

# RayPose: Ray Bundling Diffusion for Template Views in Unseen 6D Object Pose Estimation

Junwen Huang<sup>1,2</sup>Shishir Reddy Vutukur<sup>1,2</sup>  
Slobodan Ilic<sup>1</sup>Peter KT Yu<sup>3</sup>  
Benjamin Busam<sup>1,2</sup>Nassir Navab<sup>1,2</sup><sup>1</sup>Technical University of Munich <sup>2</sup>Munich Center for Machine Learning <sup>3</sup>XYZ Robotics

## Abstract

Typical template-based object pose pipelines estimate the pose by retrieving the closest matching template and aligning it with the observed image. However, failure to retrieve the correct template often leads to inaccurate pose predictions. To address this, we reformulate template-based object pose estimation as a ray alignment problem, where the viewing directions from multiple posed template images are learned to align with a non-posed query image. Inspired by recent progress in diffusion-based camera pose estimation, we embed this formulation into a diffusion transformer architecture that aligns a query image with a set of posed templates. We reparameterize object rotation using object-centered camera rays and model object translation by extending scale-invariant translation estimation to dense translation offsets. Our model leverages geometric priors from the templates to guide accurate query pose inference. A coarse-to-fine training strategy based on narrowed template sampling improves performance without modifying the network architecture. Extensive experiments across multiple benchmark datasets show competitive results of our method compared to state-of-the-art approaches in unseen object pose estimation.

## 1. Introduction

Multi-view vision is a core element for 3D perception [12]. Spatial understanding and measurements often depends on multiple cameras or temporally-varied perspectives over time to reason about the surrounding in 3D. Also for the task of object pose estimation – the prediction of rotation and translation of objects in space, multi-view constraints can be beneficial [25]. In many computer vision applications, like robotic bin picking, augmented reality, and autonomous driving, multiple cameras or acquisitions are not

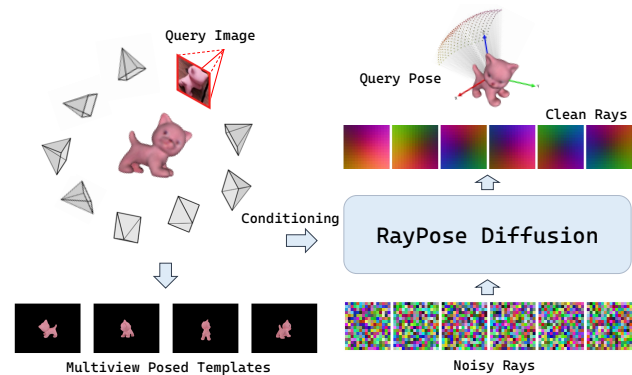


Figure 1. Given a novel object query image, our method accurately predicts the object’s 6D pose using a multiview diffusion model conditioned on a set of template images with known poses. Leveraging our proposed structured 2D pose maps, represented as bundles of rays, the diffusion model recovers the query object’s pose by progressively denoising these ray bundles.

available and the system needs to function even with a single monocular RGB image.

In object pose estimation literature, much effort has been put into learning other constraints, such as object appearance from visual data during training. Instance-based approaches [48, 53] therefore get their constraint from access to model appearance during training while category-level approaches [7, 22, 28, 54] use object shape and semantic priors. Despite the excellent results that benefit from deep learning, these approaches require training for every new object or object category from scratch and creating synthetic training data from a CAD model is also computationally expensive. To overcome per-object training, researchers have been working on unseen object pose estimation with access to textured CAD models during inference [6, 19, 26, 40, 42, 56, 60]. These advancements promise to overcome the scalability and flexibility hurdles of object-specific approaches.

These approaches are unable to access multiple views by input design and template approaches typically solve a

classification task first: which is the best template given an image query? Consecutive steps after template matching involve correspondence estimation, pose prediction, and optionally refinement [40, 46, 56]. Instead of finding the best possible posed template and then building pairwise correspondences, we think of the problem as an implicit bundle agreement among multiple views, using multiple template-query tuples to reason about 3D, with the advantage of having the template already posed.

Learning to reason about 3D from multiview inputs has been extensively studied in prior work [14, 25, 31, 45]. More recently, diffusion models have emerged as powerful tools for 3D reasoning, demonstrating remarkable generalization capabilities [2, 35, 36, 51, 52, 55, 58, 59]. Among them, PoseDiffusion [55] addresses the inverse problem of structure-from-motion by directly diffusing camera poses within a probabilistic diffusion framework, modeling the conditional distribution of poses given input images. Building upon this, recent work [59] introduces an overparameterization of camera poses using Plücker coordinates [44], representing a pose as 2D maps of ray direction and ray moment. This formulation is shown to be more compatible with diffusion processes and leads to improved accuracy in relative pose estimation. These approaches exhibit strong generalization and can infer relative camera poses even in novel scenes composed of entirely unseen images. Motivated by this capability, we propose to leverage a set of posed template images and a single query image to estimate the 6D pose of an object in the query by building on the strengths of multiview diffusion-based backbones.

Although diffusion models have shown success in relative camera pose estimation [55, 59], they are suboptimal for object pose estimation due to scale differences: camera poses are defined in a large world coordinate system, while object poses reside in a compact, object-centric space. To bridge this gap, we propose novel object-centric pose representations tailored for 6D object pose estimation. For rotation, we replace camera-centric Plücker coordinates with an object-centered formulation where rays are structured as a 2D image-aligned grid. For translation, we extend the Scale-Invariant Translation Estimation (SITE) framework [29] to generate a dense translation map. This object-centric parameterization enables more precise and disentangled reasoning about object-level 6D pose within the diffusion framework. Our structured pose diffusion framework takes a query image of an unseen object cropped from the scene and a set of posed images as templates, obtained by synthetic rendering from a CAD model, and generates precise 6D object pose predictions. We also propose a coarse-to-fine object pose estimation strategy by sampling the template with a narrower distribution based on the inputs. We evaluated our method on standard benchmark datasets from the pose estimation benchmark [50] and compared it to re-

cent methods for unseen object pose estimation. The performance of our method surpassed the results of the related works, and a detailed ablation study verified our design choices. This paper makes the following contributions:

- we formulate unseen object pose estimation as ray bundling problem between multiview templates and RGB query, which helps the network to capture the correlation between query and templates in 3D space.
- we introduce object-centric orientation and translation over-parameterization suitable for learning within diffusion framework.
- we propose a flexible diffusion-based 6D object pose framework for unseen object pose estimation that can be extended to a coarse-to-fine prediction by using different template sampling

## 2. Related Work

The benchmark for object pose estimation (BOP) [16] has long been dominated by traditional handcrafted feature matching methods based on point pair features (PPF). In recent years, learning-based approaches such as GDR-Net [53], ZebraPose [48], and SurfEmb [13] have surpassed traditional methods in performance. However, these methods are instance-specific and require training on each target object. More recently, the community has placed increasing emphasis on *unseen object pose estimation*, which focuses on estimating the pose of novel objects not encountered during training. Below, we describe different pose estimation pipelines used in this setting.

**Model-free approaches.** Without 3D model of target objects, Gen6D [34], OnePose [49], and OnePose++ [14] estimate its pose by flipping the structure from motion (SfM) at its head and matching features to align a posed object image to a test view. MFOS [27] uses posed template images as a model representation and establishes correspondences between the input query image patch and the rendered 3D bounding box of the object associated with each template image. While attractive, this leads to lower pose accuracy caused by this rough bounding box approximation of the object shape.

**Template-based approaches.** OSOP [47] and OVE6D [4] utilize a template object representation for 2D segmentation and coarse to fine matching. MegaPose [26] proposes a generic render-and-compare refinement strategy. GigaPose [40] performs a template-matching approach in two stages: 1) estimates out-of-plane rotation (2 DoF) by finding discriminative synthetic templates rendered from a CAD model and then 2) establishes correspondences to estimate the four remaining 4 DoF of the object pose.

**Foundation models with CAD model prior.** The idea of foundation models is a recent way to incorporate generic prior knowledge into pose estimation pipelines. Due to the need for an abundance of labeled data, all approaches are

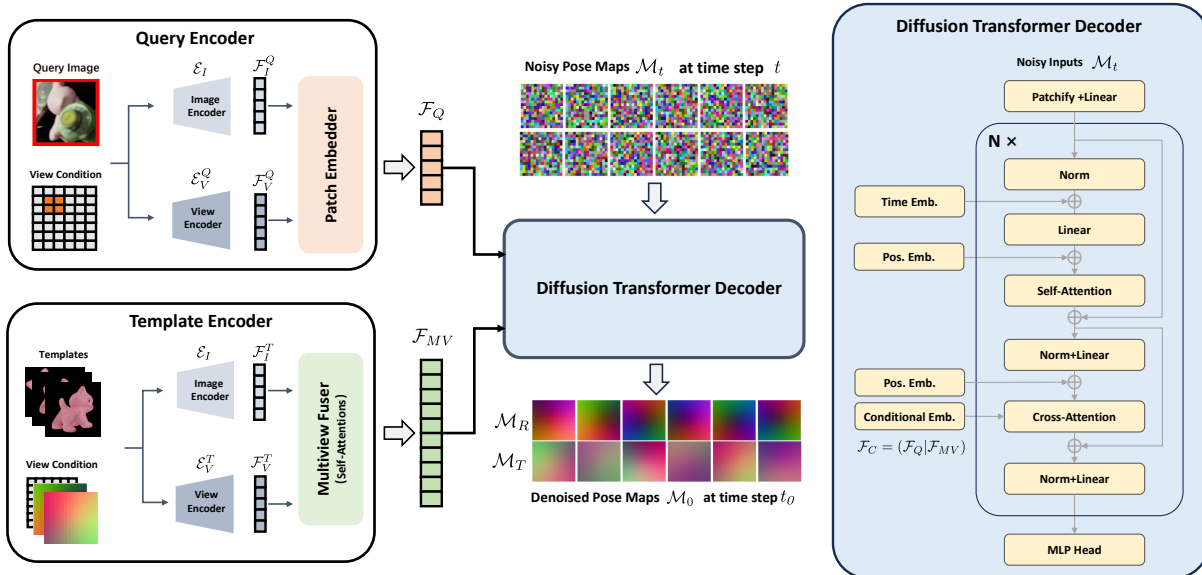


Figure 2. Pipeline overview of our method. We represent the 6D object pose using structured rotation and translation maps and employ a diffusion model to estimate the pose from random inputs. Given a query image of an unseen object and multiple template images with known object poses, our method first extracts query embeddings from a query encoder and multiview posed template embeddings from a template encoder. These embeddings serve as conditioning inputs for a diffusion transformer decoder, which is trained to denoise the object pose from random inputs. The model predicts the relative pose between the query and templates, from which the absolute 6D pose of the query object is reconstructed based on the known poses for the templates.

trained on synthetic data. Several methods learn generic 3D descriptors such as ZeroPose [6] and GCPose [60]. ZeroPose predicts poses utilizing the foundation models of ImageBind [11] and SAM [24] together with 3D-3D feature matching. GCPose [60] uses explicit knowledge of object symmetries. FoundPose [42] combines features from the foundation model DINOv2 [41] and bag-of-words retrieval for coarse matching and then uses featuremetric alignment for pose refinement. MatchU [18] and SAM6D [32] build discriminative descriptors by fusing RGB and depth information using transformers.

**Diffusion in Pose Estimation** Diffusion models reconstruct a target distribution from noise over multiple time steps, inherently capturing multimodal distributions. They are, by design, capable of capturing multimodal distributions as different noisy initializations can lead to different predictions during inference in the case of a multimodal distribution. RayDiffusion [59] denoises camera poses using ray parameterization for multiview estimation, avoiding COLMAP [45] in NeRF training but is unsuitable for object pose estimation. Object pose diffusion [17] diffuses poses in  $SE(3)$  space, excelling in synthetic data but struggling with unseen objects and real datasets. PoseDiffusion [55] addresses the SfM problem by diffusing camera poses across multiple images, implicitly performing bundle adjustment. Other methods include DiffusionNOCS [20], an RGB-D approach that diffuses NOCS maps for pose estimation, and Diff9D [33], which estimates 9D pose by diffusing

scale, translation, and rotation based on image conditioning.

## 3. Method

### 3.1. Method Overview

In this paper, we represent the 6D object pose using pose maps  $\mathcal{M}$ , which encode both orientation and translation. As illustrated in Fig. 2, we adopt a multiview diffusion transformer framework that learns to estimate object pose by denoising noisy pose maps conditioned on an input query object image and a set of reference images with known object poses (termed posed templates). We extract a query embedding  $\mathcal{F}_Q$  and the multiview template embedding  $\mathcal{F}_{MV}$  using the query and template encoders, respectively. Each encoder consists of an image encoder  $\mathcal{E}_I$  that extracts 2D image features, and a view encoder  $\mathcal{E}_V$  that encodes 6D object pose and/or 2D object location. Specifically, the multiview template features are fused using a Multiview Fuser to form the embedding  $\mathcal{F}_{MV}$ . A Diffusion Transformer Decoder is then trained to reconstruct the clean pose maps  $\mathcal{M}_0$  from noisy inputs  $\mathcal{M}_t$ , conditioned on both  $\mathcal{F}_Q$  and  $\mathcal{F}_{MV}$ . We train our model with two different template sampling strategies to obtain both coarse and fine pose predictors. For the coarse predictor, template viewpoints are randomly sampled independently of the query pose. For the fine predictor, the same model is trained with templates sampled from a narrower distribution centered around the query pose. This strategy enables coarse-to-fine pose inference during testing without any changes to the network architecture.

### 3.2. Object Pose Parameterization

The 6D object pose is defined by its rotation  $\mathbf{R} \in SO(3)$  and translation  $\mathbf{t} \in \mathbb{R}^3$ , representing the transformation from the object’s local coordinate frame to the camera coordinate system. While compact pose regression is desirable, it remains challenging for neural networks, especially in generic or cluttered scenes. Recent work [59] overparameterizes camera poses using ray directions and ray moments based on Plücker coordinates [44], which has proven effective for scene-level camera pose estimation. However, this formulation entangles camera intrinsics, rotation, and translation, limiting its effectiveness for object-level pose tasks. Specifically, inaccuracies in the predicted direction map can propagate to the translation component, hindering the centimeter-level precision required in object pose estimation. To overcome this, we propose a novel object-centric representation that maps the 6D object pose into separate 2D rotation and translation maps, enabling more accurate and disentangled learning.

**Rotation Parameterization** Camera pose estimation or novel view synthesis methods often model camera-centered rays, where rays originate from the camera center and pass through pixel coordinates in the image plane. In contrast, we introduce an object-centered ray representation, where the object center is treated as a *virtual pinhole camera*, emitting rays toward the camera coordinate system. Given the camera intrinsic matrix  $\mathbf{K} \in \mathbb{R}^{3 \times 3}$  and extrinsic parameters—rotation  $\mathbf{R} \in SO(3)$  and translation  $\mathbf{t} \in \mathbb{R}^3$ , a 3D object point  $\mathbf{x}$  is projected onto the image plane as  $\mathbf{u} = \mathbf{K}[\mathbf{R} \mid \mathbf{t}]\mathbf{x}$ . Instead of relying on this conventional image-based projection, we define a structured representation in which object-centered rays are mapped onto a normalized 2D square grid using a uniform intrinsic matrix, denoted as  $\mathbf{K} = \mathbf{K}_I$ . The set of direction vectors originating from the object center is represented as

$$\mathcal{M}_R = \{\mathbf{d}_1, \dots, \mathbf{d}_n\} \quad (1)$$

where each direction vector  $\mathbf{d}_i$  is normalized to unit length. This formulation enables us to map arbitrary rotation matrices  $\mathbf{R}$  onto a unique structured grid on the unit sphere surface. To construct the ray map, we uniformly select  $\{\mathbf{d}_i\}_{i=1}^n$  on the projected grid of the sphere surface, ensuring that each vector passes through the center of its corresponding grid cell. Consequently, we obtain a 2D grid map with the shape of  $(p \times p \times 3)$  as our rotation representation in the diffusion process. The illustration is given in the supplementary. Given the object-centered ray representation, we recover the rotation matrix  $\mathbf{R}$  by aligning the predicted ray directions with a predefined canonical frame. Let  $\mathcal{M}_R = \{\mathbf{d}_1, \dots, \mathbf{d}_n\}$  be the predicted ray set and  $\mathcal{M}_R^* = \{\mathbf{d}_1^*, \dots, \mathbf{d}_n^*\}$  the reference rays corresponding to an identity rotation  $\mathbf{R} = \mathbf{I}$ . The optimal rotation matrix  $\mathbf{R}^*$  is obtained by solving:

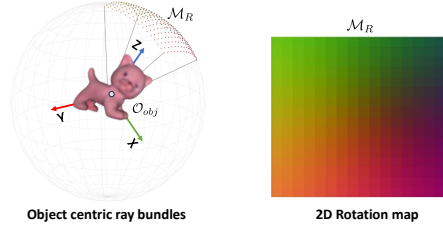


Figure 3. Visual illustration of the object-centric ray representation used for rotation prediction in our diffusion model. The rotation map  $\mathcal{M}_R$  is defined as a bundle of rays originating from the object center  $\mathcal{O}_{obj}$ , encoded as a 3-channel 2D map.

$$\mathbf{R}^* = \arg \min_{\mathbf{R} \in SO(3)} \sum_{i=1}^n \|\mathbf{R}\mathbf{d}_i^* - \mathbf{d}_i\|^2 \quad (2)$$

where  $\mathbf{R}$  is the relative rotation of the object with respect to the canonical frame. This problem can be solved using the Singular Value Decomposition (SVD) differentially, ensuring a valid rotation by enforcing  $\mathbf{R}^T \mathbf{R} = \mathbf{I}$ . This formulation allows for robust recovery of the object’s orientation and enables the diffusion process on 3D rotations from a structured 2D ray representation. **Translation Parameterization** A major challenge in estimating an object’s 6D pose from a single RGB image is minimizing translation error, particularly for previously unseen objects and scenes. Earlier work, SSD6D [23], estimates translation by locating the object centroid in 2D coordinates and comparing the bounding box scale with a pre-rendered template of the same rotation to determine object distance. However, this approach assumes the object center aligns with the bounding box center, making it sensitive to occlusion. Instance-level regression-based methods [29, 53] improve robustness by employing Scale-Invariant Translation Estimation (SITE), which predicts translation by computing the offset between the bounding box center and the object center. More recently, generalizable RGB-based methods [40, 42] estimate translation by establishing 2D correspondences between query and template images using a pre-trained feature matcher. While template depth can be rendered, these methods rely solely on one RGB image pair for correspondence extraction. In this paper, we extend SITE to a patch-level dense translation map. Given the object translation  $\mathbf{t} = [t_x, t_y, t_z]$  and the camera intrinsic matrix  $\mathbf{K}$ , the projected object centroid  $[\mathbf{o}_x, \mathbf{o}_y, 1]^T$  in image coordinates is computed as:

$$[\mathbf{o}_x, \mathbf{o}_y, 1]^T = \mathbf{K}\mathbf{t}. \quad (3)$$

We estimate the offset from each pixel  $(u, v)$  in the detected bounding box to the object centroid  $(\mathbf{o}_x, \mathbf{o}_y)$ , forming a dense normalized translation offset map:

$$\mathcal{M}_T = \left( \frac{u - \mathbf{o}_x}{w}, \frac{v - \mathbf{o}_y}{h}, \frac{t_z}{r_z} \right), \quad (4)$$

where  $w$  and  $h$  denote the bounding box width and height, and  $r_z$  is the zoom-in ratio of the bounding box. Similar

to the rotation map, we uniformly sample the pixels in the bounding box with the same shape of  $(p \times p \times 3)$  as the 2D translation map in the diffusion process. The 3D object translation is then recovered by back-projecting the estimated centroid offset using the camera intrinsics:

$$\mathbf{t}^* = r_z \cdot \mathbf{K}^{-1}[w \cdot \Delta \mathbf{o}_x + \mathbf{o}_x, h \cdot \Delta \mathbf{o}_y + \mathbf{o}_y, \Delta \mathbf{o}_z]^T. \quad (5)$$

To this end, we represent object pose as 2D pose maps  $\mathcal{M} = (\mathcal{M}_R, \mathcal{M}_T)$ . This pose representation decouples rotation and translation as well as the camera intrinsics, enabling the model to predict rotation and translation independently and enabling the use of a diffusion model to denoise the pose on two dense 2D maps.

### 3.3. Multiview Template Conditioned Diffusion

In our framework, we employ a multiview diffusion model to estimate object pose by conditioning it with the input query image and the posed templates. This network formulates the learning as a denoising process that gradually refines noisy inputs into the proposed structured pose maps.

#### 3.3.1. Diffusion Preliminaries

**Diffusion process.** The diffusion process consists of a forward (noising) and a reverse (denoising) process. Given a clean pose representation  $\mathcal{M}_0$  (either the rotation map  $\mathcal{M}_R$  or the translation map  $\mathcal{M}_T$ ), the forward process adds Gaussian noise over a fixed number of timesteps  $T$ . At each timestep  $t \in \{1, \dots, T\}$ , the pose map is perturbed as:

$$\mathcal{M}_t = \sqrt{\alpha_t} \mathcal{M}_0 + \sqrt{1 - \alpha_t} \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (6)$$

where  $\alpha_t$  is a noise schedule controlling the variance at timestep  $t$ .

**Denoising process.** The reverse process aims to recover the clean pose representation by learning to predict and remove the noise. A neural network  $\epsilon_\theta(\mathcal{M}_t, t, \mathcal{F}_c)$  is trained to estimate the noise  $\epsilon$  conditioned on an embedding  $\mathcal{F}_c$  that encodes the query and template information. The predicted pose is obtained by iteratively refining  $\mathcal{M}_t$  using the learned noise estimator:

$$\begin{aligned} \mathcal{M}_{t-1} = & \frac{1}{\sqrt{\alpha_t}} \left( \mathcal{M}_t - \sqrt{1 - \alpha_t} \epsilon_\theta(\mathcal{M}_t, t, \mathcal{F}_c) \right) \\ & + \sigma_t z, \quad z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \end{aligned} \quad (7)$$

where  $\sigma_t$  controls the stochasticity of the denoising step. This iterative process gradually refines the noisy pose representation into a structured output.

#### 3.3.2. Network Architecture

Our diffusion-based framework for 6D object pose estimation consists of three main blocks: (1) the **Query Encoder**, which extract the query image features, (2) the **Template Encoder**, which encodes and fuse the multiple posed template information, and (3) the **Diffusion Transformer Decoder**, which attends the query and templates and predicts

the denoised rotation and translation maps. The overall architecture is illustrated in Figure 2.

**Template Encoder.** The template encoder consists of three components: the *Image Encoder*  $\mathcal{E}_I$ , the *View Encoder*  $\mathcal{E}_V$ , and the *Multiview Fuser*  $\mathcal{E}_F$ . Given  $N$  object templates rendered from different viewpoints, we first employ a frozen DINOv2 [41] backbone to extract the image feature maps  $\mathcal{F}_T$ . Inspired by prior works [21, 36, 37, 61], which have demonstrated the effectiveness of Fourier encoding for camera rays in multiview scene understanding and reconstruction, we extend this idea to embed our structured pose information.

**View Encoder.** In our framework, the *View Encoder* uses three Fourier encoders to process the structured rotation map, translation map, and the normalized bounding box coordinates in the 2D image. This view embedding  $\mathcal{F}_v$  explicitly locates the objects by incorporating both their viewpoint in 3D space and their scale in 2D image coordinates, enforcing the network to learn implicit relationships across different views. The Fourier feature embedding for a scalar input  $x$  is computed as  $\gamma(x) = (x, \sin(2\pi Bx), \cos(2\pi Bx))$ , where  $B$  is a fixed frequency band that controls the resolution of the encoded features. Given the bounds  $K_r$ ,  $K_t$ , and  $K_c$  for rotation, translation, and object 2D coordinate maps, the total dimensionality of the view embedding is  $D_V = (2(K_r + K_t + K_c) + 1)d$ , where  $d$  is the number of frequency bands used for each Fourier encoding. This formulation ensures that each structured feature is transformed into a high-dimensional space, improving the model’s ability to capture fine-grained variations in object pose.

**Multiview Fuser.** The view embeddings  $\mathcal{F}_v$  and the image embeddings  $\mathcal{F}_I$  are concatenated and conditioned with view-level and patch-level positional encodings. This combined representation is then passed to the *Multiview Fuser*, which consists of  $\mathcal{N}_F$  self-attention layers to extract the fused multiview template embedding  $\mathcal{F}_{MV}$ . Following [59], we use DiT [43] blocks for cross-view information exchange and ensure that the network effectively aggregates object appearance and viewpoint-dependent geometric cues across all template images. The multiview embedding encodes the 2D and 3D priors of the object implicitly, which will be used as conditioning information to help reason the pose of the query image.

**Query Encoder.** The Query Encoder shares the same image and view encoders as the Template Encoder but processes only a single-view input. Since the pose of the query is not known, only the object’s 2D location in image coordinates is conditioned, instead of the full pose. Patch-level positional encoding is also incorporated to capture the spatial position of patches in the query image, facilitating implicit fusion with template patches. The resulting query embedding is denoted as  $\mathcal{F}_Q$ .

**Diffusion Transformer Decoder.** The fused multiview

template embedding  $\mathcal{F}_{MV}$  and the query embedding  $\mathcal{F}_Q$  serve as conditioning information in our diffusion transformer decoder, which denoises the 2D rotation and translation maps over a series of timesteps. At each step  $t$ , the noisy pose maps  $\mathcal{M}_t$  are processed alongside their conditional embeddings through a sequence of  $\mathcal{N}_D$  transformer-based diffusion blocks based on DiT [43]. Unlike DiT, which primarily uses only self-attention layers, we follow [30] to also incorporate cross-attention layers, allowing the query embedding to attend to the fused template embedding, enabling the feature exchange between different latent spaces that are constructed from a single query and multiple posed template views respectively. During training, for each time step  $t$ , the decoder learns to predict the noise component  $\epsilon_\theta(\mathcal{M}_t, t, \mathcal{F}_C)$ , where  $\mathcal{F}_C = (\mathcal{F}_Q | \mathcal{F}_{MV})$ . During inference, the pose maps are randomly initialized and iteratively denoised to obtain the final 2D rotation and translation maps. To enhance generalization, instead of directly predicting the absolute pose of the query, we estimate the relative pose between the query and templates. Specifically, the rotation maps represent the relative rotation from the template to the query, while the translation maps encode a relative depth scale between them:

$$\mathbf{t}_z^{rel} = \frac{\mathbf{t}_z^Q r_z^T}{\mathbf{t}_z^T r_z^Q} \quad (8)$$

where  $\mathbf{t}_z^{rel}$  is the predicted relative depth scale,  $\mathbf{t}_z^Q$  and  $\mathbf{t}_z^T$  are the ground-truth depths of the object center for the query and templates respectively, and  $r_z^Q$  and  $r_z^T$  are the zoom-in scales of the query and template. Finally, the query pose  $\mathbf{H}_Q = [\mathbf{R}_Q | \mathbf{t}_Q]$  is recovered from the predicted relative pose  $\mathbf{H}_{rel}$  and the ground-truth pose  $\mathbf{H}_{gt}$  of the templates:

$$\mathbf{H}_Q = \mathbf{H}_{rel} \mathbf{H}_{gt}. \quad (9)$$

Our multiview DiT-based Diffusion Transformer Decoder estimates the relative pose between a query and each template by conditioning on implicit geometric priors encoded in the templates and the query features. Each template view produces an independent pose hypothesis, supervised by its specific relative pose to the query. This design enables efficient, batch-wise probabilistic sampling from randomly selected templates, yielding diverse pose hypotheses in parallel.

### 3.4. Loss Functions

In our diffusion model, instead of training the network to estimate the noise, we follow prior work [59] and train the denoising network  $\epsilon_\theta(\mathcal{M}_t, t, \mathcal{F}_C)$  to learn the reverse diffusion process by predicting the original clean pose map  $\mathcal{M}_0$  conditioned on the noisy input  $\mathcal{M}_t$  at timestep  $t$ . The loss function is defined as:

$$\mathcal{L}_{diff} = \mathbb{E}_{t, \epsilon} [\|\mathcal{M}_0 - \epsilon_\theta(\mathcal{M}_t, t, \mathcal{F}_C)\|_2^2], \quad (10)$$

where  $t$  is uniformly sampled from  $[1, T]$  during training, and  $\mathcal{M}_t$  is the noisy version of the original pose map  $\mathcal{M}_0$  corrupted with Gaussian noise at timestep  $t$ . This training strategy naturally integrates task-specific constraints on the original rotation and translation maps.

**Rotation losses.** For the rotation map, we apply a pixel-level reconstruction loss to the target map  $\mathcal{M}_R^*$ :

$$\mathcal{L}_{recon}^R = \frac{1}{p^2} \|\mathcal{M}_R - \mathcal{M}_R^*\|_2^2, \quad (11)$$

where  $p$  denotes the spatial resolution of the rotation ray maps. Since each element in the rotation map represents a directional vector, we also employ a cosine similarity loss  $\mathcal{L}_{cos}^R$  to supervise ray directions. Given that the rotation map consists of structured ray bundles, we introduce an angle-consistency loss to enforce geometric coherence across adjacent rays in the predicted ray map. Given the predicted ray directions  $\{\mathbf{d}_i\}_{i=1}^n$  and the canonical ray set  $\{\mathbf{d}_i^*\}_{i=1}^n$ , we ensure that the predicted rays maintain consistent relative angles that reflect the intrinsic ray map geometry. For each pair of adjacent rays indexed by  $(i, j)$ , the relative angle is computed as  $\alpha_{ij} = \arccos(\mathbf{d}_i^\top \mathbf{d}_j)$ . Similarly, we precompute the reference angles  $\alpha_{ij}^*$  from the canonical rays corresponding to an identity rotation. The ray-consistency loss is then defined as:

$$\mathcal{L}_{reg}^R = \frac{1}{|\mathcal{N}_r|} \sum_{(i,j) \in \mathcal{N}_r} (\alpha_{ij} - \alpha_{ij}^*)^2, \quad (12)$$

where  $\mathcal{N}_r$  is the set of adjacent ray index pairs, and  $|\mathcal{N}_r|$  is its cardinality. This loss term encourages the network to respect intrinsic geometric constraints imposed by the projection, ensuring stable rotation estimation. The overall rotation loss is then formulated as a weighted combination of the reconstruction loss, cosine similarity loss, and ray-consistency loss:

$$\mathcal{L}^R = \lambda_{recon} \mathcal{L}_{recon}^R + \lambda_{cos} \mathcal{L}_{cos}^R + \lambda_{reg} \mathcal{L}_{reg}^R, \quad (13)$$

where  $\lambda_{recon}$ ,  $\lambda_{cos}$ , and  $\lambda_{reg}$  are hyperparameters that balance the contributions of each term. This formulation ensures that the predicted rotation map is both accurate at the pixel level and geometrically consistent with the camera’s intrinsic structure.

**Translation losses.** For the translation map, similar to the rotation ray reconstruction, we apply a pixel-level L2 loss to ensure accurate reconstruction of the dense translation offset maps:

$$\mathcal{L}_{recon}^T = \frac{1}{p^2} \|\mathcal{M}_T - \mathcal{M}_T^*\|_2^2. \quad (14)$$

Additionally, to explicitly supervise the final 3D translation prediction, we compute the object translation  $\mathbf{t}$  from the predicted translation map using the decoding formulation described in Eq. 5. We then impose an L1 loss on the

predicted translation components along each axis:

$$\mathcal{L}_{xyz}^T = \lambda_x |t_x - t_x^*| + \lambda_y |t_y - t_y^*| + \lambda_z |t_z - t_z^*|, \quad (15)$$

where  $\lambda_x, \lambda_y, \lambda_z$  are weighting the supervision strength. The overall translation loss is defined as the weighted sum of these two terms:

$$\mathcal{L}^T = \lambda_{\text{recon}} \mathcal{L}_{\text{recon}}^T + \lambda_t \mathcal{L}_{xyz}, \quad (16)$$

where  $\lambda_{\text{recon}}$  and  $\lambda_t$  are the hyperparameters to balance dense map-level supervision and explicit object-level translation regression.

**Overall loss function.** The final training loss is defined as a weighted sum of the rotation and translation losses:

$$\mathcal{L} = \lambda_{\text{rot}} \mathcal{L}^R + \lambda_{\text{trans}} \mathcal{L}^T, \quad (17)$$

where  $\lambda_{\text{rot}}$  and  $\lambda_{\text{trans}}$  are hyperparameters controlling the relative importance of rotation and translation losses.

### 3.5. Coarse to Fine Predictor Training Strategy.

We train our network with different template distributions to obtain both a coarse and fine predictor. For the coarse predictor, we use eight templates with randomly sampled poses from pre-processed scene-cropped images, ensuring diverse viewpoint coverage. The fine predictor employs online template sampling, augmenting templates for  $\pm 30^\circ$ ,  $\pm 5\text{cm}$  based on the query pose to enforce closer template-query alignment. Both predictors share the same network architecture but differ in template distributions during training. During inference, the coarse predictor provides an initial pose estimate, which is then refined by the fine predictor using a more localized template distribution.

## 4. Experiments

### 4.1. Experimental Setup

**Evaluation Metrics.** We adopt the metric Average Recall (AR) proposed by the Benchmark of Pose Estimation (BOP) [50]. The AR score is calculated with 3 pose-error functions: Visible Surface Discrepancy (VSD), Maximum Symmetry-Aware Surface Distance (MSSD), and Maximum Symmetry-Aware Projection Distance (MSPD). A pose is considered correct if the pose errors are within a pre-defined error threshold. The mean recall on the each error functions is computed over multiple error thresholds. The overall accuracy of a method is given by the Average Recall  $AR = (AR_{\text{VSD}} + AR_{\text{MSSD}} + AR_{\text{MSPD}})/3$ .

**Training and evaluation datasets.** We train our model on realistic synthetic datasets generated by Megapose [26], comprising approximately 2 million images rendered with BlenderProc [8] using objects from Google Scanned Objects [10] and ShapeNet [5]. For novel object pose estimation, we evaluate our method on five benchmark datasets:

LM-O [3], T-LESS [15], YCB-V [57], TUD-L [16], and IC-BIN [9]. Our evaluation is structured as follows: in Section 4.2, we compare our method with baselines on novel object pose estimation; in Section 4.3, we conduct ablation studies where we analyze design components, where we train and evaluate our method on LM-O dataset.

### 4.2. Compare to Baselines

We evaluate our method on five benchmark datasets: LM-O, T-LESS, TUD-L, IC-BIN, and YCB-V, which are unseen during training, and compare it with recent state-of-the-art methods that use only RGB images as input. All the methods use the same detection and segmentation results generated from CNOS [39] by default, except for OSOP [46]. As shown in Table 1, we analyze different setups, considering whether refinement and multi-hypothesis predictions are used. Our method achieves the highest average AR across all settings. In the single-prediction setting without refinement, it improves over the previous best method by 3.4% on average, with notable gains of 6.3% on LM-O and 9.2% on T-LESS. With refinement, our method continues to outperform the baselines, particularly excelling on TUD-L and LM-O. In the multi-hypothesis setting, it achieves the best performance on most datasets, especially with T-LESS dataset being improved by 3.7%. These results highlight the effectiveness of our approach in enhancing pose estimation by leveraging robust pose representations and a diffusion-based pipeline while ensuring strong generalization across diverse datasets.

### 4.3. Ablation Study

We conduct an ablation study on four key components of our approach: template ground-truth (GT) view embedding, the multiview setup, the fine-level predictor, and relative pose prediction. Each component is either removed or replaced with an alternative setup, and the results are summarized in Table 2.

**Fine predictor.** In the refinement stage, we apply both our fine predictor and MegaPose refinement. As shown in (1) – (3) of Table 2, our fine predictor improves performance by 6.3% compared to the coarse prediction. Notably, the fine predictor does not modify the network itself but instead utilizes a different template sampling strategy. Further improvements are achieved when incorporating an external refiner during the refinement stage.

**Relative pose prediction.** To enhance generalization across different scenes, camera intrinsics, and viewing conditions, we predict the relative pose between posed templates and the query, using the ground-truth template pose to infer the absolute query pose. In this ablation, we modify the network to directly predict the absolute poses of all template and query frames using a sequence of DiT blocks. As shown in (4) of Table 2, this modification results in a per-

Method	Refinement	Multi-hypo	LM-O	T-LESS	TUD-L	IC-BIN	YCB-V	Average
OSOP [47]	✗	✗	31.2	–	–	–	33.2	32.2
ZS6D [1]	✗	✗	29.8	21.0	–	–	32.4	27.7
MegaPose [26]	✗	✗	22.9	17.7	25.8	15.2	28.1	21.9
GenFlow [38]	✗	✗	25.0	21.5	30.0	16.8	27.7	24.2
GigaPose [40]	✗	✗	29.9	27.3	30.2	23.1	29.0	27.9
FoundPose [56]	✗	✗	39.6	33.8	46.7	<b>23.9</b>	45.2	37.8
Ours	✗	✗	<b>42.1</b>	<b>36.9</b>	<b>48.3</b>	21.8	<b>46.2</b>	<b>39.1</b>
<hr/>								
MegaPose [26]	✓	✗	49.9	47.7	65.3	36.7	60.1	51.9
GigaPose [40]	✓	✗	55.6	<b>54.6</b>	57.8	<b>44.3</b>	63.4	55.1
FoundPose [56]	✓	✗	55.7	51.0	63.3	43.3	66.1	55.9
Ours	✓	✗	<b>56.2</b>	53.8	<b>66.5</b>	41.6	62.8	<b>56.2</b>
<hr/>								
GenFlow [38]	✓	✓	56.3	52.3	68.4	45.3	63.3	57.1
MegaPose [26]	✓	✓	56.0	50.7	68.4	41.4	62.1	55.7
GigaPose [40]	✓	✓	59.9	57.0	64.5	46.7	66.3	58.9
FoundPose [56]	✓	✓	61.0	57.0	69.4	<b>47.9</b>	<b>69.0</b>	60.9
Ours	✓	✓	<b>62.2</b>	<b>59.1</b>	<b>70.2</b>	45.5	68.9	<b>61.2</b>

Table 1. Comparison of our method with the baselines with only RGB image inputs.

Method	GT Temp. Pose	Multiview	Relative Pose	Fine Predictor	AR
(1) <i>Fine+MegaPose</i>	✓	✓	✓	✓	<b>56.2</b>
(2) <i>Fine</i>	✓	✓	✓	✓	42.1
(3) <i>Coarse</i>	✓	✓	✓	✗	39.6
(4) <i>Absolute pose</i>	✓	✓	✗	✗	35.4
(5) <i>Single view</i>	✓	✗	✗	✗	32.5
(6) <i>w/o GT Template Pose</i>	✗	✓	✓	✗	27.8

Table 2. Ablation study for the key component of our method.

Method	Template Distribution	AR
Coarse	Random	65.58
Coarse	fixed	60.28
Fine	$\pm 90^\circ, \pm 10\text{cm}$	65.81
<b>Ours-Fine</b>	$\pm 30^\circ, \pm 5\text{cm}$	<b>67.29</b>
Fine	$\pm 15^\circ, \pm 3\text{cm}$	61.04

Table 3. Comparison of different template selection strategies for the fine predictor on LM-O dataset. The bold is our default setup.

formance drop of up to 10%, highlighting the effectiveness of relative pose prediction.

**Multi-view prediction.** We modify our pose map prediction head to perform single-view prediction, meaning the query pose is estimated directly without leveraging multiple templates. As shown in (5) of Table 2, this change leads to an approximately 18% performance drop, confirming that multi-view prediction enables the model to learn stronger implicit correspondences between the query and templates.

**Ground-truth template pose.** In our setup, we explicitly input template ground-truth pose maps as conditional information to help the network learn inherent correlations with the input query image. To evaluate its impact, we remove the ground-truth pose map from the input while retaining only the 2D position information, which is essential for predicting the relative distance from the camera. The ground-truth template pose is only used during inference to recover the absolute query pose. As shown in (6) of Table 2, removing the template ground-truth pose map results in a significant performance drop compared to the single-view setting. This finding underscores the importance of leveraging template pose priors in multi-view prediction, and also indicates the effectiveness of the view encoders.

#### 4.4. Effects of pose distribution of templates.

We evaluate our method’s sensitivity to template distribution by training with different template sampling strategies on the LM-O dataset, as shown in Table 3. For the coarse predictor, using randomly sampled templates yields better performance than the fixed-template setting. For the fine predictor, we examine different template distributions with varying pose and translation constraints. The best performance is achieved with a template distribution constrained to  $\pm 30^\circ$  in rotation and  $\pm 5\text{cm}$  in translation, which aligns with the mean error of the coarse predictor. A wider distribution ( $\pm 90^\circ, \pm 10\text{cm}$ ) performs similarly to the randomly sampled distribution used in the coarse predictor, while overly narrow constraints ( $\pm 15^\circ, \pm 3\text{cm}$ ) lead to a slight performance drop. These results underscore the importance of selecting appropriate template distributions for both coarse and fine predictors to balance generalization and fine-level accuracy.

## 5. Conclusion

In this paper, we introduced a structured representation for object pose that enables effective deployment of diffusion models for object 6D pose estimation. Instead of pairwise matching, we propose aligning object-centered rays across multiple posed templates. Our multiview diffusion model is conditioned on embeddings extracted from both the query and multiple posed template images using dedicated encoders. A coarse-to-fine strategy further improves pose accuracy without altering the model architecture. By integrating multiview diffusion models into the 6D pose estimation pipeline, our approach enables probabilistic reasoning from multiview inputs instead of explicitly reconstructing 3D. While our model achieves competitive performance, it relies on posed templates and accurate object detections. Future work would be of interest in relaxing these constraints to further generalize multiview-based pose estimation.

## References

- [1] Philipp Ausserlechner, David Habegger, Stefan Thalhammer, Jean-Baptiste Weibel, and Markus Vincze. Zs6d: Zero-shot 6d object pose estimation using vision transformers. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 463–469. IEEE, 2024. 8
- [2] Emmanuelle Bourigault and Pauline Bourigault. Mvd-iff: Scalable and flexible multi-view diffusion for 3d object reconstruction from single-view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7579–7586, 2024. 2
- [3] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *ECCV*, 2014. 7
- [4] Dingding Cai, Janne Heikkilä, and Esa Rahtu. Ove6d: Object viewpoint encoding for depth-based 6d object pose estimation. In *CVPR*, 2022. 2
- [5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 7
- [6] Jianqiu Chen, Mingshan Sun, Tianpeng Bao, Rui Zhao, Liwei Wu, and Zhenyu He. Zeropose: Cad-model-based zero-shot pose estimation, 2023. 1, 3
- [7] Yamei Chen, Yan Di, Guangyao Zhai, Fabian Manhardt, Chenyangguang Zhang, Ruida Zhang, Federico Tombari, Nassir Navab, and Benjamin Busam. Secondpose: Se (3)-consistent dual-stream feature fusion for category-level pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9959–9969, 2024. 1
- [8] Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Youssef Zidan, Dmitry Olefir, Mohamad Elbadrawy, Ahsan Lodhi, and Harinandan Katam. Blenderproc. *arXiv preprint arXiv:1911.01911*, 2019. 7
- [9] Andreas Drounoglou, Rigas Kouskouridas, Sotiris Malasiotis, and Tae-Kyun Kim. Recovering 6d object pose and predicting next-best-view in the crowd. In *CVPR*, 2016. 7
- [10] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *ICRA*, 2022. 7
- [11] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 3
- [12] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1
- [13] Rasmus Laurvig Haugaard and Anders Glent Buch. Surfemb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings. In *CVPR*, 2022. 2
- [14] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without CAD models. In *NeurIPS*, 2022. 2
- [15] Tomáš Hodaň, Pavel Haluza, Štěpán Obdržálek, Jiří Matas, Manolis Lourakis, and Xenophon Zabulis. T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. *WACV*, 2017. 7
- [16] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrike, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiri Matas, and Carsten Rother. Bop: Benchmark for 6d object pose estimation. In *ECCV*, 2018. 2, 7
- [17] Tsu-Ching Hsiao, Hao-Wei Chen, Hsuan-Kung Yang, and Chun-Yi Lee. Confronting ambiguity in 6d object pose estimation via score-based diffusion on se(3). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3456–3465, 2023. 3
- [18] Junwen Huang, Hao Yu, Kuan-Ting Yu, Nassir Navab, Slobodan Ilic, and Benjamin Busam. Matchu: Matching unseen objects for 6d pose estimation from rgb-d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10095–10105, 2024. 3
- [19] Junwen Huang, Hao Yu, Kuan-Ting Yu, Nassir Navab, Slobodan Ilic, and Benjamin Busam. Matchu: Matching unseen objects for 6d pose estimation from rgb-d images. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1
- [20] Takuya Ikeda, Sergey Zakharov, Tianyi Ko, Muhammad Zubair Irshad, Robert Lee, Katherine Liu, Rares Ambrus, and Koichi Nishiwaki. Diffusionnocs: Managing symmetry and uncertainty in sim2real multi-modal category-level pose estimation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7406–7413, 2024. 3
- [21] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver io: A general architecture for structured inputs & outputs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 5
- [22] HyunJun Jung, Shun-Cheng Wu, Patrick Ruhkamp, Hannah Schieber, Pengyuan Wang, Giulia Rizzoli, Hongcheng Zhao, Sven Damian Meier, Daniel Roth, Nassir Navab, et al. Housecat6d—a large-scale multi-modal category level 6d object pose dataset with household objects in realistic scenarios. *arXiv preprint arXiv:2212.10428*, 2022. 1
- [23] 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 (ICCV)*, pages 1521–1529, 2017. 4
- [24] 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. 3
- [25] 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, 2
- [26] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. In *CoRL*, 2022. 1, 2, 7, 8
- [27] JongMin Lee, Yohann Cabon, Romain Brégier, Sungjoo Yoo, and Jerome Revaud. Mfos: Model-free & one-shot object pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2911–2919, 2024. 2
- [28] Weihang Li, Hongli Xu, Junwen Huang, Hyunjun Jung, Peter KT Yu, Nassir Navab, and Benjamin Busam. Gce-pose: Global context enhancement for category-level object pose estimation. *arXiv preprint arXiv:2502.04293*, 2025. 1
- [29] Yisheng Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. CDPN: Coordinates-based disentangled pose network for real-time RGB six-degrees-of-freedom object pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 7678–7687, 2019. 2, 4
- [30] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchu Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 6
- [31] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. In *2024 International Conference on 3D Vision (3DV)*, 2024. 2
- [32] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. *arXiv preprint arXiv:2311.15707*, 2023. 3
- [33] Jian Liu, Wei Sun, Hui Yang, Pengchao Deng, Chongpei Liu, Nicu Sebe, Hossein Rahmani, and Ajmal Mian. Diff9d: Diffusion-based domain-generalized category-level 9-dof object pose estimation. *arXiv preprint arXiv:2502.02525*, 2025. 3
- [34] 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 *ECCV*, 2022. 2
- [35] Yuanxun Lu, Jingyang Zhang, Tian Fang, Jean-Daniel Nahmias, Yanghai Tsin, Long Quan, Xun Cao, Yao Yao, and Shiwei Li. Matrix3d: Large photogrammetry model all-in-one. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11250–11263, 2025. 2
- [36] Zhengzhe Luo, Yan Huang, Liang Wang, and Tieniu Tan. Zero-shot novel view and depth synthesis with multi-view geometric diffusion. *arXiv preprint arXiv:2303.17598*, 2023. 2, 5
- [37] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 5
- [38] Sunghill Moon, Hyeontae Son, Dongcheol Hur, and Sangwook Kim. Genflow: Generalizable recurrent flow for 6d pose refinement of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2024. 8
- [39] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Vincent Lepetit, and Tomas Hodan. Cnos: A strong baseline for cad-based novel object segmentation. In *ICCV*, 2023. 7
- [40] Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. GigaPose: Fast and robust novel object pose estimation via one correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 4, 8
- [41] 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. 3, 5
- [42] Evin Pinar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomáš Hodan. FoundPose: Unseen object pose estimation with foundation features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1, 3, 4
- [43] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 5, 6
- [44] Julius Plücker. *Analytisch-geometrische Entwicklungen*. GD Baedeker, 1828. 2, 4
- [45] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3
- [46] Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic. Osop: A multi-stage one shot object pose estimation framework. In *CVPR*, 2022. 2, 7
- [47] Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic. Osop: A multi-stage one shot object pose estimation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6835–6844, 2022. 2, 8
- [48] Yongzhi Su, Mahdi Saleh, Torben Fetzter, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari. Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. In *CVPR*, 2022. 1, 2
- [49] 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 *CVPR*, 2022. 2
- [50] Martin Sundermeyer, Tomáš Hodaň, Yann Labbe, Gu Wang, Eric Brachmann, Bertram Drost, Carsten Rother, and Jiří Matas. Bop challenge 2022 on detection, segmentation and pose estimation of specific rigid objects. In *CVPR*, 2023. 2, 7
- [51] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdifffusion: Enabling holistic multi-

- view image generation with correspondence-aware diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. [2](#)
- [52] Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdifusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. In *European Conference on Computer Vision (ECCV)*, 2024. [2](#)
- [53] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. GDR-Net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16611–16621, 2021. [1](#), [2](#), [4](#)
- [54] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. [1](#)
- [55] Jianyuan Wang, Christian Rupprecht, and David Novotny. PoseDiffusion: Solving pose estimation via diffusion-aided bundle adjustment. 2023. [2](#), [3](#)
- [56] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17868–17879, 2024. [1](#), [2](#), [8](#)
- [57] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter . Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *Robotics: Science and Systems*, 2018. [7](#)
- [58] Fuyang Zhang, Shitao Tang, Jiacheng Chen, and Yasutaka Furukawa. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2302.01329*, 2023. [2](#)
- [59] Jason Y. Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *International Conference on Learning Representations (ICLR)*, 2024. [2](#), [3](#), [4](#), [5](#), [6](#)
- [60] Heng Zhao, Shenxing Wei, Dahu Shi, Wenming Tan, Zheyang Li, Ye Ren, Xing Wei, Yi Yang, and Shiliang Pu. Learning symmetry-aware geometry correspondences for 6d object pose estimation. In *ICCV*, 2023. [1](#), [3](#)
- [61] Wenhao Zhao, Yuchao Dai, Hongdong Li, and Zhibo Rao. Towards zero-shot scale-aware monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. [5](#)