Pose Anything Anywhere: Model-free Object Poses from Arbitrary References

Hongli Xu*[1], Jiaqi Hu*[1,3], Junwen Huang*†[1,2], Boyang Zhong[1],
Peter KT Yu[4,5], Nassir Navab[1,2], Benjamin Busam[1,2], Slobodan Ilic[1,3]

* Equal contribution † Corresponding author

[1]Technical University of Munich [2]Munich Center for Machine Learning (MCML)
[3]Siemens AG [4]XYZ Robotics [5]ROBOX

Abstract. Estimating the 6D pose of unseen objects is a fundamental yet challenging problem for open-world robotics and embodied perception. Model-based methods are accurate but depend on CAD assets or heavy onboarding, while most model-free approaches are still limited to pairwise single-anchor matching and thus fail under occlusion and large viewpoint changes with low query–reference overlap. Therefore, we present **PANY**, a unified model-free framework that seamlessly supports both RGB and RGB-D inputs, operates on one or sparse pose-free reference views, and generalizes effectively to novel objects. Built on a multi-view transformer geometry backbone, PANY moves beyond pairwise matching by learning view-consistent geometry and cross-view alignment cues that remain stable under wide baselines and limited overlap. When additional unposed assist views are available, PANY aggregates them via pose-graph canonical registration to increase geometric coverage and reinforce the final pose. Extensive experiments show that PANY achieves state-of-the-art performance across multiple benchmarks, substantially outperforming existing model-free methods, improving pose accuracy by **+12%** on YCB-V and over **+20%** on LM-O. Furthermore, PANY consistently performs well under both single-reference and sparse-reference settings, demonstrating strong robustness in real-world environments.

1 Introduction

With decades of progress in autonomous systems, modern embodied AI is increasingly expected to perceive, reconstruct, and orient previously unseen objects. A key requirement is estimating an object's 6D pose for downstream interaction such as grasping, placement, and tool use. However, many existing pipelines rely on strong prerequisites (e.g., accurate CAD models, object-specific assets, or dense posed image sequences for reconstruction), which significantly limits scalability and usability in open-world deployment.

We study 6D object pose estimation in a practical operational regime where supervision and assets are minimal at inference time. Given a query image and

[0] * Equal contribution. The first three authors are listed in random order.

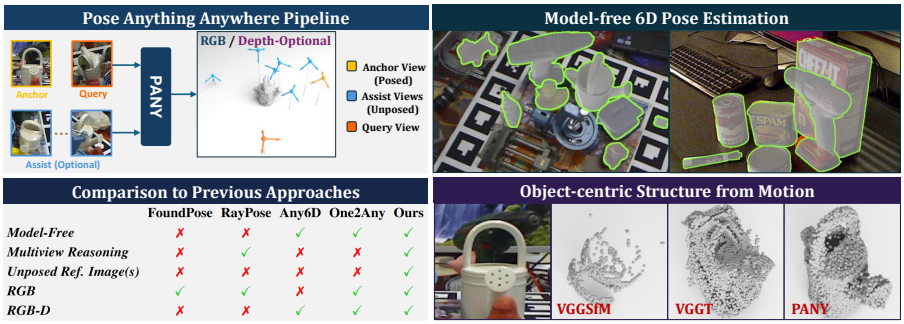

Fig. 1: PANY is the first model-free approach that simultaneously handles single- and sparse-view references and is compatible with RGB/RGBD input. Compared to previous approaches [19, 26, 33, 42], PANY jointly infers object geometry and 6D pose from sparse and unposed reference views. It remains robust under large viewpoint changes and occlusions, and maintains stability in object-centric scenarios where VGGSfM [55] and VGGT [54] fail.

an *arbitrary, unordered, and sparsely sampled* set of reference images of the same object instance, the goal is to recover the query pose in a canonical object-centric coordinate system, *without CAD models and without explicit reconstruction*. From an information perspective, a single reference view often provides insufficient coverage of the object under occlusion and large viewpoint change (e.g., missing the back side), leading to intrinsic ambiguity in correspondence and pose. Therefore, additional views can systematically reduce this uncertainty by increasing geometric coverage and establishing cross-view constraints. In our target setting, such extra observations come as *pose-free assist views*: they are unposed, unconstrained snapshots whose camera poses are unknown and should not require annotation, yet they can act as intermediates to bridge wide-baseline gaps.

Existing approaches fall into three broad groups according to the reference modality: **(A) Model-based pipelines** rely on CAD models or object-specific 3D assets to generate templates or renderings for pose estimation [12, 20, 31, 40, 42, 51]. While highly accurate, they require curated 3D libraries and are less scalable for open-world objects. **(B) Reconstruction-first pipelines** remove the need for CAD models by reconstructing an object from multiple views (e.g., SfM/NeRF/3DGS) and then applying a model-based solver [13, 22, 26, 28, 36, 48, 60]. Although effective, this direction typically incurs heavy capture or computation and does not directly support rapid onboarding from sparse, unordered references. **(C) Pairwise model-free methods** estimate pose by matching a query image to a single reference via dense correspondences or learned pose inference [7,10,30,33,43,53,56]. However, under large viewpoint change and occlusion, the query–reference overlap can be weak, making the problem intrinsically ambiguous and limiting robustness in the arbitrary sparse-reference regime. Recent geometry foundation models [54] provide strong multi-view geometry estimation from RGB inputs, but are trained with tracking-style supervision and often yield

view-dependent representations, which are insufficient for stable *object-centric canonical alignment* across wide baselines and unordered reference sets (Fig. 1).

Motivated by those observations, we present **PANY**, a model-free system for estimating object pose from arbitrary sparse references with optional pose-free assist views. In contrast to pairwise model-free approaches that rely on a single reference–query match, PANY performs multi-view reasoning across sparse and unordered observations to resolve pose ambiguity under occlusion and wide viewpoint changes. Unlike reconstruction-first pipelines that require explicit object reconstruction before pose estimation, PANY directly infers object-centric canonical alignment from sparse views without reconstructing a full 3D model. PANY builds upon a geometry foundation model and adapts it from scene-level prediction to object-centric canonical alignment. To reduce the ambiguity induced by limited overlap, PANY learns geometry-aware cross-view correspondences that remain consistent under wide-baseline viewpoint changes and occlusion. When additional unposed views are available, PANY further aggregates them as intermediates via a pose-graph-based canonical registration, increasing geometric coverage and reinforcing the final pose even when no single reference overlaps well with the query.

Extensive experiments on several public benchmarks demonstrate the effectiveness of our framework design, enabling PANY to generalize across modalities (RGB/RGB-D) and reference configurations(single reference or arbitrary sparse references with pose-free assist views), setting a new state of the art on several challenging benchmarks. Our key contributions are:

- We formalize pose estimation from **arbitrary sparse, unordered references** (optionally with pose-free assist views) as a practical operational regime, and propose **PANY** to address it without CAD models or onboarding stage.
- PANY adapts geometry foundation models to **object-centric canonical alignment** by learning geometry-aware, spatially consistent cross-view correspondences for robust pose reasoning under weak texture, occlusion, and wide baselines.
- PANY introduces an optional **multi-view inference** procedure that aggregates pose-free assist views via pose-graph canonical registration, improving robustness when no single reference view provides sufficient overlap.

2 Related Work

Recently, model-based unseen object pose estimation methods have shown remarkable progress, achieving performance comparable to overfitting approaches on public benchmarks [17,49]. With access to the CAD models, model-based approaches can generate a large set of templates to perform 2D-3D alignment [1, 19,37,37,40,42,46], direct 3D-3D matching [5,6,20,31], or adopt a render-and-compare strategy [25,29,50,60] to refine the estimated pose iteratively. However, their reliance on accurate CAD models as references restricts their applicability

in general daily scenarios where such models are unavailable. In this section, we review recent advances in **model-free novel object pose estimation**, focusing on approaches based on object reconstruction, relative pose estimation, and multiview reconstruction with foundational models.

Approaches based on object reconstruction. Instead of relying on accurate CAD models, some approaches reconstruct the 3D shape of the object from multi-view RGB images and align the input 2D image with the reconstructed 3D representation. Reconstruction can be achieved through traditional structure-from-motion (SfM) pipelines [13, 14, 34, 48], or implicitly using neural representations such as Neural Radiance Fields (NeRFs) [8, 28, 58–60] and 3D Gaussian Splatting (3DGS) [4, 21, 36]. However, these methods typically require video sequences or dense sets of posed images and involve a lengthy onboarding process for each new object. Recent studies aim to relax these constraints. NOPE [38] uses a single RGB image and matches it against synthetic templates generated from multiple viewpoints via a generative model. GigaPose [40] employs an image-to-3D model [35] to recover object geometry, followed by model-based pose estimation protocols, but still requires size initialization. Any6D [26] further extends this idea by jointly estimating object scale and pose from only a single RGB-D reference image. OnePoseViaGen [10] uses a powerful image-to-3D framework to reconstruct high-quality object 3D models. Although these methods take a single image as the anchor view, their pose estimation pipelines remain fundamentally model-based, relying on template matching against reconstructed object models. As a result, their performance is highly sensitive to the quality of the generated models, which can introduce geometric and appearance hallucinations when the object is only partially observed.

Approaches based on relative pose estimation. Extensive research [23,30, 43,47,56,63,64] has addressed relative camera pose estimation between query and reference images using dense descriptor matching [45, 47], pose diffusion [56, 63], or learning-based bundle adjustment [30, 64]. While effective for camera localization, these methods struggle to generalize to object-level pose estimation, where the target occupies smaller image regions and scale recovery is required. To overcome this, model-free approaches directly predict relative pose from a reference RGB or RGB-D image without explicit 3D reconstruction. Oryon [7] extends this paradigm to open-vocabulary settings, One2Any [26] predicts Reference Object Coordinate (ROC) maps from RGB-D pairs, and 3DAHV [67], DVMNet [66], and SingRef6D [53] achieve single-image generalization through shared 2D–3D embeddings or pretrained matchers [47]. Despite these advances, accurately reasoning geometric correlations under large viewpoint changes or discontinuities remains an open challenge.

Grounding models for visual geometry reasoning. Recent works have explored multi-view architectures to jointly learn geometry and semantics for reconstruction and pose estimation from unposed RGB inputs. DUSt3R [57] and

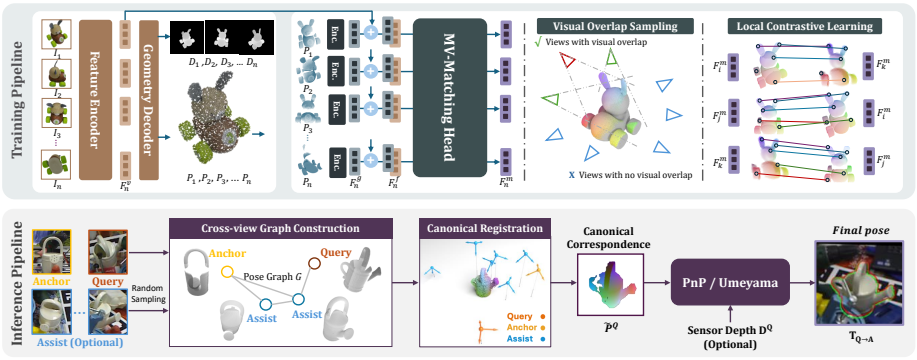

Fig. 2: Overview of the PANY framework. (i) Given a query image and multiple reference views (at least one posed anchor image and optionally a few unposed assist views). Our network takes all views as input, jointly learns local geometry, dense 3D correspondences, and models spatial-consistent view representations for robust object pose estimation. (ii) During inference, the assist views act as intermediates to bridge limited visual overlap between the query and anchor views, enabling reliable 3D reasoning under large viewpoint variations.

MASt3R [27] predict dense aligned point maps and camera parameters from uncalibrated image pairs. Fast3R [62] scales this idea to hundreds of unordered, unposed images through a transformer-based multi-view formulation, enabling efficient large-scale inference. VGGT [54] further unifies multi-view camera pose, depth, point map, and correspondence prediction within a single framework. Despite the impressive performance of grounded visual geometry models on scene-level inputs, scaling them to the object level remains challenging due to their view-dependent representations, especially when objects are only partially observed and exhibit significant viewpoint variation and occlusion, which hinder consistent geometric reasoning. Building on the strong geometry foundation model, our method eliminates view dependency in the predicted point maps and learns cross-view, spatially consistent 3D representations for robust pose alignment under large viewpoint changes and cluttered scenes.

3 Method

Our goal is to estimate the 6D pose of an object from sparse reference observations without requiring CAD Models or onboarding stage. The key challenge arises when the query and reference views exhibit limited visual overlap due to occlusion or large viewpoint changes. To address this challenge, we propose a canonical multi-view alignment framework that combines three key components: (1) a multi-view geometry backbone that predicts object-centric 3D structure from sparse views, (2) a cross-view correspondence learning module that establishes geometry-consistent matches across wide-baseline viewpoints, and (3) a pose-graph aggregation mechanism that integrates pose-free assist views to

increase geometric coverage and resolve pose ambiguity. Together, these components enable robust pose reasoning from arbitrary sparse references, even when no single reference sufficiently overlaps with the query.

3.1 Pipeline Overview.

Our method takes an arbitrary query view of an object and conditions on a sparse set of reference views, among which the anchor view defines the object's canonical pose, while the remaining assist views are pose-free. As illustrated in Figure 2, the framework comprises two stages: a training pipeline and an inference pipeline. During training, we employ a transformer backbone to predict dense point maps and jointly train a cross-view 3D matching head using contrastive learning on local point features. By enforcing cross-view geometric constraints through dense 3D correspondences, the model learns spatially consistent and robust multi-view representations. When the query and anchor views share little visual overlap, direct pose estimation becomes ambiguous. To resolve this ambiguity, we leverage pose-free assist views as intermediates that increase geometric coverage across viewpoints. We aggregate these views through a pose-graph registration procedure that aligns local reconstructions into a consistent canonical coordinate frame at inference time.

3.2 Task Formulation.

As illustrated in Figure 2, given a RGB(-D) *query image* denoted as I^{Q} and K *reference images* $\{I_i^{\mathrm{ref}}\}_{i=1}^{K}$, our goal is to estimate the object pose $\mathbf{T}^{Q}$ in the query image. Among $\{I_i^{\mathrm{ref}}\}_{i=1}^{K}$, one of which has a predefined canonical pose $\mathbf{T}^{A}$, termed the *anchor view* I^{A}, the remaining randomly captured pose-free reference views are denoted as *assist views* $\{I_i^{\mathrm{U}}\}_{i=1}^{M}$, which optionally providing complementary 3D contextual cues to enhance pose reasoning robustness. The query, anchor, and assist images form the complete set of inputs $\mathcal{I} = \{I^{\mathrm{Q}}\} \cup \{I^{\mathrm{A}}\} \cup \{I_i^{\mathrm{U}}\}$, with the total number of views $N = 2 + M, M \geq 0$. Our network takes $\mathcal{I} = \{I_i\}_{i=1}^{N}$ as input and estimates the relative object transformation between the query and anchor views $\mathbf{T}_{Q\to A}$ by leveraging the pose-free assist views. The final object pose in the query image can be solved by $\mathbf{T}^{Q} = \mathbf{T}^{A}\mathbf{T}_{Q\to A}^{-1}$.

3.3 Multi-view Geometric Reasoning

Given N input images $\{I_i\}_{i=1}^{N}$, each image is first patchified and encoded into tokens using a DINOv2 [41] encoder. These tokens are then processed by alternating frame-wise intra-attention within each image and global cross-attention across all images, allowing the model to capture both local appearance features and cross-view geometric correlations. The resulting features for each image, termed as *visual embedding* $\{F_i^{v}\}_{i=1}^{N}$ are decoded using DPT-based upsampling heads to produce dense geometric predictions, including depth maps $\hat{D}_i \in \mathbb{R}^{H\times W}$ and point maps $\hat{P}_i \in \mathbb{R}^{3\times H\times W}$.

$$f : \{I_i\}_{i=1}^{N} \longrightarrow \{(\hat{D}_i, \hat{P}_i)\}_{i=1}^{N}. \tag{1}$$

3.4 Cross-view 3D Matching

Dense 3D point maps provide a bridge across views, yet the associated features remain view-dependent, making local matching unstable under wide-baseline changes. Without explicit cross-view consistency constraints, independently extracted embeddings may drift and lead to unreliable correspondences To address these issues (Figure 2), we introduce a *cross-view 3D matching model* with three components aligned with the following subsections: (1) a geometry-aware feature encoding and fusion module that converts per-view point maps into 3D-aligned tokens and produces rotation-invariant descriptors; (2) a local contrastive objective that supervises cross-view correspondences on the resulting 3D descriptors; (3) an end-to-end joint training scheme that couples geometry prediction and correspondence learning to obtain consistent multi-view 3D representations.

Cross-view 3D Matching Model. Given the predicted dense point map $\hat{P}_i$ for each view I_i, we encode local geometry using a geometric transformer backbone [31,44]. A geometric encoder with rotation-invariant positional embedding extracts geometric features F_i^g from $\hat{P}_i$. In parallel, per-pixel visual features F_i^v are obtained from the visual backbone and lifted to 3D through coordinate indexing with $\hat{P}_i$. The two modalities are fused to form tokens F_i^f.

The fused tokens together with their 3D coordinates are processed by a geometry-aware matching head

$$F_i^m = \mathcal{T}_{\text{match}}\Big(F_i^f,\, \hat{P}_i\Big), \tag{2}$$

where $\mathcal{T}_{\text{match}}(\cdot)$ denotes a geometric transformer that alternates sparse geometric attention and dense linear attention [31]. The resulting descriptors F_i^m are rotation-invariant and geometry-aligned, enabling robust cross-view correspondence learning under viewpoint and appearance variations.

Local Contrastive Learning. Wide-baseline viewpoint changes often produce view-dependent representations that hinder reliable cross-view correspondence. To address this issue, we train the matching head using local contrastive supervision defined on the geometry-aware descriptors. Given a set of K input views with point sets $\{P_i\}_{i=1}^K$ and descriptors $\{F_i^m\}_{i=1}^K$, we compute cross-view attention matrices between view pairs as

$$A_{ij} = F_i^m (F_j^m)^\top, \tag{3}$$

where $A_{ij} \in \mathbb{R}^{N_i \times N_j}$ measures descriptor similarity between points from views i and j. For each valid view pair (i, j), the attention matrix is supervised using a bidirectional cross-entropy objective

$$\mathcal{L}_{ij} = \text{CE}(A_{ij}, \hat{Y}_i) + \text{CE}(A_{ij}^\top, \hat{Y}_j), \tag{4}$$

where $\hat{Y}_i$ and $\hat{Y}_j$ denote the ground-truth correspondences. The final matching loss aggregates all valid pairs

$$\mathcal{L}_{\text{match}} = \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \mathcal{L}_{ij}. \tag{5}$$

Since the predicted point maps lie in a shared coordinate frame, the correspondence label for a point $p_i \in P_i$ is defined by nearest-neighbor search

$$k^* = \arg\min_k \|p_i - p_{j,k}\|_2, \tag{6}$$

and

$$y_i = \begin{cases} 0, & \|p_i - p_{j,k^*}\|_2 \geq \delta_{\text{dis}} \\ k^*, & \text{otherwise}, \end{cases} \tag{7}$$

where δ_{dis} is a distance threshold that determines whether a valid correspondence exists.

Joint Learning for Geometry and Matching To jointly enable accurate 3D geometry prediction and cross-view correspondence learning from multi-view RGB inputs, We jointly optimize the geometry prediction module and the correspondence learning module in an end-to-end manner. Our training objective combines supervision for both geometric reconstruction and cross-view contrastive learning. The geometry head predicts dense depth maps $\hat{D}_i$ and point maps $\hat{P}_i$, which are supervised using ℓ_1 regression losses against their ground-truth counterparts:

$$\mathcal{L}_{\text{depth}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{D}_i - \overline{D}_i\|_1, \quad \mathcal{L}_{\text{point}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{P}_i - \overline{P}_i\|_1. \tag{8}$$

For cross-view correspondence supervision, we adopt the local contrastive loss described in Section 3.4. The final objective integrates all components:

$$\mathcal{L} = \lambda_d \mathcal{L}_{\text{depth}} + \lambda_p \mathcal{L}_{\text{point}} + \lambda_m \mathcal{L}_{\text{match}}, \tag{9}$$

where λ_d, λ_p, and λ_m are balancing weights. This joint objective enforces accurate per-view geometry prediction while learning geometry-aware embeddings that remain consistent across multiple views.

3.5 Addressing Large-view Gap

Large viewpoint gaps between the query and anchor views often lead to limited visual overlap, making direct pose estimation unreliable. To mitigate this issue, we leverage pose-free assist views as intermediate observations that increase geometric coverage across viewpoints. Our pipeline is trained with multi-view inputs, enabling flexible inference with an arbitrary number of views and ensuring 3D-consistent point map predictions across them.

Pose Graph Construction. Given a query view I^Q, an anchor view I^A with canonical pose $\mathbf{T}^A$, and M pose-free assist views $\{I_i^U\}_{i=1}^M$, the network predicts dense point maps $\hat{P}_i$ and cross-view correspondences across all views. We construct a pose graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node represents a view and its predicted point cloud. An edge $(i, j) \in \mathcal{E}$ is created if the number of valid correspondences between two views exceeds a threshold η. For each connected pair,

we estimate a similarity transformation $\mathbf{T}_{i\leftarrow j}$ using robust point-set alignment with RANSAC and Umeyama estimation.

Canonical Alignment. To recover a globally consistent reconstruction, we align all predicted point clouds into the anchor-centric coordinate frame. Starting from the anchor node, we traverse the pose graph and compose the estimated transformations along the graph path:

$$\mathbf{T}_{i\to A} = \prod_{(k,l)\in\pi(i,A)} \mathbf{T}_{k\leftarrow l}. \tag{10}$$

The predicted point clouds are then registered into the canonical frame as

$$\tilde{P}_i = \mathbf{T}_{i\to A}\hat{P}_i.$$

This multi-view alignment aggregates geometric information from multiple views and resolves pose ambiguity that cannot be addressed by pairwise matching alone.

Final Pose Estimation. Once the query point cloud is registered into the canonical coordinate system, the final object pose in the query image is recovered as

$$\mathbf{T}^Q = \mathbf{T}^A\mathbf{T}^{-1}_{Q\to A}. \tag{11}$$

The relative transformation $\mathbf{T}_{Q\to A}$ is estimated from the canonical point clouds using Kabsch–Umeyama alignment when depth is available, or via a PnP optimization using predicted 3D–2D correspondences for RGB-only inputs.

4 Experiments

4.1 Implementation Details

Input images are pre-processed by cropping and resizing to a resolution of 518×518. During training, the VGGT [54] feature backbone is frozen; we apply LoRA [18] with a rank of 8 and an alpha value of 32 to the frozen components. The proposed matching head is trained from scratch, jointly with the LoRA finetuning. We set the loss weights to $\lambda_d = 1.0$, $\lambda_p = 1.0$, and $\lambda_m = 0.05$ for balancing geometry and matching objectives. The contrastive InfoNCE temperature is set to $\tau = 0.1$. For each training sample, given a query image, we randomly select 2 to 8 reference views, resulting in a total of 3 to 9 input images per sample. Among these references, at least two are required to be positive reference views, with the positive view threshold $\theta_{\text{OFP}} = 10°$. For cross-view graph construction, we set the minimum correspondence threshold to $\eta = 500$ valid 3D pairs from 2048 samples per view pair. The model is trained for 15 epochs with a batch size of 16, where each epoch consists of 500 iterations. We train the model on two NVIDIA A100 GPUs for approximately 36 hours. We use the AdamW optimizer [24] with a learning rate of 1e-4 and employ a warm-up strategy during the first epoch.

4.2 Dataset

Training dataset. To enhance the generalization ability of our model to diverse real-world objects, we train our model on the large-scale synthetic Omni6DPose dataset [65], which contains more than 5,000 CAD objects spanning 149 categories. For training, we extract object-centric crops and construct paired reference-query samples, resulting in over 2 million RGB-D images with accurate ground-truth poses.
Test benchmarks. To evaluate the robustness and generalization ability of our method, we test our method on 5 real-world datasets that have never been seen during training. We report results on LINEMOD [15], YCB-Video [61], Toyota-Light [16], LM-O [2], and Real275 [52] test set.
Evaluation metrics. We follow the BOP benchmark [16] and previous methods [7, 26, 33, 42] to evaluate the pose estimation performance on single-reference and sparse-reference setups. For fair comparison with the previous methods, we adopt the Average Recall (**AR**), **ADD-0.1d**, and **ADD AUC** as our evaluation metrics.

4.3 Experiment Setup

To demonstrate the flexibility and scalability of our approach, we compare against recent state-of-the-art methods under varying input configurations. Following Oryon [7], which samples image pairs for testing, we evaluate our method alongside RGB [30, 33, 43, 53, 56] and RGB-D [11, 26, 32, 33] baselines on the Real275 [52] and Toyota-Light [16] datasets. We also follow the One2Any [33] protocol that directly uses the first view as the anchor image, evaluating our method on LINEMOD [15] and YCB-Video [61]. For the multi-reference setup, we further compare with recent template-based methods on the LM-O [3] dataset.

4.4 Comparison to Single-Reference Methods

Evaluation on Real275 and Toyota-Light datasets. We compare our approach with baselines that only take a single RGB or RGB-D reference image as input on the Real275 [52] and Toyota-Light [16] datasets, as summarized in Table 1. For fair comparison, we follow the evaluation protocol established by Oryon [7], and evaluate 2,000 query–reference image pairs per test set, assuming ground-truth masks are available. As shown in Table 1, our method consistently outperforms all alternative approaches across all metrics when only a single reference view is provided. Specifically, compared to methods that estimate depth from the input RGB image in SingRef6D [53], generate 3D geometry from images in Any6D [26], or perform point cloud registration from visual embeddings in Oryon [7], our model achieves more accurate query-reference alignment even with a single reference image, demonstrating the robustness and generalizability of our 3D reasoning pipeline.
Evaluation on YCB-V and LINEMOD datasets. Following One2Any [33], we evaluate our method on the occluded YCB-V [61] and LINEMOD [15] datasets,

Table 1: Comparison with state-of-the-art methods on **Real275** [52] and **Toyota-Light** [16] datasets. **All the methods use the same query-reference image pairs**.

Method	Mod.	Real275 [52]		Toyota-Light [16]	
		AR	ADD-0.1d	AR	ADD-0.1d
PoseDiffusion [56]	RGB	9.2	0.8	7.8	1.2
RelPose++ [30]	RGB	22.8	11.9	30.9	11.6
LatentFusion [43]	RGB	22.6	9.6	28.2	10.2
One2Any [43]	RGB	26.6	4.8	22.7	3.7
SingRef6D [53]	RGB	28.7	11.6	31.7	**13.1**
PANY (Ours)	RGB	**59.3**	**27.8**	**37.5**	10.8
ObjectMatch [11]	RGBD	26.0	13.4	9.8	5.4
Oryon [7]	RGBD	46.5	34.9	34.1	22.9
Any6D [26]	RGBD	51.0	53.5	43.3	32.2
One2Any [33]	RGBD	54.9	41.0	42.0	34.6
PANY (Ours)	RGBD	**81.8**	**89.3**	**51.1**	**39.8**

Table 2: Pose estimation performance on **YCB-Video** [61] and **LINEMOD** [9] datasets. **All approaches use the first image from each scene as the reference view.** Methods marked with * require additional CAD models or ground-truth translations.

Method	Ref. Images	YCB-V [61] ADD AUC	LINEMOD [9] ADD-0.1d
FS6D [14] + ICP	16	42.1	91.5
FoundationPose* [60]	1 (CAD)	76.1	48.3
NOPE [38]*	1 + GT Trans	25.1	16.3
Oryon [7]	1	7.4	9.8
One2Any [33]	1	84.4	52.6
PANY (Ours)	1	92.5	55.3
PANY (Ours)	1 + 8 unposed	**94.6**	**87.4**

using only the first view as the posed reference and assuming available ground-truth masks. YCB-V poses challenges due to severe occlusions, while LINEMOD contains textureless and irregular objects captured in nearly 360° video sequences, making single-view reference particularly difficult. We adopt FoundationPose [60] results from [33] for comparison. Our PANY (single-ref.) uses only one anchor image as reference, whereas PANY (multi-ref.) incorporates eight unposed assist views. Table 2 shows that our method outperforms all baselines, and qualitative results in Figure 3 highlight robustness to occlusion, textureless surfaces, and large viewpoint changes. Those experiments demonstrate that incorporating assist views significantly boosts performance, validating the effectiveness of our multi-view reasoning pipeline under challenging conditions.

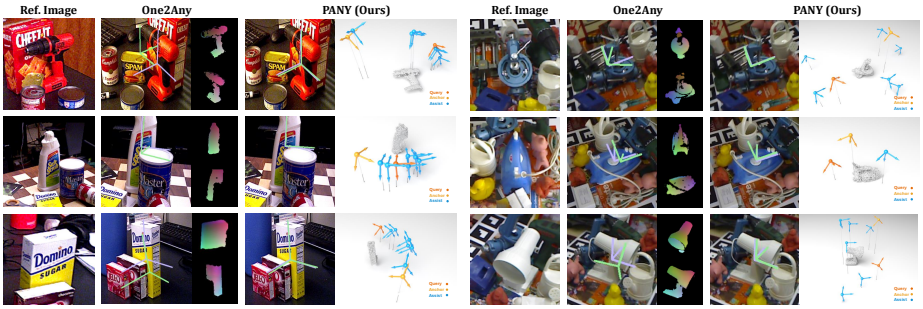

Fig. 3: Quantative Comparison between One2Any [33] and Ours. The predicted poses are displayed in green and ground truth poses are in pink. The second column shows the pose estimation results of One2Any [33] through pair-wise prediction; the last column shows how we leverage unposed references to conduct multi-view reasoning under large viewpoint changes.

4.5 Performance of Multi-view Pose Reasoning

To evaluate the effectiveness of our multi-view pose reasoning, we conduct experiments on the challenging LM-O dataset [2], which features severe occlusions and large viewpoint changes. Following prior model-based unseen object pose estimation works, we use CNOS [39] masks instead of ground-truth segmentation. For each LM-O object, we select eight sparse reference views from LINEMOD [15], with only the first view defining the canonical pose. We compare against state-of-the-art model-based RGB-only methods in Figure 4, and also reproduce FoundPose [42] under the same setting with additional in-plane rotations for fairness. As shown in Table 3, our method surpasses all model-based baselines that rely on dense CAD templates, despite using only eight posed anchors. Compared to One2Any [33], which uses a single reference, our approach achieves nearly double the accuracy, demonstrating the strength of our multi-view reasoning and geometry-aligned design.

Table 3: Evaluation on **LM-O** [15] dataset compared to model-based methods with RGB image inputs and all methods use CNOS [39] mask as segmentation mask.

Method	# Posed Ref.	Model-free	AR	Time(s)
MegaPose [25]	520	✗	22.9	2.53
GenFlow [37]	208	✗	25.0	-
GigaPose [40]	162	✗	29.9	11.53
FoundPose [42]	57	✗	39.7	1.7
RayPose [19]	8	✗	42.1	**0.71**
FoundPose [42]	8	✓	34.7	1.7
PANY (Ours)	1 + 8 unposed	✓	**42.3**	**0.75**
One2Any [33]	1	✓	18.2	**0.09**
Any6D [26]	1	✓	28.6	-
PANY (Ours)	1	✓	**36.9**	0.19

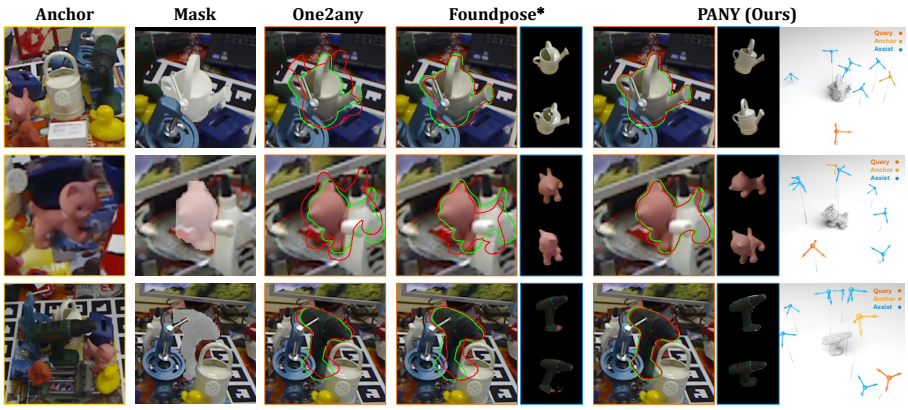

Fig. 4: Qualitative comparison between One2Any [33], FoundPose [42], and PANY. Green and red contours indicate ground-truth and predicted poses. Yellow, orange, and blue mark the anchor view, query view, and assist views. FoundPose* additionally uses assist views with pose annotations.

4.6 Ablation Study

Model design. We analyze the contribution of each component of our canonical multi-view alignment framework on the Real-275 [52] dataset under a controlled setting with one anchor and one query image (no assist views). Results are summarized in Table 4b. *(1) Geometry-only baseline.* As a minimal baseline, we directly align the predicted query and anchor point maps without explicit correspondence learning. This setting evaluates whether geometry prediction alone is sufficient for reliable pose recovery. *(2) Visual correspondence without geometric fusion.* We then introduce a cross-view matching module operating on visual embeddings only, without incorporating explicit geometric features. This improves setup (1) by 7.6% in ADD-0.1d, indicating that learning cross-view correspondences already provides stronger alignment cues than geometry prediction alone. *(3) Full model.* Our complete framework further fuses geometric point features with visual embeddings to jointly learn geometry-aware correspondences in a shared 3D feature space. This design yields a 17.8% improvement over the baseline.

Overall, these results show that the gains primarily come from the proposed cross-view correspondence learning and geometry-aware feature fusion, which together improve the spatial consistency of predicted geometry and enable more reliable pose alignment.

Backbone-only baseline. To verify that the gains do not simply come from using a stronger geometry backbone, we evaluate backbone-only variants using DUSt3R, MASt3R, and VGGT to predict geometry while keeping the same downstream solver. As shown in Table 4a, replacing the backbone alone is insufficient, and our full system consistently outperforms all backbone-only baselines.

Effect of anchor view sampling. We evaluate robustness to anchor selection on LM-O [2] by randomly sampling anchor images three times per object.

Table 4: Controlled ablation under identical inference settings. (a) Backbone-only baselines: replacing the geometry backbone (DUSt3R/MASt3R/VGGT) while keeping the same solver and without matching or aggregation is insufficient to match our performance. (b) Model design ablation on Real275: under identical single-anchor settings (no assist views), progressively adding matching and geometric fusion yields consistent gains, demonstrating that improvements are not explained by backbone replacement alone.

Method	AR	ADD-0.1d
vanilla DUSt3R	27.8	16.8
vanilla MASt3R	29.2	17.6
vanilla VGGT	43.6	33.4
Ours Full	**51.1**	**39.8**

(a) Backbone-only baselines.

Design Variant	AR	ADD-0.1d
Geometry-only	56.6 (-4.6%)	23.6 (-15.1%)
+ Matching	57.6(-2.9%)	25.4 (-8.6%)
Full (Matching+Fusion)	**59.3**	**27.8**

(b) Model design ablation.

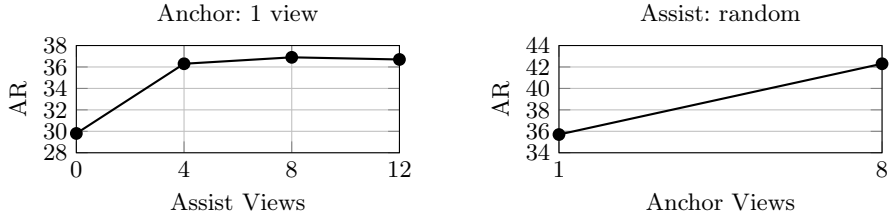

Fig. 5: Impact of assist views and anchor views on LM-O.

The resulting AR scores remain nearly identical, indicating that the method is insensitive to anchor choice.

Effect of the number of assist views. We vary the number of assist views during inference on LM-O and average over three random samples per setting. Performance improves with additional views and saturates beyond four, motivating our default use of eight assist views.

Limitations. Our method may struggle in cases with extreme occlusion or very limited query reference overlap, where reliable correspondences cannot be established. Strongly symmetric objects can also introduce pose ambiguity when geometric cues alone are insufficient to resolve orientation.

5 Conclusions

We present PANY, a scalable framework for model-free pose estimation from sparse RGB references. By jointly learning geometry-aware representations and cross-view correspondences, PANY enables robust object-centric alignment under large viewpoint changes and limited visual overlap. Experiments across multiple benchmarks demonstrate strong generalization across objects and reference configurations. Overall, PANY provides a practical step toward robust multi-view reasoning for real-world robotic perception.

References

1. Ausserlechner, P., Haberger, D., Thalhammer, S., Weibel, J.B., Vincze, M.: Zs6d: Zero-shot 6d object pose estimation using vision transformers. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 463–469. IEEE (2024)
2. Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J., Rother, C.: Learning 6d object pose estimation using 3d object coordinates. In: European conference on computer vision. pp. 536–551. Springer (2014)
3. Brachmann, E., Rother, C.: Learning less is more-6d camera localization via 3d surface regression. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4654–4662 (2018)
4. Cai, D., Heikkilä, J., Rahtu, E.: Gs-pose: Generalizable segmentation-based 6d object pose estimation with 3d gaussian splatting. In: 2025 International Conference on 3D Vision (3DV). pp. 1001–1011. IEEE (2025)
5. Caraffa, A., Boscaini, D., Hamza, A., Poiesi, F.: Freeze: Training-free zero-shot 6d pose estimation with geometric and vision foundation models. In: European Conference on Computer Vision. pp. 414–431. Springer (2024)
6. Chen, J., Sun, M., Bao, T., Zhao, R., Wu, L., He, Z.: Zeropose: Cad-model-based zero-shot pose estimation. arXiv preprint arXiv:2305.17934 **2**(6), 7 (2023)
7. Corsetti, J., Boscaini, D., Oh, C., Cavallaro, A., Poiesi, F.: Open-vocabulary object 6d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18071–18080 (2024)
8. Di Felice, F., Remus, A., Gasperini, S., Busam, B., Ott, L., Tombari, F., Siegwart, R., Avizzano, C.A.: Zero123-6d: Zero-shot novel view synthesis for rgb category-level 6d pose estimation. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 14204–14211. IEEE (2024)
9. Esposito, M., Busam, B., Hennersperger, C., Rackerseder, J., Navab, N., Frisch, B.: Multimodal us–gamma imaging using collaborative robotics for cancer staging biopsies. International journal of computer assisted radiology and surgery **11**(9), 1561–1571 (2016)
10. Geng, Z., Wang, N., Xu, S., Ye, C., Li, B., Chen, Z., Peng, S., Zhao, H.: One view, many worlds: Single-image to 3d object meets generative domain randomization for one-shot 6d pose estimation. arXiv preprint arXiv:2509.07978 (2025)
11. Gümeli, C., Dai, A., Nießner, M.: Objectmatch: Robust registration using canonical object correspondences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13082–13091 (2023)
12. Haugaard, R.L., Buch, A.G.: 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. pp. 6749–6758 (2022)
13. He, X., Sun, J., Wang, Y., Huang, D., Bao, H., Zhou, X.: Onepose++: Keypoint-free one-shot object pose estimation without cad models. Advances in Neural Information Processing Systems **35**, 35103–35115 (2022)
14. He, Y., Wang, Y., Fan, H., Sun, J., Chen, Q.: Fs6d: Few-shot 6d pose estimation of novel objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6814–6824 (2022)
15. Hinterstoisser, S., Holzer, S., Cagniart, C., Ilic, S., Konolige, K., Navab, N., Lepetit, V.: Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In: 2011 international conference on computer vision. pp. 858–865. IEEE (2011)

16. Hodan, T., Michel, F., Brachmann, E., Kehl, W., GlentBuch, A., Kraft, D., Drost, B., Vidal, J., Ihrke, S., Zabulis, X., et al.: Bop: Benchmark for 6d object pose estimation. In: Proceedings of the European conference on computer vision (ECCV). pp. 19–34 (2018)
17. Hodan, T., Sundermeyer, M., Labbe, Y., Nguyen, V.N., Wang, G., Brachmann, E., Drost, B., Lepetit, V., Rother, C., Matas, J.: Bop challenge 2023 on detection segmentation and pose estimation of seen and unseen rigid objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5610–5619 (2024)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
19. Huang, J., Vutukur, S.R., Yu, P.K., Navab, N., Ilic, S., Busam, B.: Raypose: Ray bundling diffusion for template views in unseen 6d object pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9102–9112 (2025)
20. Huang, J., Yu, H., Yu, K.T., Navab, N., Ilic, S., Busam, B.: Matchu: Matching unseen objects for 6d pose estimation from rgb-d images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10095–10105 (2024)
21. Jin, Y., Prasad, V., Jauhri, S., Franzius, M., Chalvatzaki, G.: 6dope-gs: Online 6d object pose estimation using gaussian splatting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8032–8043 (2025)
22. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
23. Kim, J., Park, J., Lee, K., Cho, N.I.: Refpose: Leveraging reference geometric correspondences for accurate 6d pose estimation of unseen objects. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6447–6456 (2025)
24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
25. Labbé, Y., Manuelli, L., Mousavian, A., Tyree, S., Birchfield, S., Tremblay, J., Carpentier, J., Aubry, M., Fox, D., Sivic, J.: Megapose: 6d pose estimation of novel objects via render & compare. arXiv preprint arXiv:2212.06870 (2022)
26. Lee, T., Wen, B., Kang, M., Kang, G., Kweon, I.S., Yoon, K.J.: Any6d: Model-free 6d pose estimation of novel objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11633–11643 (2025)
27. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European conference on computer vision. pp. 71–91. Springer (2024)
28. Li, F., Vutukur, S.R., Yu, H., Shugurov, I., Busam, B., Yang, S., Ilic, S.: Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2123–2133 (2023)
29. Li, Y., Wang, G., Ji, X., Xiang, Y., Fox, D.: Deepim: Deep iterative matching for 6d pose estimation. In: Proceedings of the European conference on computer vision (ECCV). pp. 683–698 (2018)
30. Lin, A., Zhang, J.Y., Ramanan, D., Tulsiani, S.: Relpose++: Recovering 6d poses from sparse-view observations. In: 2024 International Conference on 3D Vision (3DV). pp. 106–115. IEEE (2024)
31. Lin, J., Liu, L., Lu, D., Jia, K.: Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27906–27916 (2024)

32. Lin, J., Wei, Z., Ding, C., Jia, K.: Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks. In: European Conference on Computer Vision. pp. 19–34. Springer (2022)
33. Liu, M., Li, S., Chhatkuli, A., Truong, P., Van Gool, L., Tombari, F.: One2any: One-reference 6d pose estimation for any object. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6457–6467 (2025)
34. Liu, Y., Wen, Y., Peng, S., Lin, C., Long, X., Komura, T., Wang, W.: Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. In: European Conference on Computer Vision. pp. 298–315. Springer (2022)
35. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9970–9980 (2024)
36. Matteo, B., Tsesmelis, T., James, S., Poiesi, F., Del Bue, A.: 6dgs: 6d pose estimation from a single image and a 3d gaussian splatting model. In: European Conference on Computer Vision. pp. 420–436. Springer (2024)
37. Moon, S., Son, H., Hur, D., Kim, S.: Genflow: Generalizable recurrent flow for 6d pose refinement of novel objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10039–10049 (2024)
38. Nguyen, V.N., Groueix, T., Ponimatkin, G., Hu, Y., Marlet, R., Salzmann, M., Lepetit, V.: Nope: Novel object pose estimation from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17923–17932 (2024)
39. Nguyen, V.N., Groueix, T., Ponimatkin, G., Lepetit, V., Hodan, T.: Cnos: A strong baseline for cad-based novel object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2134–2140 (2023)
40. Nguyen, V.N., Groueix, T., Salzmann, M., Lepetit, V.: Gigapose: Fast and robust novel object pose estimation via one correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9903–9913 (2024)
41. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
42. Örnek, E.P., Labbé, Y., Tekin, B., Ma, L., Keskin, C., Forster, C., Hodan, T.: Foundpose: Unseen object pose estimation with foundation features. In: European Conference on Computer Vision. pp. 163–182. Springer (2024)
43. Park, K., Mousavian, A., Xiang, Y., Fox, D.: 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. pp. 10710–10719 (2020)
44. Qin, Z., Yu, H., Wang, C., Guo, Y., Peng, Y., Xu, K.: Geometric transformer for fast and robust point cloud registration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11143–11152 (2022)
45. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4938–4947 (2020)
46. Shugurov, I., Li, F., Busam, B., Ilic, S.: Osop: A multi-stage one shot object pose estimation framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6835–6844 (2022)

47. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8922–8931 (2021)
48. Sun, J., Wang, Z., Zhang, S., He, X., Zhao, H., Zhang, G., Zhou, X.: Onepose: One-shot object pose estimation without cad models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6825–6834 (2022)
49. Sundermeyer, M., Hodaň, T., Labbe, Y., Wang, G., Brachmann, E., Drost, B., Rother, C., Matas, J.: Bop challenge 2022 on detection, segmentation and pose estimation of specific rigid objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2785–2794 (2023)
50. Tremblay, J., Wen, B., Blukis, V., Sundaralingam, B., Tyree, S., Birchfield, S.: Diff-dope: Differentiable deep object pose estimation. arXiv preprint arXiv:2310.00463 (2023)
51. Wang, G., Manhardt, F., Tombari, F., Ji, X.: Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16611–16621 (2021)
52. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: 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. pp. 2642–2651 (2019)
53. Wang, J., Zhu, H., Guo, H., Al Mamun, A., Xiang, C., Lee, T.H.: Singref6d: Monocular novel object pose estimation with a single rgb reference. Advances in Neural Information Processing Systems **38**, 25400–25437 (2026)
54. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
55. Wang, J., Karaev, N., Rupprecht, C., Novotny, D.: Vggsfm: Visual geometry grounded deep structure from motion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21686–21697 (2024)
56. Wang, J., Rupprecht, C., Novotny, D.: Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9773–9783 (2023)
57. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20697–20709 (2024)
58. Wen, B., Bekris, K.: Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8067–8074. IEEE (2021)
59. Wen, B., Tremblay, J., Blukis, V., Tyree, S., Müller, T., Evans, A., Fox, D., Kautz, J., Birchfield, S.: Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 606–617 (2023)
60. Wen, B., Yang, W., Kautz, J., Birchfield, S.: Foundationpose: Unified 6d pose estimation and tracking of novel objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17868–17879 (2024)
61. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. arXiv preprint arXiv:1711.00199 (2017)

62. Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M.: Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21924–21935 (2025)
63. Zhang, J., Lin, A., Kumar, M., Yang, T.H., Ramanan, D., Tulsiani, S.: Cameras as rays: Pose estimation via ray diffusion. In: International Conference on Learning Representations. vol. 2024, pp. 23345–23366 (2024)
64. Zhang, J.Y., Ramanan, D., Tulsiani, S.: Relpose: Predicting probabilistic relative rotation for single objects in the wild. In: European Conference on Computer Vision. pp. 592–611. Springer (2022)
65. Zhang, J., Huang, W., Peng, B., Wu, M., Hu, F., Chen, Z., Zhao, B., Dong, H.: Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In: European Conference on Computer Vision. pp. 199–216. Springer (2024)
66. Zhao, C., Zhang, T., Dang, Z., Salzmann, M.: Dvmnet: Computing relative pose for unseen objects beyond hypotheses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20485–20495 (2024)
67. Zhao, C., Zhang, T., Salzmann, M.: 3d-aware hypothesis & verification for generalizable relative object pose estimation. In: International Conference on Learning Representations. vol. 2024, pp. 45563–45578 (2024)