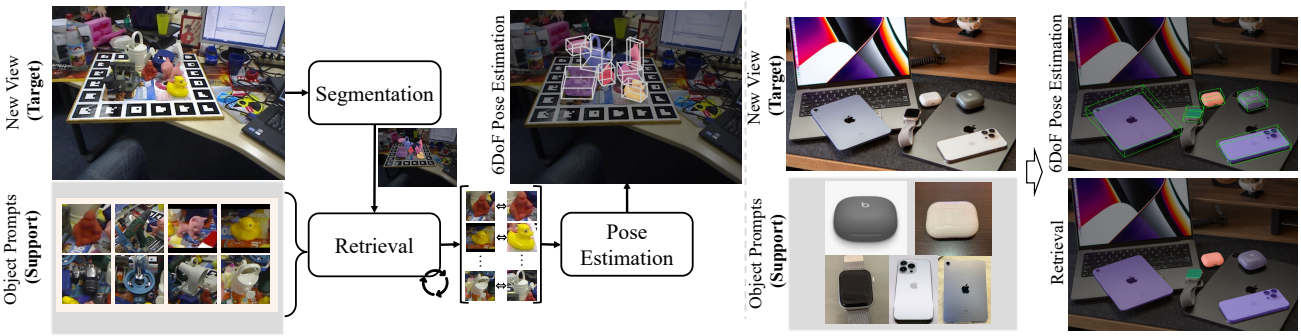
POPE: 6-DoF Promptable Pose Estimation of Any Object, in Any Scene, with One Reference

Zhiwen Fan*[2], Panwang Pan*†‡[1], Peihao Wang*[2], Yifan Jiang[2], Dejia Xu[2], Zhangyang Wang‡[2]

* Equal contribution † Project leader ‡ Corresponding author

[1]ByteDance [2]University of Texas at Austin



(a). Framework of ***Promptable Object Pose Estimator***. (b). 6DoF pose estimation of in-the-wild object.

Figure 1. Promptable object Pose Estimator (**POPE**) is a zero-shot object 6DoF pose estimation method, which predicts the relative pose between given object prompts (**support view**) and the object in any new view (**target view**). Our framework recognizes the object prompts in the target under any scene (as shown in a cluttered scene in (**a**)) and estimates the relative pose for any category with only one support image. POPE also exhibits capability on any object pose estimation (as shown in (**b**)).

Abstract

Despite the significant progress in six degrees-of-freedom (6DoF) object pose estimation, existing methods have limited applicability in real-world scenarios involving embodied agents and downstream 3D vision tasks. These limitations mainly come from the necessity of 3D models, closed-category detection, and a large number of densely annotated support views. To mitigate this issue, we propose a general paradigm for object pose estimation, called ***Promptable Object Pose Estimation (POPE)****. The proposed approach POPE enables zero-shot 6DoF object pose estimation for any target object in any scene, while only a single reference is adopted as the support view. To achieve this, POPE leverages the power of the pre-trained large-scale 2D foundation model, employs a framework with hierarchical feature representation and 3D geometry principles. Moreover, it estimates the relative camera pose between object prompts and the target object in new views, enabling both two-view and multi-view 6DoF pose estimation tasks. Comprehensive experimental results demonstrate that POPE exhibits unrivaled robust performance in zero-shot settings, by achieving a significant reduction in the averaged Median Pose Error by* ***52.38%*** *and* ***50.47%*** *on the LINEMOD [22] and OnePose [54] datasets, respectively. We also conduct more challenging testings in causally captured images (see Figure 1), which further demonstrates the robustness of POPE.*

1. Introduction

Robotic systems and augmented reality/virtual reality (AR/VR) applications have become ubiquitous across numerous industries, facilitating the execution of intricate tasks ot offering immersive user experiences. Describing the status of objects, particularly their six degrees-of-freedom (6DoF) poses, is a crucial step towards achieving in-depth scene understanding and delicate interactions. More importantly, given the diverse nature of real-world scenarios, it is essential to have a method that can operate on arbitrary object assets.

However, enabling object 6DoF pose estimation on unseen objects using simple and easy-to-obtained references is challenging. Traditional instance-level [28, 30, 33, 46, 56,

62, 66] or category-level [1, 12, 59] pose estimators exhibit limitations in handling diverse objects, as they are specifically designed for particular instances or categories. These design principles restrict their generalization capabilities to unseen instances or categories during testing, due to their reliance on CAD models or a well-defined category-level canonical space. Later, tremendous efforts have been devoted to addressing the aforementioned challenges by adopting structure-from-motion (SfM [51]) techniques [20, 54], reducing the number of support views [38], or leveraging depth maps and self-supervised trained Vision Transformers [19]. A detailed visual comparison is summarized in Figure 2.

A straightforward way to accomplish 6DoF object pose estimation with a single support view is to estimate relative poses [52, 67] by performing 2D-2D matching between query and reference images. However, dense matching on arbitrary objects is highly unstable, especially for wide-baseline camera views or a clustered background. Besides the difficulties in image matching, another substantial issue in real-world scenes arises from the potential for heavy occlusion of the target object, which makes it hard to be detected. Previous methods propose to adopt off-the-shelf detectors [27] for specific instances/categories, or design a correlation-based object detector on a small scale dataset [38]. Consequently, their robustness when dealing with novel objects in diverse scenes are not guaranteed.

To tackle the obstacles of open-world detection on arbitrary target objects and the robust 2D-2D matching, a promising avenue is leveraging the power of the foundation model that is trained on a vastly large-scale dataset. Recently, the community has witnessed the emerging properties of these foundation models on few-shot or even zero-shot generalization, crossing from language [5, 15] to vision [9, 29, 44].

These advancements have shed light on the under-explored problem of zero-shot object pose estimation - the tantalizing possibility of making no assumption on the object category and using only one reference image. Specifically, the newly arising capability of performing zero-shot segmentation across various image data domains [29] and non-parametric instance-level recognition [44] have shown potential in addressing these challenges.

In this paper, we introduce a novel task named ***Promptable Object Pose Estimation*** to tackle the challenge of estimating the 6DoF object pose between the given object prompts (a single image for each instance, used as support) and any new captured viewpoint with complex backgrounds (target). Our proposed model, called ***POPE***, consists of four main features in one unified pipeline: (i) *Segment Objects* generates a set of valid segmentation proposals for any image at a new viewpoint; (ii) *Retrieve Objects* constructs object-level matching between object prompts and segmented object proposals at two views; (iii) *Pose Objects* estimate the relative pose by utilizing the matched correspondences between paired object images; and (iv) *Online Refinement for Arbitrary View-number* triggers a coarse-to-fine pose estimation process with efficient 2D-2D global matching and 2D-3D local matching for retrieved objects, on newly target views. We outline our contributions below:

- We establish a new and challenging task: ***Promptable Object Pose Estimation***, which aims to estimate the pose of an object in wild scenarios, with no assumptions on object category and using only one reference image.
- To tackle this problem, we propose a 6DoF pose foundation model, **POPE**, that seamlessly integrates the power of pre-trained foundation models and 3D geometry principles for high-quality segmentation, hierarchical object retrieval, and robust image matching, to enable accurate object pose estimation in diverse and uncontrolled environments.
- For evaluation, we introduce a large-scale test dataset containing diverse data sources. POPE outperforms existing generalizable pose estimators and demonstrates remarkable effectiveness for both promptable pose estimation and downstream 3D-vision tasks.

2. Related Works

Large-scale Pre-trained 2D Foundation Models. Models trained on large-scale datasets, demonstrating the scaling effect with data-parameter balance, are regarded as foundation models. Recently, we witnessed that the foundation models [5] demonstrated strong generalization capability, serving as the base model in a wide range of tasks [5]. For example, CLIP [48] utilizes contrastive learning to construct a joint embedding space of text and image modalities. Similarly, self-supervised models such as DINO [9] and DINOv2 [44] show emerging properties for learning robust visual features. Segment-Anything Model (SAM) [29] demonstrates promptable segmentation ability that supports interactive segmentation with visual promptings such as points and bounding boxes. In this paper, we achieve the goal of promptable object pose estimation by harnessing the power of foundation models. We build a system that integrates the essence of SAM and DINO to help POPE handle cluttered real scenes by performing dense segmentation and instance-level matching.

Generalizable Object Pose Estimator. Early approaches for estimating the 6DoF pose of objects build instance-level [16, 23, 65] or category-level [10, 12, 12, 14, 17, 19, 34, 35, 57, 59, 60] frameworks. They usually require perfect instance-specific CAD models or well-established canonical space of specific categories. Thus, the methods only work on specific instances and categories. They cannot generalize to novel instances/categories that are unseen during training. The recent advances in generalizable object pose estimators can be divided into two categories based on whether a

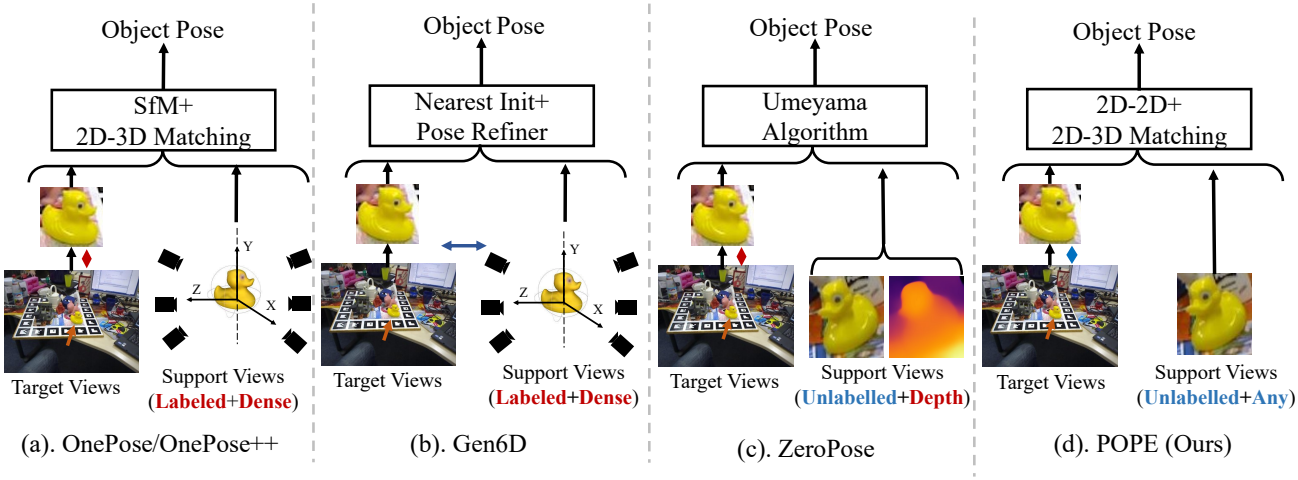


Figure 2. **Comparing POPE with previous frameworks.** We provide a detailed comparison between POPE and prior works, including (a) OnePose/OnePose++ [20, 54] which relies on a large number of posed support views and the corresponding bounding box; (b) Gen6D [38] that replace 2D-3D matching pipeline with a refiner network; and (c) ZeroPose [19] which further utilized depth maps. Different from all these methods, the proposed method POPE eliminates the need for densely annotated support views and enables accurate object retrieval in new viewpoints without relying on any assumptions about the object's category. Here, ♦ denotes a close-category detector, ↔ means a correlation-based detector, and ♦ denotes a open-world detector.

3D model is utilized. One line of work adopts high-quality 3D objects through shape embedding [13, 47, 63], template matching [2, 21, 55, 61] and rendering-and-comparison approaches [6, 32, 43, 66]. The other approaches instead aim to avoid the need of 3D objects and utilize depth map [45], object mask [36, 45, 64] and reference images [20, 38, 54]. Specifically, Gen6D [38] first detects the target object and initializes a pose estimate from dense reference views. Then, Gen6D refines the pose using feature volume and a 3D neural network. OnePose [54] and OnePose++ [20] construct a sparse point cloud from the RGB sequences of all support viewpoints and then determine the object poses by matching the target view with the sparse point cloud. However, these works still require dense support views, i.e. $\geq$32 views, where each view needs to be annotated with ground-truth poses. We argue the requirement of dense support views is not practical for real-world applications. To this end, we propose the paradigm of promptable pose estimation, where we only use one support view as the reference. We turn the 6DoF object pose estimation task into relative pose estimation between the retrieved object in the target view and the support view. Thus, we do not have any hypothesis of object category, achieving generalizable object pose estimation.

Two-view Object Pose Estimation. The methods of estimating the relative camera pose between two views can be classified into two categories: i) correspondence-based methods, and ii) direct pose regression methods. The correspondence-based methods establish cross-view pixel-level correspondences, and the pose can be recovered by solving the fundamental matrix [41]. The methods establish the correspondences based on hand-crafted features, e.g. SIFT [40], and SURF [3], or using learned features [18, 25, 31, 37, 50, 53]. Some of the methods also incorporate robust estimation methods [4], or the synergy between shape reconstruction and pose estimation [26]. Another category of methods learns cues for pose estimation in an end-to-end manner [7, 24, 49, 67]. For example, RelPose [67] builds an energy-based framework for handling pose ambiguity. The 8-Point Transformer [49] incorporates the inductive bias of the 8-point algorithm into transformer designs. FORGE [24] leverages 3D feature volumes to alleviate the ambiguity of learning on 2D features. In this work, we stick to the classic correspondence-based method because of its better generalization ability on novel instances/categories. Different from prior works establishing image-level correspondence (matching the support image with the entire target image) [50, 53], we propose a coarse-to-fine paradigm. We first build instance-level correspondence by matching the prompt object (shown in the support image) with segmented object instances in the target image, which identifies the highly possible regions of the prompt object. Then we establish fine-grained dense correspondence between the support image and the identified regions in the target image, which avoids noisy matching with cluttered background regions.

3. Propmtable Object Pose Estimation Task

Generalizable 6DoF object pose estimators play a crucial role in robotics and 3D vision tasks by accurately determining the position and orientation of novel objects in 3D space, without the need for fine-tuning.

However, current methods [19, 20, 38, 54, 64] have limitations. They can only handle cases where an off-the-shelf

detector is used for closed-category object separation from the background [19, 20, 54]. Additionally, the number of support views required for a robotics system to grasp an object is often uncertain due to occlusions, object appearance variations, and sensor limitations [64]. Furthermore, the tedious requirement of pose annotation [20, 38, 54] or depth maps [19] in the support view makes it challenging to scale up and generalize to various scenes. These limitations hinder the deployment of existing pose estimators in diverse and uncontrolled scenes. To address these challenges, we propose to decompose the 6DoF object pose estimation problem into relative object pose estimation. This approach reduces the reliance on absolute pose annotation and allows for easy extension from two-view to multiple-view scenarios. Moreover, we introduce an *Open-world Detector* that is category-agnostic and robust to occlusion and pose variation.

3.1. Task Definition

We introduce a novel task of ***Propmtable Object Pose Estimation (POPE)***. The primary goal of this task is to estimate the relative poses for all objects in one scene image according to a series of (single-view) reference image prompts. Specifically, our POPE model receives an arbitrary scene image and a sequence of arbitrary reference images as the input. As the output, POPE simultaneously detects all the objects from the scene and annotates their poses according to the references.

Why Promptable? The use of object prompts allows for higher interactivity and flexibility, enabling end users to indicate their interest in specific objects through prompts such as object images or even abstract sketches. The promptable setting eliminates the reliance on predefined categories or assumptions regarding the size and shape of objects, resulting in a more generalizable approach that can be applied to any object as long as it is included in the set of object prompts.

Why Single-View Prompt? We argue that in most user cases, only single-image references are presented and preferred. On the one hand, consistent images captured for the same object from different angles barely exist in the wild and web collection. On the other hand, estimating 6DoF pose with multiple views requires additional calibration of the reference views resulting in a chicken-egg problem. Enabling high-performance two-view geometry also frees the robotic agent from acquiring a CAD model and benefits 3D reconstruction with fewer views. Despite estimating the poses through only one reference view being a challenging setting, fortunately, it can be endowed with the prevalent foundation models which enable robust feature representation for both detection and matching. In addition, single-reference pose estimation can be served as a starting point for multi-view geometry. Our POPE pipeline can be seamlessly integrated into a multi-view progressive reconstruction pipeline, which consistently boosts pose estimation and reconstruction accuracy starting with a set of unposed images from scratch.

3.2. Preliminary of Two-view Pose Estimation.

The task of estimating the relative camera poses from two separate images, without a 3D CAD model, is referred to as two-view object pose estimation. Classic geometric vision theory suggests that the camera poses and depth maps can be computed from image matching points alone, without any additional information [39].

Given a set of image matching points $\mathbf{x}_i$ and $\mathbf{x}'_i$ in homogeneous coordinates, along with a known camera intrinsic matrix $\mathbf{K}$, the task of two-view object pose estimation is to find the camera rotation matrix $\mathbf{R}$, translation vector $\mathbf{t}$, and corresponding 3D homogeneous point $\mathbf{X}_i$. The goal is to satisfy the equations $\mathbf{x}_i = \mathbf{K}\left[\mathbf{I}|\mathbf{0}\right]\mathbf{X}_i$ and $\mathbf{x}'_i = \mathbf{K}\left[\mathbf{R}|\mathbf{t}\right]\mathbf{X}_i$ for all i. A classical method to solve this problem consists of three steps: computing the essential matrix $\mathbf{E}$ from the image matching points, extracting the relative camera pose $\mathbf{R}$ and $\mathbf{t}$ from $\mathbf{E}$, and triangulating the matching points to get $\mathbf{X}_i$. The essential matrix can be solved using at least 5 matching points [42], and $\mathbf{R}$ and $\mathbf{t}$ can be computed from $\mathbf{E}$ using matrix decomposition. There is a scale ambiguity for relative camera pose estimation, and the 3D point $\mathbf{X}_i$ can be computed with a global scale ambiguity.

3.3. Modular Approach to Zero-shot Promptable Object Pose Estimation

Directly applying a two-view image matching framework between a prompt image and a complex target containing the same object is prone to failure. This is because a complex scene can have numerous noisy matches, especially when limited to only two observations. Hence, in this paper, we propose a modular approach to address this problem by breaking it down into multiple steps. First, we formulate an *Open-world Detector* that segments and identifies the queried object prompts in the target image. Next, we establish correspondences with new views, refining incorrect object retrievals and solving the task of relative pose estimation.

Open-world Object Detector. In this paper, we propose a robust and general detector that conditions on the user-provided object prompt image I_P and the image in the target view I_T, without making any assumptions about object categories. The proposed detector aims to obtain the matched object mask in the target view, by generating all K valid masks $\mathcal{M} = \{m^1, m^2, ..., m^K\}$ within I_T using automatic object mask generation from a segmentation model[29], and retrieving the masked object image with the best global image properties. Specifically, we generate densely uniform points on the image lattice as prompts for promptable segmentation model (SAM) [29] to obtain $\mathcal{M}$, which represents

Object Prompt (**Support**) | Proposal-1 (**Target**) | Proposal-2 (**Target**) | Proposal-3 (**Target**)

Figure 3. **Failed matches.** Relying solely on the similarity score of the [CLS] token for global representation can lead to inaccurate matches, especially in clustered scenarios. This motivates us to incorporate local descriptor information for improved retrieval.

the object segments. The next goal is to retrieve the masked object image in the target view I_T by establishing the relationship between the object prompt image I_P and the masked object image set $\mathcal{I}_T^K = \{I_T^1, I_T^2, ..., I_T^K\}$, given one object prompt image I_P and K object segments in the target. However, we cannot guarantee that the image pairs have enough texture [11] or sufficient image content overlapping for local feature matching of the open-world objects. Inspired by recent progress in self-supervised pre-trained Vision Transformer (ViT) models [9], we employ the retrieval augmented data engine in the DINO-v2 model [44] to perform robust global retrieval. Here, we utilize the embedded [CLS] token to capture global image properties and construct a cosine similarity matrix of shape $1 \times K$ via the inner product of the [CLS] tokens: $\mathcal{S}(P, T, k) = \langle CLS_P, CLS_T(k) \rangle$, which reveals the object relationship between the prompt image I_P and the k_{th} masked image in set $\mathcal{I}_T^K$. By finding the highest score within the matrix, we retrieve the matched image of the same object in two views. Moreover, extending from a single prompt image to multiple ones (e.g., M) can be easily achieved by scaling up the similarity matrix to $M \times K$.

Hierarchical Retrieval Refinement with Local Descriptors. However, despite being trained on a large-scale dataset, DINO-v2 may generate high similarity scores for objects with similar appearances, resulting in erroneous global object-level retrieval (last column of Figure 3). This, in turn, can negatively impact the accuracy of the pose estimation stage. In order to address this issue, we propose a fine-grained approach that incorporates local descriptors to enhance the retrieval process and provide more reliable object matches. Specifically, we leverage local descriptors to summarize the similarities of local visual patterns, including edges, corners, and textures. These descriptors complement the potentially erroneous retrievals obtained solely from global representations. To implement this approach, we consider the Top-K proposals generated by DINO-v2, ranking the similarity scores in descending order. We then establish image correspondences using a transformer-based local feature estimation framework [53] when using natural RGB images as prompt. The predicted confidence matrix $\mathcal{P}_c$ represents the matching probabilities for all correspondences. To determine the confidence level of the matches, we introduce a confidence criterion based on a threshold value σ. We select and count the matches with confidence scores higher than the threshold in the total number of matches n. This criterion is defined as Criteria $= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(c_i \geq \sigma)$, where c_i denotes the confidence score of the i-th match, and $\mathbb{1}$ is the indicator function that returns 1 if its argument is true, and 0 otherwise. The proposal with the largest criteria score among the Top-K proposals is selected as the best-matched pair, providing a more reliable estimation of the object pose.

Pose Estimation. With dense correspondences established across the best-matched views, we proceed to estimate the relative pose of the cameras. This pose estimation involves determining the rotation $\mathbf{R} \in \mathrm{SO}(3)$ and the translation vector $\mathbf{t} \in \mathbb{R}^3$ by matching descriptors, computing the essential matrix, and applying RANSAC to handle outliers and ensure reliable results [42]. It is important to note that our method is capable of recovering the relative rotation accurately. However, the predicted translation is up-to-scale, similar to other relative pose estimator [19, 38]. This limitation arises from the fact that recovering the absolute translation (or object scale) is an ill-posed problem when only considering two views, as it is susceptible to scale-translation ambiguity. To address this issue, we employ the PnP algorithm and utilize the bounding box of the prompted object in the uncropped support view to recovering the scale of translation.

4. Experiments

We initially demonstrate our approach for achieving zero-shot 6DoF object pose estimation on four different datasets using a two-view scenario. Subsequently, we validate the proposed open-world detector by assessing its segmentation and retrieval accuracy. Finally, in order to adapt POPE to multi-view pose estimation and evaluate the accuracy of multiple-view pose, we visualize the performance using additional input target frames and assess the pose on the task of novel view synthesis.

4.1. Evaluation Setup

Datasets. We evaluate our method on four widely used 6DoF object pose estimation datasets, to test the zero-shot transferbility of **POPE** without any finetuning. **The LINEMOD Dataset [22]** is a standard benchmark dataset for 6DoF object pose estimation with the ground-truth CAD model. The LINEMOD dataset consists of images of thirteen low-textured objects under cluttered scenes and varying lighting conditions. **The YCB-Video Dataset [8]** consists of 92 RGBD videos of 21 YCB objects, with medium clustered background and ground-truth CAD model for evaluation. **The OnePose Dataset [54]** contains around 450 real-world

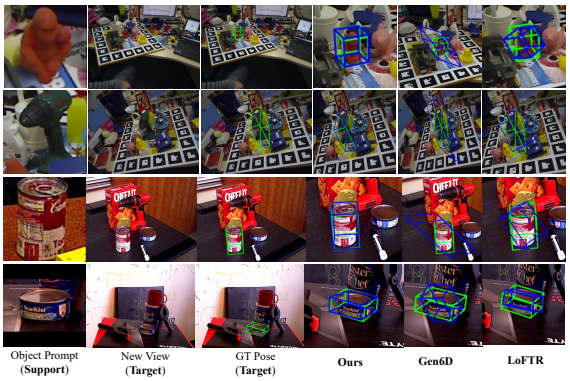
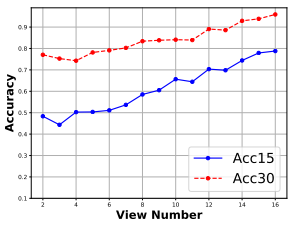
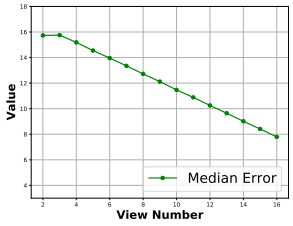


Figure 4. **Qualitative results on the LINEMOD [22] and YCB-Video [8] datasets.** Ground-truth poses are visualized with green boxes, while estimated poses are represented by blue boxes. Gen6D performs poorly compared to correspondence-based methods, as it relies on a closed initial view for relative pose estimation. LoFTR tends to produce noisy matching results when using the object prompt and new view directly as input. POPE demonstrates robustness against complex and cluttered backgrounds by employing open-world object detection and finding correspondences through cropped and centralized images.

video sequences of 150 objects with rich textures and simple background. Each frame is annotated with camera poses and 3D bounding box. **The OnePose++ Dataset [20]** supplements the original OnePose dataset with 40 household low-textured objects. As the pose distribution in different datasets varies, we organize the manage the test set in a balanced distribution from $0°$ to $30°$. Overall, the test set contains 5796 pairs on LIMEMOD, 2843 pairs on YCB-Video, 2751 pairs on OnePose, and 3166 pairs on OnePose++.



(a) The accuracy under 15°(Acc15) and 30°(Acc30).



(b) The Median Error with different view number.

Figure 5. We present plots illustrating the accuracy and median error as the number of views increases from 2 to 16.

Model Selection and Baselines. We compared our proposed POPE method with two other approaches: LoFTR [53], an image-matching based method that directly performs correspondence matching for pose estimation, and Gen6D [38], which utilizes a correlation-based network to discover object boxes, find pose initialization, and refine the relative object pose. We excluded the comparison with OnePose and OnePose++ as they are unable to generate point clouds from a single support view. In POPE, we utilize pre-trained models for different tasks: the *Segment Anything* model [29] with a ViT-H architecture for object mask generation, the *DINO-v2* model [44] pre-trained with ViT-S/14 for object proposal generation, and the *LoFTR* model [53] pre-trained with indoor scenes for natural image-based image matching. We set σ as 0.9 and the K as 3 in the experiments. It is important to note that the evaluated promptable object pose estimation does not rely on labeled examples for fine-tuning, including the pose in the support view and object masks, for any objects in real-world environments.

Evaluation. We report the median error for each pair of samples, along with the accuracy at $15°$ and $30°$, following the standard practice in relative object pose estimation [19]. The accuracy metrics represent the percentage of predictions with errors below these thresholds. In the main draft, our evaluation primarily focuses on the two-view settings, while we provide additional results on downstream applications (multiple-view pose estimation, novel view synthesis).

4.2. Comparisons

Results on LINEMOD and YCB-video datasets. We present the overall average median error and pose accuracy under different thresholds in Table 1. Due to space limitations, we include the full table in the supplementary materials and demonstrate the median error for five instances in this section. It is evident from the results that the proposed **POPE** consistently outperforms other methods across all metrics, exhibiting a significant margin over each instance. The qualitative results, visualized in Figure 4, highlight important observations. Gen6D [38] heavily relies on accurate initialization for pose refinement and struggles in single-reference scenarios. LoFTR [53] fails to provide accurate matches when handling clustered scenes with object occlusions, resulting in inaccurate box predictions. It is important to note that the visualization of object boxes incorporates ground-truth translation to address scale ambiguity.

Results on OnePose and OnePose++ datasets. In addition to the dataset containing multiple objects in cluttered scenes, we also evaluate the proposed framework on recently introduced one-shot object pose estimation datasets. Unlike previous approaches that rely on pose or box annotations, we conduct zero-shot two-view pose estimation without such annotations. The results in Table 1 demonstrate that **POPE** achieves a smaller median error in the relative object pose estimation task for both datasets. As the pose gap increases, LoFTR can improve its accuracy by utilizing the entire image for matching, incorporating more textural details from the background while still performing on par with our method. Visualizations are provided in Figure 6.

Scaling from 2-view to Multi-view Promptable Pose Estimation (POPE) To address the requirement for sparse-

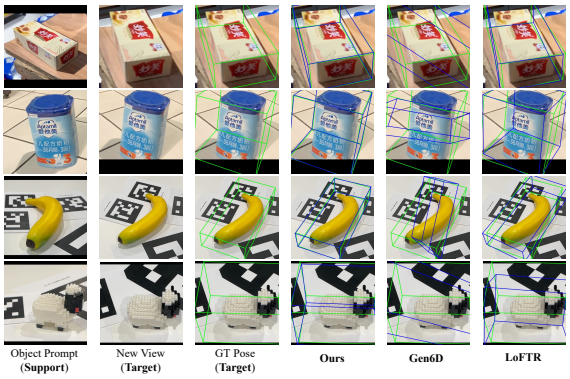


Figure 6. **Qualitative results on the OnePose [54] and OnePose++ [20] datasets.** Ground-truth poses are depicted with green boxes, while estimated poses are represented by blue boxes. Gen6D performs poorly compared to correspondence-based methods due to the significant pose gap between the support and target views. LoFTR is susceptible to the presence of similar patterns between the object and background (last row). In contrast, our proposed POPE exhibits strong generalization ability on both textured and textureless single object datasets.

view datasets in real-world scenarios, we have expanded our method from 2-view promptable pose estimation (POPE) to accommodate multi-view scenarios. Initially, we utilize the image matching results obtained from the 2-view POPE. We utilize the semi-dense correspondences from LOFTR [53] to reconstruct a semi-dense point cloud using COLMAP [51].

To introduce a new target viewpoint, we randomly select an image and perform object segmentation in a promptable manner. This enables us to retrieve the object's identity and exclude any negative effects caused by the clustered background. Subsequently, we conduct image matching between the prompt image and the newly added object image, register it, and extract correspondences between the new image and the semi-dense point cloud. The pose of the new object image is estimated by solving PnP. Finally, we update the sparse point cloud by minimizing reprojection errors and perform back-projection to obtain an optimized, accurate object point cloud, as well as updated object poses.

To demonstrate the scalability of our method, we visualize the performance curve by randomly increasing the number of views. Figure 5 illustrates that the overall accuracy significantly improves as more visual information is incorporated.

Novel View Synthesis, an Application of POPE Our next objective is to validate the accuracy of our predicted pose estimation and demonstrate its practical applicability in downstream applications. To this end, we employ the estimated multi-view poses obtained from our POPE model, in combination with a pre-trained and generalizable Neural Radiance Field (GNT) [58].

Specifically, we configure the GNT with a source view

Dataset	Method	All Categories			Per Category (Med. Err ↓)				
		Med. Err (↓)	Acc30 (↑)	Acc15 (↑)	Eggbox	Can	Iron	Hole.	Camera
LINEMOD [22]	Gen6D [38]	44.855	0.364	0.096	31.781	30.407	30.094	45.288	35.970
	LoFTR [53]	33.036	0.562	0.324	16.887	17.585	17.904	31.782	22.550
	Ours	**15.731**	**0.770**	**0.483**	**10.530**	**12.699**	**13.157**	**14.779**	**15.102**
		Med. Err (↓)	Acc30 (↑)	Acc15 (↑)	008	003	005	006	010
YCB-Video [8]	Gen6D [38]	54.477	0.232	0.077	45.461	80.992	50.587	66.999	50.597
	LoFTR [53]	19.5419	0.686	0.478	15.359	36.942	17.832	28.999	17.475
	Ours	**13.9411**	**0.801**	**0.544**	**7.787**	**18.385**	**14.171**	**20.100**	**15.428**
		Med. Err (↓)	Acc30 (↑)	Acc15 (↑)	Mfmi.	Oreo	Taip.	Diyc.	Tee
OnePose++ [20]	Gen6D [38]	35.428	0.411	0.158	16.963	16.612	17.787	19.132	19.867
	LoFTR [53]	9.012	0.891	0.703	4.077	3.938	4.147	5.041	5.312
	Ours	**6.273**	**0.896**	**0.728**	**1.765**	**1.203**	**2.147**	**2.769**	**3.799**
		Med. Err (↓)	Acc30 (↑)	Acc15 (↑)	Yell.	Teab.	Oran.	Gree.	Inst.
OnePose [54]	Gen6D [38]	17.785	0.893	0.389	35.811	31.536	36.829	30.609	48.317
	LoFTR [53]	4.351	**0.963**	**0.918**	9.773	6.488	9.439	7.3482	17.136
	Ours	**2.155**	0.962	0.911	**5.470**	**3.194**	**8.044**	**3.967**	**16.492**

Table 1. We conduct experiments on zero-shot two-view object pose estimation on LINEMOD dataset, and report Median Error and Accuracy at 30º, 15º averaged across all 13 categories. We also report Accuracy at 30º broken down by class for an illustrative subset of categories.

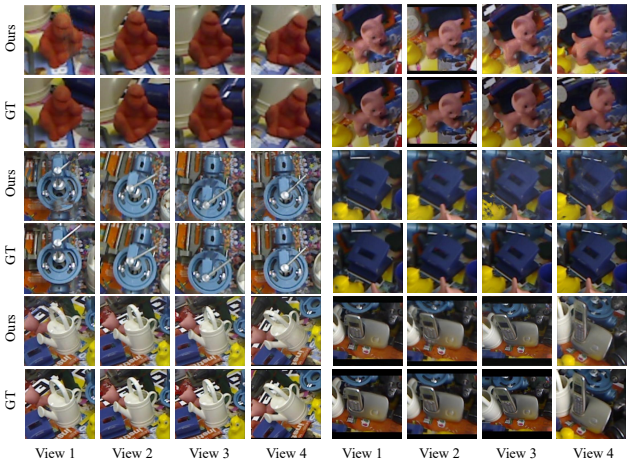


Figure 7. **Application: Novel View Synthesis**. In the domain of novel view synthesis, we employ the poses obtained from our POPE model in conjunction with a pre-trained, generalizable Neural Radiance Field (GNT) [58]. The GNT is configured with a source view number of 10, utilizing the ground truth poses for the purpose of warping. Subsequently, we generate 16 new viewpoints by rendering the scenes using the estimated poses derived from our POPE model.

number of 10 and utilize ground truth poses for source view warping. Subsequently, we leverage the estimated poses from our POPE model to generate new viewpoints based on the obtained POPE poses. Notably, our rendered results exhibit a remarkable resemblance to the ground-truth color image, as depicted in Figure 7, validating the precision of our estimated poses.

These findings provide compelling evidence supporting the accuracy and reliability of our pose estimation method, paving the way for its effective implementation in diverse downstream applications.

Promptable Object Pose Estimation in Arbitrary Scene We provide supplementary visual examples in Figure 8 and

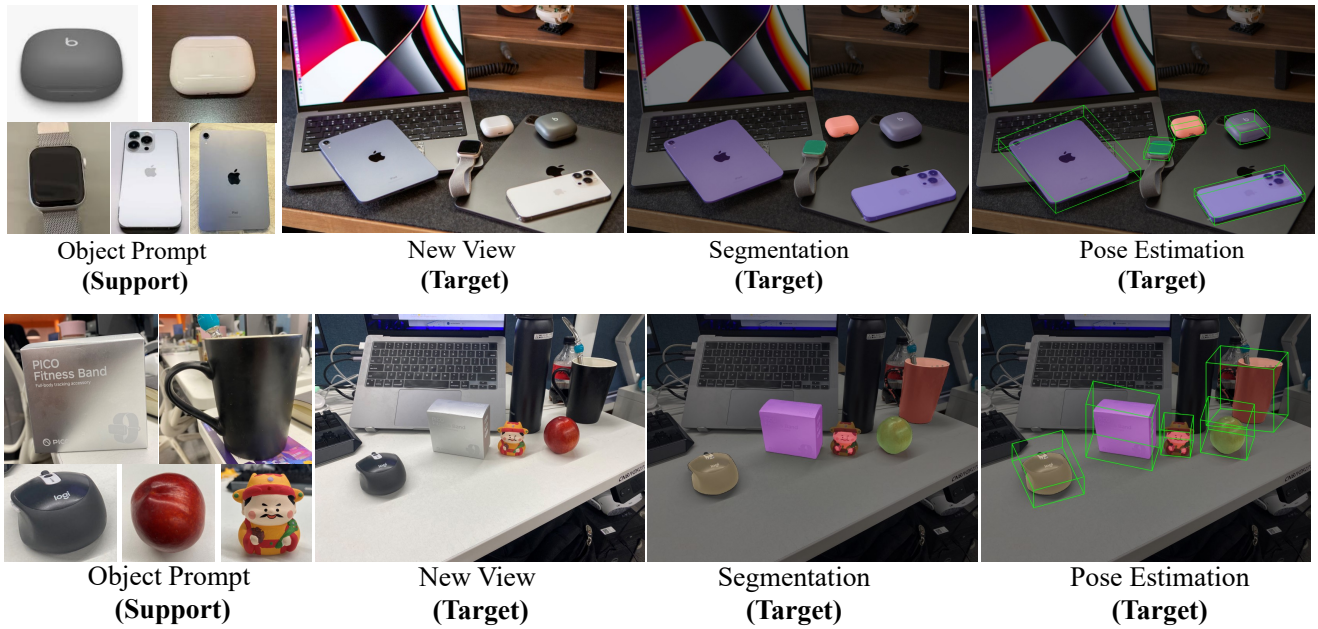


Figure 8. **Visual Examples of Promptable Object Pose Estimation for Indoor Test Cases.** we present visual examples showcasing the retrieved object masks and the estimated relative poses in the context of promptable object pose estimation for indoor test cases.

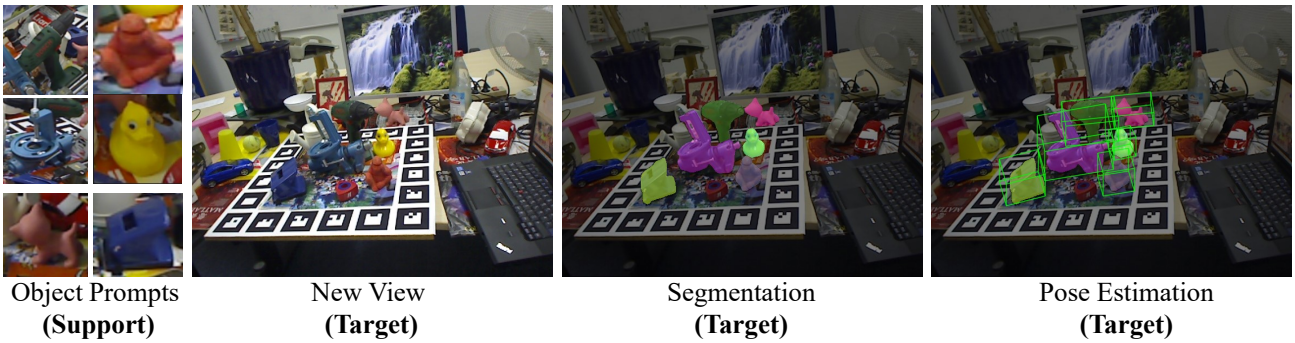


Figure 9. **Visual Examples of Promptable Object Pose Estimation for LINEMOD-Occ Test Cases.** we present visual examples showcasing the retrieved object masks and the estimated relative poses in the context of promptable object pose estimation on the occlusion subset of LINEMOD.

Figure 9 to further illustrate the effectiveness of our promptable 6DoF object pose estimation method. This approach leverages a prompt image containing the object of interest, and our algorithm, **POPE**, demonstrates the ability to recognize objects of various categories through segmentation and retrieval processes, ultimately achieving accurate estimation of relative object poses.

5. Conclusion

In this paper, we present Promptable Object Pose Estimator (POPE), a zero-shot solution for estimating the 6DoF object pose in any scene with only one support image. Our solution highlights the potential of leveraging 2D pre-trained foundation models to lift the typical object pose estimation to generalize in a more practical paradigm. It features a modular design that decomposes the promptable object pose estimation into several steps. We demonstrate the scalability of our proposed solution to use single support image as prompt under extreme clustered scenes, the extension to multiple viewpoints, and the validation on novel view synthesis. Several potential directions for future work exist, including distilling large-scale foundation models into smaller ones for enabling real-time inference, and incorporating single-view depth information from a monocular depth estimator to enhance zero-shot accuracy. We envision that our solution will enable users to generate photorealistic 3D assets for augmented or virtual reality applications using only a few images, even as sparse as two.

References

[1] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7822–7831, 2021. 2

[2] Vassileios Balntas, Andreas Doumanoglou, Caner Sahin, Juil Sock, Rigas Kouskouridas, and Tae-Kyun Kim. Pose guided rgbd feature learning for 3d object pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 3856–3864, 2017. 3

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. *Lecture notes in computer science*, 3951:404–417, 2006. 3

[4] Paul A. Beardsley, Philip H. S. Torr, and Andrew Zisserman. 3d model acquisition from extended image sequences. In *European Conference on Computer Vision*, 1996. 3

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[6] Benjamin Busam, Hyun Jun Jung, and Nassir Navab. I like to move it: 6d pose estimation as an action decision process. *arXiv preprint arXiv:2009.12678*, 2020. 3

[7] Ruojin Cai, Bharath Hariharan, Noah Snavely, and Hadar Averbuch-Elor. Extreme rotation estimation using dense correlation volumes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14566–14575, 2021. 3

[8] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pages 510–517. IEEE, 2015. 5, 6, 7

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2, 5

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

[11] Wei Chen, Yu Liu, Weiping Wang, Erwin M Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S Lew. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 5

[12] Xu Chen, Zijian Dong, Jie Song, Andreas Geiger, and Otmar Hilliges. Category level object pose estimation via neural analysis-by-synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 139–156. Springer, 2020. 2

[13] Meghal Dani, Karan Narain, and Ramya Hebbalaguppe. 3dposelite: a compact 3d pose estimation using node embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1878–1887, 2021. 3

[14] Xinke Deng, Junyi Geng, Timothy Bretl, Yu Xiang, and Dieter Fox. icaps: Iterative category-level object pose and shape estimation. *IEEE Robotics and Automation Letters*, 7(2): 1784–1791, 2022. 2

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[16] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, and Federico Tombari. So-pose: Exploiting self-occlusion for direct 6d pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12396–12405, 2021. 2

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

[18] Ufuk Efe, Kutalmis Gokalp Ince, and A. Aydin Alatan. Dfm: A performance baseline for deep feature matching. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4279–4288, 2021. 3

[19] Walter Goodwin, Sagar Vaze, Ioannis Havoutis, and Ingmar Posner. Zero-shot category-level object pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, pages 516–532. Springer, 2022. 2, 3, 4, 5, 6

[20] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without cad models. *arXiv preprint arXiv:2301.07673*, 2023. 2, 3, 4, 6, 7

[21] Stefan Hinterstoisser, Stefan Holzer, Cedric Cagniart, Slobodan Ilic, Kurt Konolige, Nassir Navab, and Vincent Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *2011 international conference on computer vision*, pages 858–865. IEEE, 2011. 3

[22] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I 11*, pages 548–562. Springer, 2013. 1, 5, 6, 7

[23] Tomas Hodan, Daniel Barath, and Jiri Matas. Epos: Estimating 6d pose of objects with symmetries. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 11703–11712, 2020. 2

[24] Hanwen Jiang, Zhenyu Jiang, Kristen Grauman, and Yuke Zhu. Few-view object reconstruction with unknown categories and camera poses. *ArXiv*, abs/2212.04492, 2022. 3

[25] Wei Jiang, Eduard Trulls, Jan Hendrik Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6187–6197, 2021. 3

[26] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, P. Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129:517–547, 2020. 3

[27] Glenn JOCHER, K Nishimura, T Minerva, and R Vilariño. Yolov5 [https://github. com/ultralytics/yolov5]. *Accessed March*, 7:2021, 2020. 2

[28] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the IEEE international conference on computer vision*, pages 1521–1529, 2017. 1

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 4, 6

[30] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 574–591. Springer, 2020. 1

[31] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Zi-Ping Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16242–16251, 2022. 3

[32] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 683–698, 2018. 3

[33] Zhigang Li, Gu Wang, and Xiangyang Ji. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7678–7687, 2019. 1

[34] Jiehong Lin, Hongyang Li, Ke Chen, Jiangbo Lu, and Kui Jia. Sparse steerable convolutions: An efficient learning of se (3)-equivariant features for estimation and tracking of object poses in 3d space. *Advances in Neural Information Processing Systems*, 34:16779–16790, 2021. 2

[35] Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3560–3569, 2021. 2

[36] Yunzhi Lin, Thomas Müller, Jonathan Tremblay, Bowen Wen, Stephen Tyree, Alex Evans, Patricio A Vela, and Stan Birchfield. Parallel inversion of neural radiance fields for robust pose estimation. *arXiv preprint arXiv:2210.10108*, 2022. 3

[37] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5967–5977, 2021. 3

[38] Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 298–315. Springer, 2022. 2, 3, 4, 5, 6, 7

[39] H Christopher Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293 (5828):133–135, 1981. 4

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

[41] Anastasiya Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *NIPS*, 2017. 3

[42] David Nistér. An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence*, 26(6):756–770, 2004. 4, 5

[43] Brian Okorn, Qiao Gu, Martial Hebert, and David Held. Zephyr: Zero-shot pose hypothesis rating. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14141–14148. IEEE, 2021. 3

[44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 5, 6

[45] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter 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*, pages 10710–10719, 2020. 3

[46] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4561–4570, 2019. 1

[47] Giorgia Pitteri, Aurélie Bugeau, Slobodan Ilic, and Vincent 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. 3

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[49] Chris Rockwell, Justin Johnson, and David F Fouhey. The 8-point algorithm as an inductive bias for relative pose prediction by vits. *arXiv preprint arXiv:2208.08988*, 2022. 3

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

[51] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2, 7

[52] Samarth Sinha, Jason Y Zhang, Andrea Tagliasacchi, Igor Gilitschenski, and David B Lindell. Sparsepose: Sparse-view camera pose regression and refinement. *arXiv preprint arXiv:2211.16991*, 2022. 2

[53] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 3, 5, 6, 7

[54] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6825–6834, 2022. 1, 2, 3, 4, 5, 7

[55] Martin Sundermeyer, Maximilian Durner, En Yen Puang, Zoltan-Csaba Marton, Narunas Vaskevicius, Kai O Arras, and Rudolph Triebel. Multi-path learning for object pose estimation across domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13916–13925, 2020. 3

[56] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 292–301, 2018. 1

[57] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 530–546. Springer, 2020. 2

[58] Mukund Varma, Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, and Zhangyang Wang. Is attention all that nerf needs? In *The Eleventh International Conference on Learning Representations*, 2022. 7

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

[60] Yilin Wen, Xiangyu Li, Hao Pan, Lei Yang, Zheng Wang, Taku Komura, and Wenping Wang. Disentangled implicit shape and pose learning for scalable 6d pose estimation. *arXiv preprint arXiv:2107.12549*, 2021. 2

[61] Paul Wohlhart and Vincent Lepetit. Learning descriptors for object recognition and 3d pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3109–3118, 2015. 3

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

[63] Yang Xiao, Xuchong Qiu, Pierre-Alain Langlois, Mathieu Aubry, and Renaud Marlet. Pose from shape: Deep pose estimation for arbitrary 3d objects. *arXiv preprint arXiv:1906.05105*, 2019. 3

[64] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2020. 3, 4

[65] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330. IEEE, 2021. 2

[66] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1941–1950, 2019. 2, 3

[67] Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 592–611. Springer, 2022. 2, 3