

Extreme Two-View Geometry From Object Poses with Diffusion Models

Yujing Sun*

The University of Hong Kong
Hong Kong SAR, China
yjsun@cs.hku.hk

Caiyi Sun*

Beijing Institute of Technology
Beijing, China
caiysisun.scy@gmail.com

Yuan Liu*

The Hong Kong University of Science and Technology
Hong Kong SAR, China
yuanly@ust.hk

Yuxin Ma

Shanghai Technology University
Shanghai, China
mayuxin@shanghaitech.edu.cn

Siu Ming Yiu

The University of Hong Kong
Hong Kong SAR, China
smyiu@cs.hku.hk

Abstract

Human has an incredible ability to effortlessly perceive the viewpoint difference between two images containing the same object, even when the viewpoint change is astonishingly vast with no co-visible regions in the images. This remarkable skill, however, has proven to be a challenge for existing camera pose estimation methods, which often fail when faced with large viewpoint differences due to the lack of overlapping local features for matching. In this paper, we aim to effectively harness the power of object priors to accurately determine two-view geometry in the face of extreme viewpoint changes. In our method, we first mathematically transform the relative camera pose estimation problem to an object pose estimation problem. Then, to estimate the object pose, we utilize the object priors learned from a diffusion model Zero123 [35] to synthesize novel-view images of the object. The novel-view images are matched to determine the object pose and thus the two-view camera pose. In experiments, our method has demonstrated extraordinary robustness and resilience to large viewpoint changes, consistently estimating two-view poses with exceptional generalization ability across both synthetic and real-world datasets. Code is released at <https://github.com/scy639/Extreme-Two-View-Geometry-From-Object-Poses-with-Diffusion-Models>.

Keywords: 6-DoF pose estimation, Camera pose, Ob-

ject pose, Diffusion model

1. Introduction

Relative camera pose estimation is a fundamental task in the realm of computer vision, with numerous applications spanning various cutting-edge techniques. However, estimating the relative poses of two views with extreme viewpoint changes presents a formidable challenge, especially when the co-visible regions of the two views are textureless. This difficulty stems from the lack of distinctive features to establish reliable correspondences between the two views, which is essential for accurate pose estimation. As the demand for high-quality 3D reconstruction [44], augmented reality [37], and other computer vision applications continue to grow, addressing this challenge is paramount for unlocking the full potential of these advanced techniques and pushing the boundaries of what is possible in the field of computer vision.

Though the widely adopted feature matching has difficulty in estimating camera poses of two views with extreme viewpoint changes, humans possess a remarkable ability to estimate extreme two-view poses when a common object is present in both views. This innate skill suggests that leveraging such object priors could hold the key to designing a new algorithm for two-view pose estimation, overcoming the limitations of current feature-matching techniques, and enabling more accurate and robust estimation of camera poses in challenging scenarios.

How to utilize such object priors remains an open problem. Recent works RelPose [81], RelPose++ [31] and SparsePose [65] utilize the object prior by training a trans-

* denotes joint first author

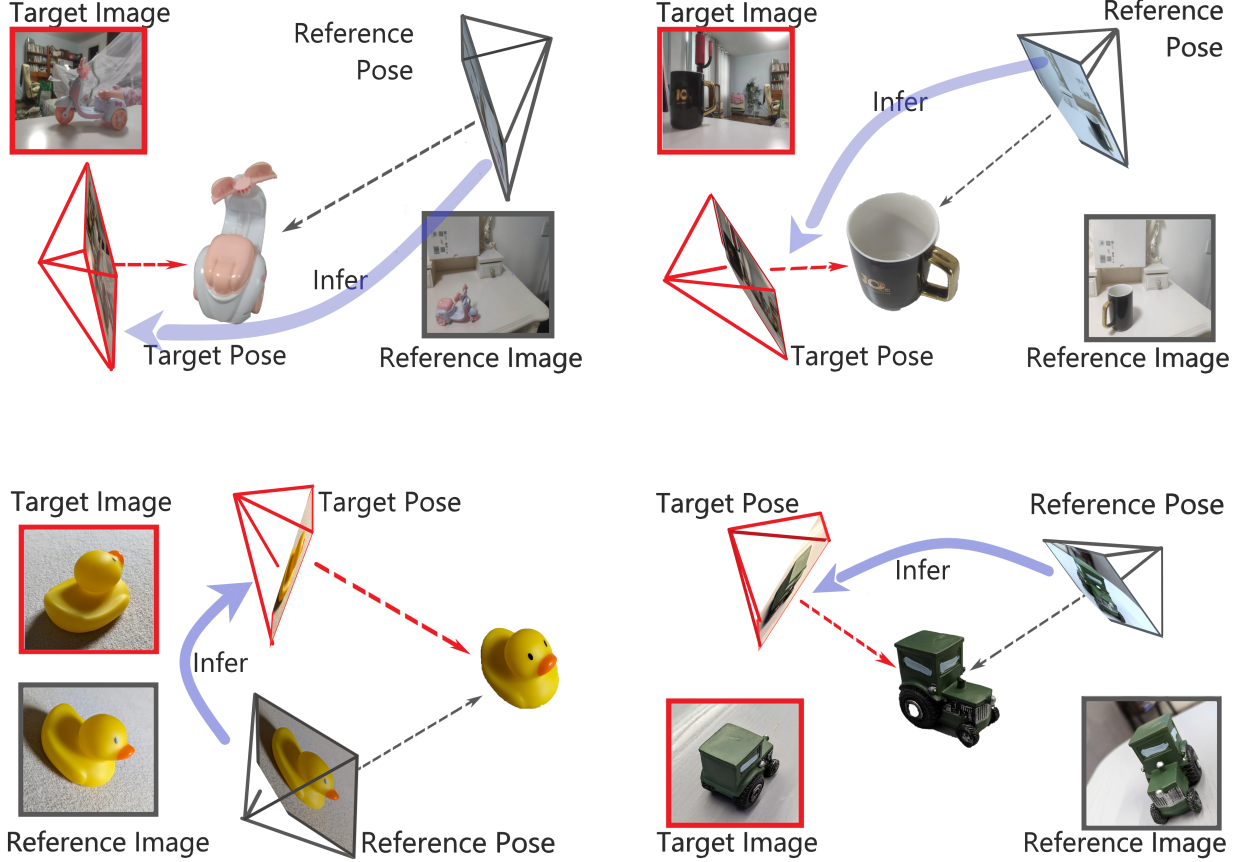


Figure 1. Our method can accurately infer the extreme relative camera pose of two images containing a co-visible object even without any overlap regions for correspondence estimation. Our method is based on a diffusion generative model to hallucinate the unseen sides of the object and match the hallucinated images with query images to estimate relative camera poses. The estimated extreme camera poses can be used in downstream applications, e.g. visual odometry.

former on the CO3D dataset [56] to regress or score the relative pose of two views of the same object. Though some promising results are achieved, we find that using the transformer to model the object prior leads to less generalization ability, especially on datasets of different domains from the CO3D dataset.

Recently, diffusion models [58] that have been trained on billion-scale datasets [62] have shown to be capable of generating high-quality images. Such a generation ability indicates these diffusion models learn meaningful priors over natural images. Based on the pretrained diffusion models, many recent methods [35, 36, 38, 63] have finetuned to generate diverse and high-quality images of arbitrary objects. This means these models encode robust object priors, which are much more generalizable than the transformer trained only on the CO3D dataset. Thus, these diffusion models are very promising tools for solving the extreme two-view camera pose estimation with object prior.

Since these diffusion models are designed for generating images of objects, how to utilize them in the two-view

pose estimation still remains unexplored. These diffusion models all assume that the image is strictly looking at the object when generating novel-view images for the object, as shown in Fig. 2. However, the input images may just contain the object but not strictly look at the object. Even worse, these diffusion models implicitly define a canonical object coordinate system in the model, which makes it even more difficult to find the poses and intrinsic matrices of the generated images in the canonical object coordinate system. Meanwhile, we have to derive the relative transformation between the camera coordinate system and the canonical object coordinate system used in diffusion models.

In this paper, we aim to build a novel framework, as shown in Fig. 1, to estimate extreme two-view camera poses by utilizing the object prior from a diffusion model. In order to utilize the diffusion mode in a pose estimation framework, we first propose a new formulation of two-view pose estimation by transforming the camera pose estimation problem into an object pose estimation problem. Then, on this object pose estimation problem, we generate multi-

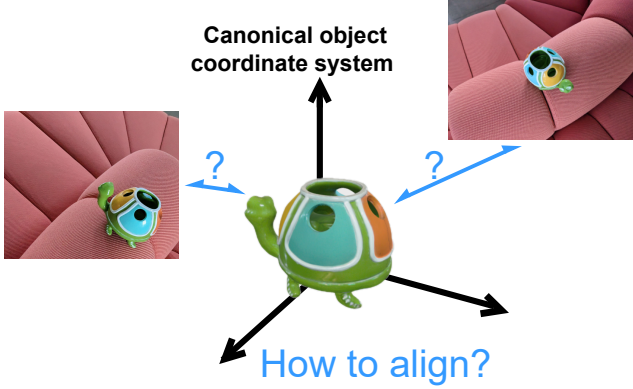


Figure 2. Challenges in applying the object prior from diffusion models, e.g. Zero123 [35], to relative pose estimation. On one hand, input images may not look at the object while Zero123 and common object pose estimations require the object to be located at the image center. On the other hand, Zero123 implicitly defines a canonical object coordinate inside, which brings difficulty in aligning the input images to this implicit canonical coordinate object system.

ple images of this object captured from different viewpoints by the Zero123 [35] model. On these generated images, we propose a novel and effective way to define their object poses and intrinsic matrices. Then, we match the other input image against the generated images to determine the plausible object poses, similar to Gen6D [37]. Finally, the determined object poses are then mapped back to determine the two-view camera pose of the input image pair.

Extensive experiments demonstrate that our method is able to predict accurate relative camera poses on image pairs with extreme viewpoint changes. On both synthetic and real datasets, our method outperforms baseline matching-based methods [59, 66] and the transformer-based pose regressor [31, 81] by a large margin. Our method shows strong generalization ability on the in-the-wild datasets and can be used in improving the closure optimization of visual odometry (VO).

In summary, the contribution of our paper is as follows:

- We propose a novel pose estimation algorithm for image pairs with extreme viewpoints change, based on diffusion models and image matching.
- To enable the utilization of object priors in diffusion models, we mathematically transform the relative pose estimation problem into an object pose estimation problem.
- The proposed framework significantly improves the accuracy of extreme two-view pose estimation on both synthetic and real datasets and shows promising results in combination with a VO method.

2. Literature Review

Feature Matching Based Pose Estimation In the scenarios of estimating poses from sets of images or video streams [39, 51], traditional methods typically revolve around finding correspondences between specifically designed local features [3, 40, 69]. Recently, with the development of deep learning techniques, neural networks have been employed and explored to boost the matching accuracy and robustness [10, 32, 59, 66]. Such matching-based pose estimation has been widely used in the downstream applications, including Structure-from-Motion (SfM) [60], Multi-View Stereo (MVS) [61] and Simultaneously localization and mapping (SLAM) [6, 46, 47]. However, in challenging scenarios where images are sparsely captured, these methods often struggle to accurately estimate poses due to the limited availability of effective features. Consequently, these approaches are not well-suited or perform poorly under such sparse-views setting.

Dense/multi-Views Based Pose Estimation Another main class of pose estimation is based on multiple or dense reference views. To achieve the goal, different priors are employed, such as category-specific knowledge, [26, 28, 41, 70, 75], and temporal locality for SLAM and VO applications [68, 72, 77].

More recently, category-agnostic priors [7, 19, 30, 37, 53, 67, 83, 84] have been presented to generalize beyond object categories and achieved improved performance on the dense-views setting. BundleTrack [73] and BundleSDF [74] can perform object pose tracking given a monocular RGBD video sequence. And PF-LRM [71] performed 3D reconstruction with joint predictions of pose and shape. Some methods [29, 54, 78] also performed 3D reconstruction at first, which needs multiple input images. However, methods falling in this category rely on the dense input reference views to provide enough objective priors, which cannot be obtained with a sparse-views setting.

Sparse-Views Based Pose Estimation The proposed method is most relevant to literature belonging to this category. It is noticeable that the limited overlapping between views makes these setups much more challenging. Many methods have been proposed to estimate object pose based on sparse views but require additional object 3D model [5, 11, 12, 15, 17, 50, 52, 55, 64, 76] or depth maps [21, 22, 33, 79]. When it comes to using only two RGB images for wide-baseline pose estimation, direct regression approaches [43, 57] typically do not work well. More recently, by unifying optimizing pairs of relative rotations, RelPose [81] achieved satisfactory accuracy on the CO3D [56] dataset. Thereafter, transformer-based approach Relpose++ [31] and SparsePose [65] further achieved improved accuracy. NOPE [48] achieved the same goal via discriminative embedding prediction with UNet. GigaPose [49] also presented to estimate objects posed

with a single image by utilizing a pre-trained model [38]. Meanwhile, E. Arnold et al. [2] proposed a map-free re-localization with one single photo of a scene to enable instant, metric-scaled re-localization, which is also relevant literature. ID-Pose [9] reversed the denoising process of the diffusion model [35] to estimate the relative pose. However, most of the above works conduct experiments (training and testing) within the same dataset. In experiments, we observe that their performance dropped extremely when applying the pre-trained model to other datasets, indicating a limited generalization ability. Compared with previous sparse-views or single-view-based approaches, the proposed framework demonstrates much stronger generalization capability and greatly outperforms them on in-the-wild datasets.

It is notable that another mainstream on single-view pose estimation assumes a fixed set of categories, such as humans [25, 42] and also other predefined object categories [1, 8, 16]. But our method is category-agnostic and depends on no category-related priors. Meanwhile, there are also works related to the sparse-views-based pose estimation but with a rather different objective [24].

Diffusion Models Image Diffusion Models [20, 35, 45, 58, 82] have been extensively employed in image generation tasks, leveraging neural networks to denoise images through the estimation and removal of noise values that are blended into the image. These models gradually remove noise from pure noise, resulting in the generation of clear and high-quality images. In order to enhance efficiency and stability, latent diffusion models [35, 58] focused on denoising the latent representations of the image, and thereby achieved reducing model complexity. Recent advancements, [45, 82], have enabled these models to be conditioned on additional inputs and provided improved control over the generated images. One notable example of a latent diffusion model is Stable Diffusion [58], which achieves high-quality image generation by training on an extensive dataset of image-text pairs. Most recently, by fine-tuning Stable Diffusion [58] on a collection of rendered pose-annotated images [13], Zero-1-to-3 [35] and Syndreamer [36] achieved impressive results in generating novel views of an object by utilizing an input image of the object and information about its relative pose. In this paper, we attempt to effectively take advantage of diffusion models to generate useful object priors and to further guide the follow-up pose inference process.

3. Method

Our method aims to estimate the relative camera pose of two images that have an extreme viewpoint change and a co-visible object as shown in Fig. 1. Given two images $I^{(1)}$ and $I^{(2)}$ with known intrinsic matrices $\mathbf{K}^{(1)}$ and $\mathbf{K}^{(2)}$ respectively, our target is to estimate the relative rigid transformation $[\mathbf{R}^{(12)}; \mathbf{t}^{(12)}]$ with $\mathbf{x}_c^{(2)} = \mathbf{R}^{(12)}\mathbf{x}_c^{(1)} + \mathbf{t}^{(12)}$

where $\mathbf{x}_c^{(1)}$ and $\mathbf{x}_c^{(2)}$ are the points in the coordinate systems of two cameras respectively. Since the extreme viewpoint change here prevents us from building reliable correspondences to solve the camera pose, we estimate the relative camera poses utilizing object priors.

The overview of our pipeline is shown in Fig. 3. We first transform the input images into object-centric images $I_v^{(1)}$ and $I_v^{(2)}$ in Sec. 3.1. Then, we feed the image $I_v^{(1)}$ to a diffusion model [35] to generate a set of images on novel viewpoints $\mathcal{G} = \{I_g^{(i)} | i = 1, 2, 3, \dots, N\}$ in Sec. 3.2. Finally, we match the $I_v^{(2)}$ with the generated image set \mathcal{G} to determine the relative pose between two images in Sec. 3.3. In the following, we first derive the underlying geometric relationship between the input image and the object-centric image.

3.1. Object-Centric Images

Virtual object-centric camera. We assume that the object is segmented out for both images. On each image, we denote the 2D object center as \mathbf{c} and the object is bounded by a 2D square bounding box of size s , as shown in Fig. 4. Since the object may not be located at the center of the input image, we need to create a new virtual camera that roughly looks at the object by rotation. In the following, we derive the construction of the intrinsic matrices $\mathbf{K}_v^{(1)}$ and $\mathbf{K}_v^{(2)}$, the rotation matrices $\mathbf{R}_v^{(1)}$ and $\mathbf{R}_v^{(2)}$, and corresponding object-centric images $I_v^{(1)}$ and $I_v^{(2)}$. The virtual cameras are constructed for two input views respectively, so that in the rest of this section, we omit the superscript for simplification.

Rotation matrix. The rotation $\mathbf{R}_v = \mathbf{R}_{v,y}\mathbf{R}_{v,x}$ is computed by rotating around the y -axis $\mathbf{R}_{v,y}$ and x -axis $\mathbf{R}_{v,x}$ as shown in Fig. 4 to make the camera look at the 2D object center \mathbf{c} , which means that

$$\mathbf{x}_v = \mathbf{R}_v\mathbf{x}_c, \quad (1)$$

where \mathbf{x}_v is the coordinate of a point in the new virtual camera while \mathbf{x}_c is its coordinate in the input camera.

Intrinsic matrix. Assume that the virtual object-centric camera is constructed to look at the object center \mathbf{c} with a new virtual image size of s_v . We determine the new focal length $f_v = s_v\sqrt{f^2 + \|\mathbf{c}\|_2^2}/s$ for this virtual camera, where f is the focal length of the input image. Then, we construct a new intrinsic matrix \mathbf{K}_v with a focal length f_v and a principle point $(s_v/2, s_v/2)$.

Object-centric warping. Since the camera is transformed by a pure rotation, according to multiview geometry [18], the resulting object-centric image and the input image are related by a homography transformation \mathbf{H}

$$\mathbf{H} = \mathbf{K}_v\mathbf{R}_v\mathbf{K}^{-1}. \quad (2)$$

We use this homography transformation to warp the input image I to get the object-centric image I_v . We transform

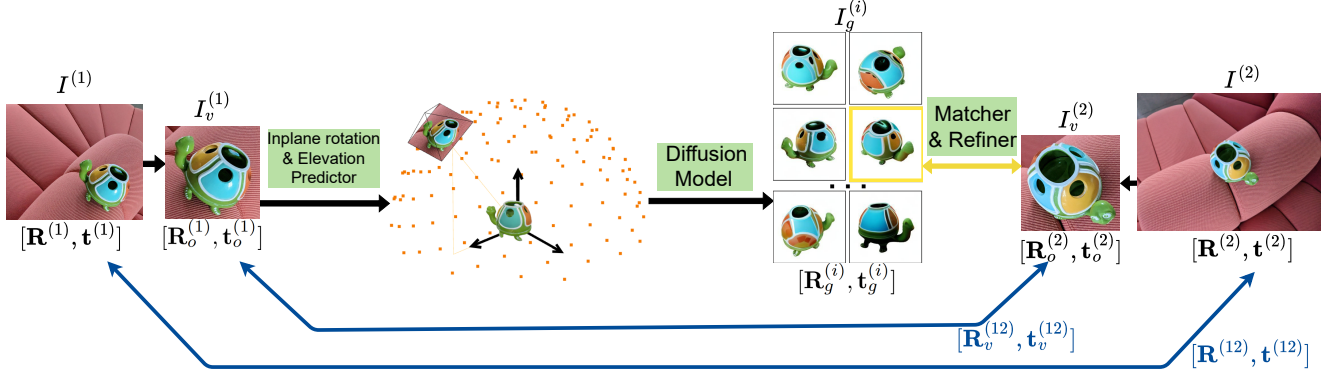


Figure 3. The overview of our pipeline.

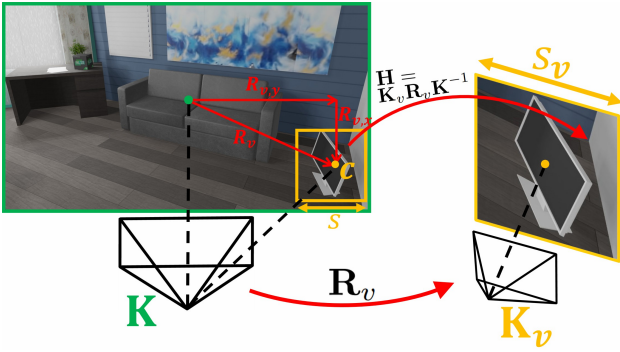


Figure 4. Transformation to object-centric images. It happens when an input image is not looking at the target object. Hence, we transform the image so that it looks at the center of the target object by a homography transformation, which leads to a new pose and a new intrinsic matrix.

both images into object-centric images $I_v^{(1)}$ and $I_v^{(2)}$. Therefore, we only need to work on object-centric images and determine their relative transformation $[\mathbf{R}_v^{(12)}; \mathbf{t}_v^{(12)}]$ with

$$\mathbf{x}_v^{(1)} = \mathbf{R}_v^{(12)} \mathbf{x}_v^{(2)} + \mathbf{t}_v^{(12)}, \quad (3)$$

which can be utilized to determine $[\mathbf{R}^{(12)}; \mathbf{t}^{(12)}]$ by Eq. 1.

3.2. Object Priors from Diffusion Models

In this section, we aim to utilize the diffusion model Zero123 [35] to generate novel-view images $\mathcal{G} = \{I_g^{(i)} | i = 1, 2, 3, \dots, N\}$ of the object. These generated images will be used in the next section to determine the relative pose.

Zero123 [35]. By fine-tuning the large-scale Stable Diffusion [58] model on millions of objects, Zero123 [35] learns strong object priors and can generate highly plausible novel-view images from a given image of an object. Given an input image of the object, a delta azimuth, and a delta elevation, Zero123 is able to generate an image on a rotated viewpoint with the given azimuth and elevation changes from the viewpoint of the input view, as shown in

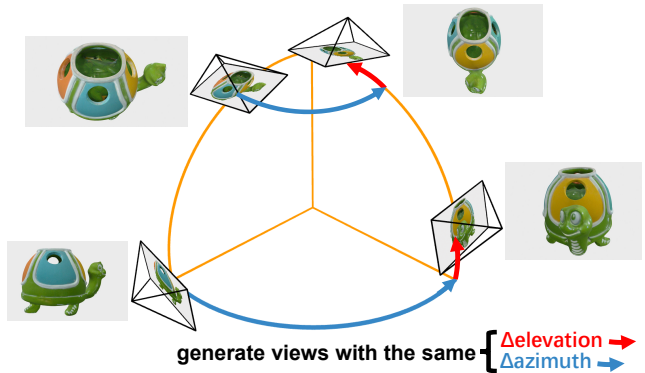


Figure 5. Zero123 [35] is able to rotate a given image by a given $\Delta\text{azimuth}$ and a $\Delta\text{elevation}$ in the object canonical coordinate. However, Zero123 assumes that the Y+ direction (UP) of the image is aligned with gravity direction. Moreover, given the same rotation angle ($\Delta\text{azimuth}$ and $\Delta\text{elevation}$), the actual rotated angle is related to the elevation angle of the input image. This requires us to estimate the inplane rotation and the canonical azimuth of the input image.

Fig. 5. Zero123 assumes the input image is correctly oriented, which means that the canonical up direction of the object is consistent with the Y direction of the camera coordinate system of the input image. Note that the delta azimuth and delta elevation in Zero123 are also based on the canonical azimuth and elevation of the object. As shown in Fig. 5, if we apply Zero123 to generate images of the same delta azimuth and delta elevation but on two different images of different canonical elevations, we rotate different angles in the 3D space. We call the coordinate implicitly defined in Zero123 as the canonical object coordinate \mathbf{x}_o . In the following, we illustrate how we generate a set of novel-view images using the image $I_v^{(1)}$.

Canonical object viewpoint of $I_v^{(1)}$. We want to use the Zero123 model to generate a set of novel-view images whose viewpoints are evenly distributed on the upper hemisphere, as shown in Fig. 6 (a). Note that these viewpoints

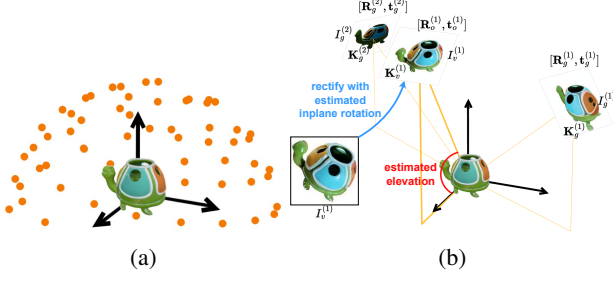


Figure 6. (a) We want to generate images on the viewpoints in the canonical object coordinate system. (b) We estimate the inplane rotation and elevation of the image $I_v^{(1)}$ and generate images on the given viewpoints using Zero123. For all the generated images and image $I_v^{(1)}$, we estimate their new intrinsic matrix and their extrinsic poses in this canonical object coordinate system.

are defined in the canonical object coordinate system with predefined elevations and azimuths. Thus, in order to compute the delta azimuths and delta elevations between $I_v^{(1)}$ and the target novel-view images, we need to compute the canonical azimuth and elevation of the image $I_v^{(1)}$. Meanwhile, we notice that the input image $I_v^{(1)}$ may not be correctly oriented with an in-plane rotation as shown in Fig. 6 (b), so we also need to determine an in-plane rotation for $I_v^{(1)}$. Since the azimuth can be arbitrarily defined, we directly set the canonical azimuth of $I_v^{(1)}$ to 0° . Thus, we only need to determine the in-plane rotation and the elevation of $I_v^{(1)}$. We extend the elevation determination algorithm of One-2-3-45 [34] to determine both the elevation and the in-plane rotations, generate a set of hypotheses on the elevations and the in-plane rotations, and then select the elevation and the in-plane rotation that can produce the most number of points in the triangulation. More details are included in the supplemental materials. With the estimated azimuth, elevation, and in-plane rotation, we rectify the input image $I_v^{(1)}$ with the in-plane rotation to be correctly oriented, compute the delta azimuths and elevations, and finally generate predefined novel-view images using Zero123 using the rectified $I_v^{(1)}$.

From viewpoints to object poses. In order to utilize the generated images to estimate the relative pose between $I_v^{(1)}$ and $I_v^{(2)}$, we need to explicitly determine the intrinsic matrices and object poses of the generated images and $I_v^{(1)}$ in the canonical object coordinate system. We already derived the intrinsic matrix $\mathbf{K}_v^{(1)}$ for $I_v^{(1)}$ and we assume all the generated images have the same intrinsic as $I_v^{(1)}$, i.e. $\mathbf{K}_g^{(i)} = \mathbf{K}_v^{(1)}$. To get the object poses $[\mathbf{R}_g^{(i)}, \mathbf{t}_g^{(i)}]$ and $[\mathbf{R}_o^{(i)}, \mathbf{t}_o^{(i)}]$ of the generated images and $I_v^{(1)}$, we already know their elevations and azimuths in the canonical object coordinate system, which helps us to determine the rotations $\mathbf{R}_g^{(i)}$ and $\mathbf{R}_o^{(i)}$. Then, assuming that all the generated

images and $I_v^{(1)}$ look at the origin of the canonical object coordinate, we only need to determine the distance d to compute the offsets $\mathbf{t}_g^{(i)}$ and $\mathbf{t}_o^{(i)}$. We further assume that the target object is located at the origin inside a sphere with a unit-length radius and the projections of the unit sphere onto every image just inscribe the boundaries of these images, which helps us determine the distance from the origin to the camera center. Thus, with $[\mathbf{R}_g^{(i)}, \mathbf{t}_g^{(i)}]$ and $[\mathbf{R}_o^{(i)}, \mathbf{t}_o^{(i)}]$, we have

$$\mathbf{x}_v^{(1)} = \mathbf{R}_o^{(1)} \mathbf{x}_o + \mathbf{t}_o^{(1)}, \quad (4)$$

$$\mathbf{x}_g^{(i)} = \mathbf{R}_g^{(i)} \mathbf{x}_o + \mathbf{t}_g^{(i)}. \quad (5)$$

where \mathbf{x}_o is the object coordinate and $\mathbf{x}_g^{(i)}$ is the camera coordinate of i -th generated image.

3.3. Object pose estimation

In this section, we match the image $I_v^{(2)}$ with the posed generated image set $\mathcal{G} = \{(I_g^{(i)} \mathbf{R}_g^{(i)}, \mathbf{t}_g^{(i)} | i = 1, \dots, N)\}$ to determine the object pose $[\mathbf{R}_o^{(2)}, \mathbf{t}_o^{(2)}]$ of $I_v^{(2)}$ with

$$\mathbf{x}_v^{(2)} = \mathbf{R}_o^{(2)} \mathbf{x}_o + \mathbf{t}_o^{(2)}. \quad (6)$$

Then, by combining Eq. 4 and Eq. 6, we can get Eq. 3 to compute the relative camera pose between the object-centric images $I_v^{(1)}$ and $I_v^{(2)}$.

To achieve this purpose, we follow the work Gen6D [37] to first select a generated image with the most similar viewpoint and then apply a feature volume-based refiner to iteratively refine the final pose.

Viewpoint selection. We compare $I_v^{(2)}$ with every generated image $I_g^{(i)}$ to compute a matching score. The generated image with a higher score is supposed to be similar to the input image. Similar to [37], we also share the information among all generated views by global normalization layers and transformers before computing the final matching scores. We regard the viewpoint of the most similar generated image as the viewpoint of $I_v^{(2)}$, which provides a coarse estimation of the object pose $[\mathbf{R}_o^{(2)}, \mathbf{t}_o^{(2)}]$. We then refine the estimated pose with a feature volume-based refiner.

Feature volume-based pose refinement. We use the feature volume-based refiner in Gen6D [37]. Given the coarse object pose estimation, we first find the K nearest generated images as reference images for the current refinement step. A 3D feature volume will be constructed using the 2D CNN features extracted from $I_v^{(2)}$ and all reference images. Then, a 3D CNN is applied to regress a pose residual in the form of a similarity transformation to update the input coarse estimated pose. This refinement process is iteratively applied for several steps to get an accurate estimation of the object pose $[\mathbf{R}_o^{(2)}, \mathbf{t}_o^{(2)}]$.

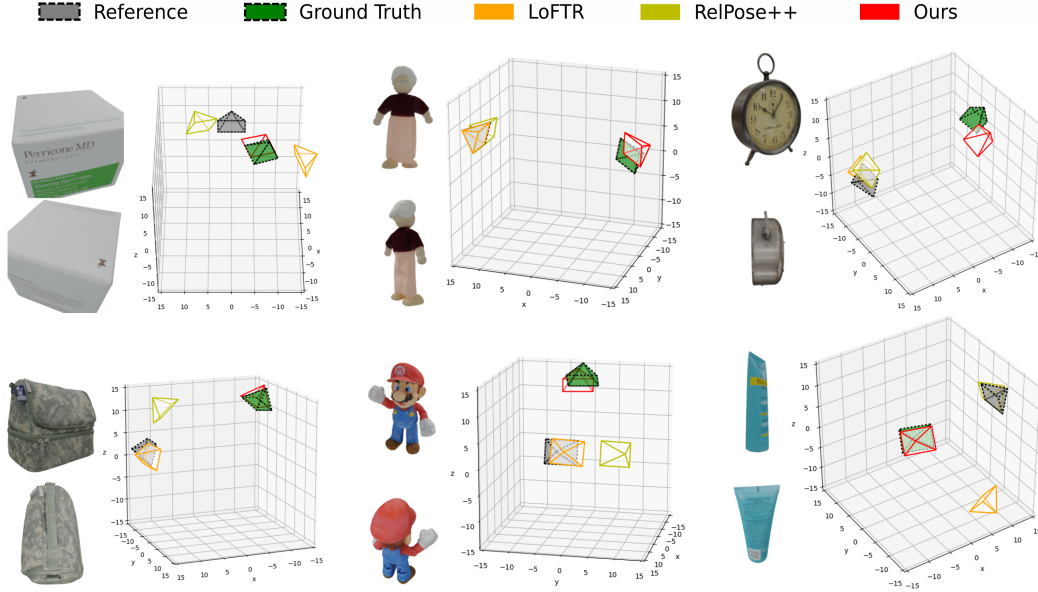


Figure 7. Visual Comparisons on the GSO dataset [14].

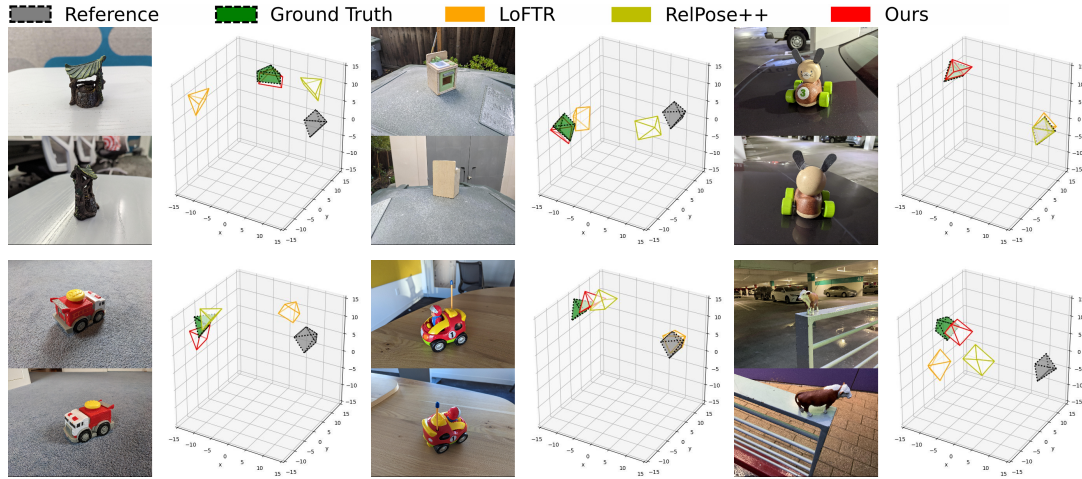


Figure 8. Visual Comparison on the Navi dataset [23].

4. Experiments and Discussions

4.1. Experimental Setups

Implementation Details We utilize Zero123 to generate 128 novel view images I_g for each reference image. When estimating inplane rotation and elevation, the number of diffusion inference steps is 75, and when generating I_g , this number is 50. During the inference of a query image, the selector selects from all 128 generated images. In the refinement stage, we apply the refiner iteratively 3 times.

Baselines To evaluate the effectiveness of the proposed framework, we compare our method with baseline methods including RelPose++ [31], ID-Pose [9], Relative Pose Regression (RPR) [2], LoFTR [66], SIFT [40]-

ZoeDepth [4]-based PnP method and Procrustes method. 3DAHV [85], RelPose++ [31] and RPR [2] regress the relative poses between images. 3DAHV [85] introduces a 3D-aware verification module that explicitly applies 3D transformations to the 3D object representations learned from two input images. It evaluates multiple pose hypotheses and selects the most reliable one as the predicted pose. RelPose++ is a transformer-based model specially designed for objects by training on the CO3D [56] dataset. RPR first regresses the 3D-to-3D correspondences from image features and then solves for the pose with a differentiable solver, which is trained end-to-end on large-scale image pair datasets of common scenes. LoFTR [66] is a dense 2D-to-2D correspondence estimator with a transformer, on

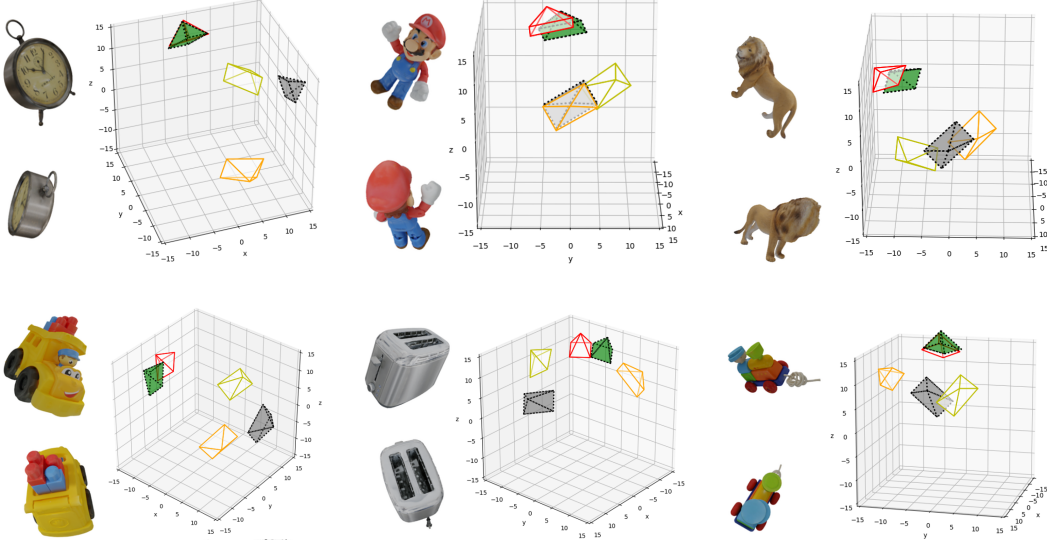


Figure 9. Visual Comparison on the rotated GSO dataset [14].

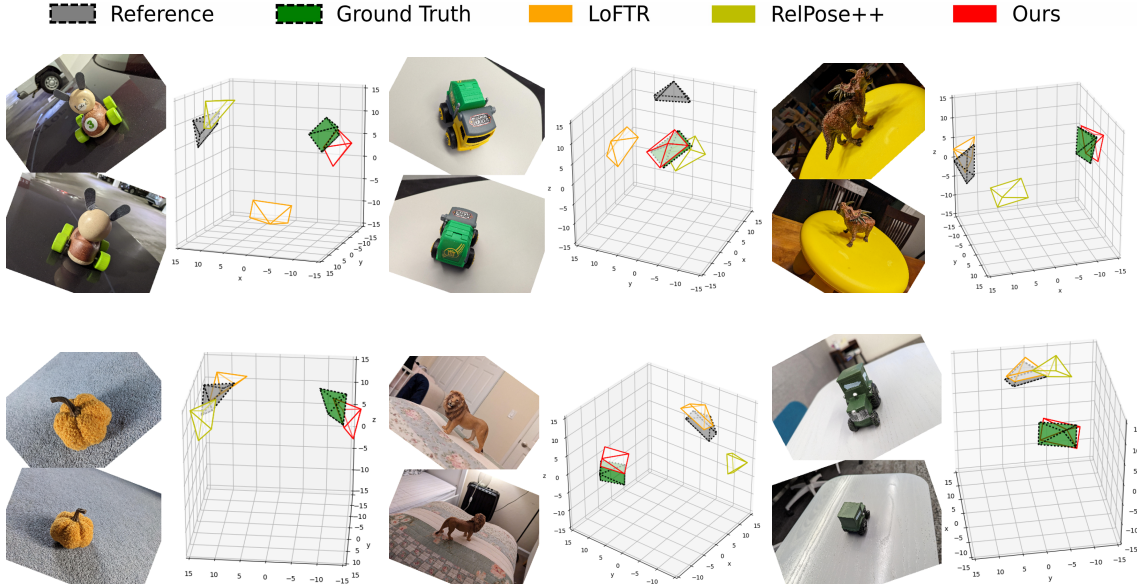


Figure 10. Visual comparisons on the rotated NAVI dataset [23]

which we solve the relative pose by the RANSAC algorithm. SIFT [40]-ZoeDepth methods first build 2D-to-2D correspondences by SIFT matching and then compute the relative poses by transforming these 2D-to-2D correspondences into 2D-to-3D (PnP-based) or 3D-to-3D (Procrustes-based) correspondences. ID-Pose [9] reverses the denoising process of Zero123 [35] to estimate the relative pose. We adopt the official implementation of 3DAHV, RelPose++, ID-Pose and LoFTR with their officially pretrained models. While for RPR and SIFT-ZoeDepth-based method, we adopt the implementation from [2].

Metrics To evaluate the predicted 6-DoF poses $\{R_i, t_i\}$ of an object in a target view, we report rotation ac-

curacy and translation accuracy with the standard two-view geometry evaluation metrics, following SuperGlue [59] and OrderAwareNet [80]. As 3DAHV [85] can not estimate the translation, we only report its rotation accuracy.

4.2. Datasets

In order to verify the effectiveness and generalization ability of our framework, we evaluate our method and baseline methods on two datasets, the GSO [14] dataset and the Navi [23] dataset. The GSO dataset is a synthetic object dataset containing about 1k 3D-scanned household objects. From the GSO dataset, we select 23 objects and render 21 images on each object for evaluation. The rendered image

pairs are set to have extremely large viewpoint changes. The Navi dataset is a real object dataset, which contains images of the same object captured in different environments and viewpoints with accurate calibrations for both intrinsic and extrinsic matrices. The Navi dataset contains 36 objects, among which we select 27 objects for evaluation. Besides, to show our method is more robust to the inplane rotation, we also add randomly inplane rotation between -45 to 45 degrees to each input image for GSO and Navi dataset. We call the resulting datasets the rotated GSO dataset and the rotated Navi dataset.

4.3. Metrics

To evaluate baseline methods and our method, we choose the accuracy under a specific degree as the metrics, which is similar to [31, 66, 81]. For the rotation accuracy, we compute the relative rotation between the ground-truth rotation \mathbf{R}_{gt} and the predicted rotation \mathbf{R}_{pr} by $\mathbf{R}_{gt}^T \mathbf{R}_{pr}$ and then transform this rotation into the axis-angle form. The resulting rotation angle of this rotation matrix is regarded as the rotation error in angle. The translation contains a scale ambiguity, so we compute the translation error as the angle between the normalized ground-truth translation \mathbf{t}_{gt} and the normalized predicted translation \mathbf{t}_{pr} by $\arccos \mathbf{t}_{gt}^T \mathbf{t}_{pr}$. We report translation and rotation accuracy under 15° and 30° .

4.4. Comparisons with baseline methods

Comparisons. We evaluate our method and baselines on two datasets, GSO [14] and Navi [23]. The qualitative results are shown in Fig. 7 and Fig. 8. While quantitative comparison results are provided in Table 1. As we can see, the matching based method LoFTR [66] and SIFT [40] does not perform well on both datasets, because the extreme viewpoint change results in very small overlap regions to estimate correspondences. While three regression-based methods, RPR [2], 3DAHV [85] and RelPose++ [31] show limited generalization ability. For ID-Pose [9], which also utilizes the pre-trained diffusion model [35], gives lower accuracy compared to ours. Additionally, it requires over 30 seconds per query image, as it requires multiple forward and backward passes of Zero123 for each query image. Our method, which effectively and efficiently combines the object prior from diffusion models with the object pose estimator, achieves the best performance on all metrics, which outperforms baselines by a large margin. More comparisons on the rotated Navi [23] Dataset, as well as visualization of correspondences of LoFTR/SIFT ZoeDepth methods and of depth maps estimated by ZoeDepth, are included in the supplemental materials.

Robustness to inplane rotations. We also provided results on the rotated GSO and Navi datasets to show the robustness to inplane rotations. The quantitative results are shown in Table 2 and qualitative results are shown in

Fig. 9 and Fig. 10. As we can see, 3DAHV [85] and Relpose++ [31] show worse robustness to the inplane rotations, while our method achieves stronger robustness by estimating an inplane rotation in the pipeline.

4.5. Analysis

In this section, we conduct analysis on each module proposed in this paper to demonstrate their effectiveness.

Inplane rotation and elevation prediction. As stated in Sec. 3.2, we need to estimate an inplane rotation and an elevation for the input image $I_v^{(1)}$ so that we can correctly generate images on the desired viewpoints by Zero123 [35]. To show this necessity, we directly treat the inplane rotation as 0° and elevation as 45° , and evaluate the performances on the rotated datasets. The results are listed in Table 3. As the results shown, it is vital to estimate both inplane rotation and elevation to get a correct relative pose estimation in our pipeline.

The number of generated images. By default, we generate 128 images with Zero123 to estimate the relative pose. To show how our method performs with different numbers of generated images, we conduct an experiment on both test datasets, as shown in Table 4. As we can see, the performance is almost the same with only 64 or 32 images but drops reasonably with only 8 or 16 images due to the sparsity.

The number of refinement iterations. We further study how the number of iterations in the pose refinement in Sec. 3.3 affects the quality of two-view pose estimation. We show the results on the GSO dataset with refinement number ranging from 0 to 8. As we can see, the performance increases with 0 to 3 iterations, but does not further increase with more refinement steps.

Sensitiveness to input masks. To show that our method is robust to the input mask of the object, we manually add dilations to the input mask by 0% to 10%, and then evaluate the results on the NAVI dataset. As we can see, adding 2% dilation to the masks almost does not change the performance at all, while further adding 10% dilation will decrease the performance reasonably. Note that with recent SAM [27] could produce very accurate object masks on arbitrary, so that the errors of mask are not very large in most cases.

4.6. Application in Visual Odometry

We show an application of our relative pose estimation in the visual odometry (VO) task. We use the ORB-SLAM2 [6] track a car to travel around a street and revisit a crossroad with a parking car beside the road as shown in Fig. 11. In this case, we use our algorithm to estimate the relative pose between two views that are co-visible to this car, and then we add the estimated relative pose to the pose graph to optimize it. It can be observed that with our rela-

Table 1. The quantitative comparison results on the two testing datasets.

Method	NAVI				GSO			
	Rotation Accuracy		Translation Accuracy		Rotation Accuracy		Translation Accuracy	
	15°	30°	15°	30°	15°	30°	15°	30°
SIFT [32]+ZoeDepth [4]+PnP	19.66	25.55	12.47	25.25	7.17	13.04	5.65	15.65
SIFT [32]+ZoeDepth [4]+Procrustes	16.65	26.72	9.60	24.31	5.00	14.78	3.48	14.57
Map-free-loc RPR [2]	17.23	33.99	13.81	35.92	4.57	15.87	3.91	16.52
LoFTR [66]	16.59	27.99	20.74	30.38	<u>20.65</u>	29.57	<u>29.57</u>	<u>36.52</u>
Relpose++ [31]	24.33	40.05	24.84	42.71	15.65	32.17	20.22	32.61
ID-Pose [9]	10.09	36.66	-	-	20.43	<u>40.43</u>	-	-
3DAHV [85]	<u>28.23</u>	<u>48.34</u>	-	-	16.74	38.91	-	-
Ours	43.16	66.47	50.64	72.41	40.43	57.61	42.39	60.65

Table 2. The quantitative comparison results on the two rotated testing datasets.

Method	Rotated NAVI				Rotated GSO			
	Rotation Accuracy		Translation Accuracy		Rotation Accuracy		Translation Accuracy	
	15°	30°	15°	30°	15°	30°	15°	30°
SIFT [32]+ZoeDepth [4]+PnP	18.09	23.90	13.68	22.08	7.83	14.35	5.65	17.61
SIFT [32]+ZoeDepth [4]+Procrustes	15.47	23.95	8.11	19.31	5.00	13.70	2.83	14.78
Map-free-loc RPR [2]	8.69	17.58	7.78	22.23	2.61	9.78	5.87	18.91
LoFTR [66]	<u>13.14</u>	21.28	<u>17.38</u>	28.97	<u>14.78</u>	18.70	<u>16.52</u>	<u>24.35</u>
3DAHV [85]	12.50	<u>30.72</u>	-	-	6.30	<u>18.71</u>	-	-
Relpose++ [31]	12.10	26.46	12.74	<u>29.20</u>	6.52	16.96	6.74	20.00
Ours	29.81	49.34	35.36	60.27	31.30	48.04	36.30	56.74

tive pose estimation, the tracked cameras are more accurate and more close to the ground-truth.

4.7. Runtime Analysis

All experiments are conducted on an Intel(R) Xeon(R) Gold 5220R CPU @ 2.20GHz and a single NVIDIA GeForce RTX 3090 GPU. For the configurations, we used in

the experiment, when the Zero123 model has already been loaded to GPU, building from a reference image costs eight minutes. After that, for each query image, it takes 0.18s to estimate the pose relative to the reference image. For a faster configuration with a rotation accuracy at 30-degree decrease of only 0.01 on rotated testing dataset, it takes 125s to build from a reference image, including 58s for inplane

Table 3. Ablation studies on inplane rotations and elevation predictions.

Method	Rotated NAVI				Rotated GSO			
	Rotation Accuracy		Translation Accuracy		Rotation Accuracy		Translation Accuracy	
	15°	30°	15°	30°	15°	30°	15°	30°
Without predict inplane rotation	22.26	41.60	26.89	51.80	19.13	32.17	19.57	41.09
Without predict elevation	22.34	44.33	27.34	55.87	18.91	36.09	21.74	50.87
Without both	20.57	40.12	26.41	52.85	11.52	29.35	18.48	37.83
With both	29.81	49.34	35.36	60.27	31.30	48.04	36.30	56.74

Table 4. Ablation studies on the number of generated images.

The number of generated images	NAVI				GSO			
	Rotation Accuracy		Translation Accuracy		Rotation Accuracy		Translation Accuracy	
	15°	30°	15°	30°	15°	30°	15°	30°
8	16.98	40.54	23.69	47.45	16.52	38.04	23.91	45.43
16	29.92	54.32	36.18	62.07	31.74	51.52	36.30	58.70
32	42.09	62.81	46.16	68.69	37.39	54.57	39.13	59.13
64	43.10	63.93	48.19	71.61	41.09	55.22	42.83	60.65
128	43.16	66.47	50.64	72.41	40.43	57.61	42.39	60.65

Table 5. Ablation studies on the number of refinement iterations on the GSO datasets.

#Refine	GSO			
	Rotation Accuracy		Translation Accuracy	
	15°	30°	15°	30°
0	26.96	52.39	31.74	57.17
1	36.30	55.00	38.04	60.00
2	40.22	56.74	42.39	60.22
3	40.43	57.61	42.39	60.87
4	41.09	57.83	42.39	60.65
5	41.52	57.17	42.39	61.30
6	40.65	57.39	43.26	61.74
7	40.87	56.96	43.04	61.30
8	40.43	57.17	43.70	61.74

Table 6. Ablation studies on sensitiveness to input masks on the NAVI dataset. We use different dilation kernels to dilate the input masks and evaluate the results on inaccurate masks.

Dilation	NAVI			
	Rotation Accuracy		Translation Accuracy	
	15°	30°	15°	30°
10%	36.38	56.43	36.18	64.31
5%	33.06	56.11	38.69	67.37
2%	42.23	61.70	44.61	72.65
0%	43.16	66.47	50.64	72.41

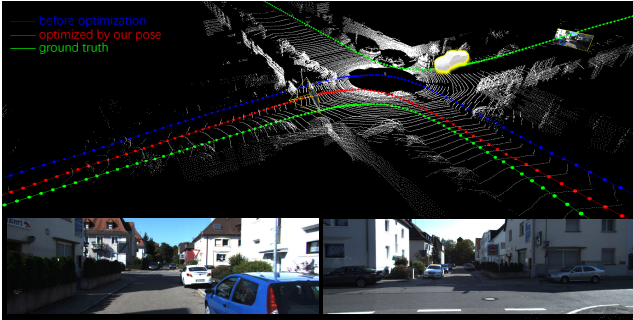


Figure 11. We add an edge to the pose graph based on estimated relative pose from reference image (left) to query image (right), and subsequently employ g2o to perform pose graph optimization.

rotation and elevation prediction, and 67s for novel-view image generations. If the object in reference image is correctly oriented, we can skip the inplane rotation prediction, then the building process only takes 73s under such a faster configuration.

4.8. Limitations

A failure example of our method on a symmetric object is shown in Fig. 12. However, precisely predicting the poses of symmetrical objects is really an ill-posed problem in the extreme two-views setting. Although Relpose++ [31] attempted to handle object symmetries, they rely on addi-

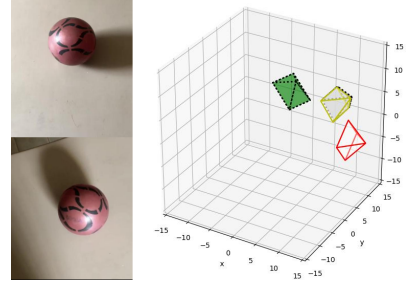


Figure 12. A failure example on a symmetric object.

tional views (more than 2 views) to avoid ambiguity. Rel-Pose++ [31] also fails in the symmetric example as shown in Fig. 12

5. Conclusion

In this paper, we introduced a new algorithm to estimate relative camera poses with extreme viewpoint changes. The key idea of our method is to utilize the object prior learned from large-scale 2D diffusion model Zero123 [35], which is able to generate novel-view images of an object. However, since Zero123 has a canonical coordinate system implicitly defined in its model and the image may not look at the object, we cannot directly apply Zero123. To address this challenge, we first propose a new formulation of the two-view pose estimation as an object pose estimation problem and correctly define the object poses for both input images and the generated images. Finally, we match the other image against the generated images to get an object pose estimation, which helps us determine the relative two-view camera poses. Extensive experiments on the GSO [14] and the NAVI [23] dataset have demonstrated the effectiveness of our design. We have also shown an application of our method in a Visual Odometry system.

Acknowledgments

This work is supported by HKU-SCF FinTech Academy, HKRGC Theme-based research scheme project T35-710/20-R, and SZ-HK-Macau Technology Research Programme #SGDX20210823103537030.

References

- [1] S. Abdulwahab, H. A. Rashwan, M. A. Garcia, M. Jabreel, S. Chambon, and D. Puig. Adversarial learning for depth and viewpoint estimation from a single image. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2947–2958, 2020. 4
- [2] E. Arnold, J. Wynn, S. Vicente, G. Garcia-Hernando, A. Monszpart, V. Prisacariu, D. Turmukhambetov, and E. Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *European Conference on Computer Vision*, pages 690–708. Springer, 2022. 4, 7, 8, 9, 10

- [3] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Computer Vision—ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7–13, 2006. Proceedings, Part I* 9, pages 404–417. Springer, 2006. 3
- [4] S. F. Bhat, R. Birkel, D. Wofk, P. Wonka, and M. Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 7, 10
- [5] D. Cai, J. Heikkilä, and E. Rahtu. Ove6d: Object view-point encoding for depth-based 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6803–6813, June 2022. 3
- [6] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós. Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021. 3, 9
- [7] P. Castro and T.-K. Kim. Posematcher: One-shot 6d object pose estimation by deep feature matching. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2140–2149, 2023. 3
- [8] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 4
- [9] W. Cheng, Y.-P. Cao, and Y. Shan. Id-pose: Sparse-view camera pose estimation by inverting diffusion models. *arXiv preprint arXiv:2306.17140*, 2023. 4, 7, 8, 9, 10
- [10] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker. Universal correspondence network. *Advances in neural information processing systems*, 29, 2016. 3
- [11] X. Cui, N. Li, C. Zhang, Q. Zhang, W. Feng, and L. Wan. Silhouette-based 6d object pose estimation. In F.-L. Zhang and A. Sharf, editors, *Computational Visual Media*, pages 157–179, Singapore, 2024. Springer Nature Singapore. 3
- [12] M. Dani, K. Narain, and R. Hebbalaguppe. 3dposelite: A compact 3d pose estimation using node embeddings. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1877–1886, 2021. 3
- [13] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 4
- [14] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 7, 8, 9, 11
- [15] G. Feng, T.-B. Xu, F. Liu, M. Liu, and Z. Wei. Nvr-net: Normal vector guided regression network for disentangled 6d pose estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(2):1098–1113, 2024. 3
- [16] D. Fu, S. Han, B. Liang, and W. Li. The 6d pose estimation of the aircraft using geometric property. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7):3358–3368, 2023. 4
- [17] A. Grabner, Y. Wang, P. Zhang, P. Guo, T. Xiao, P. Vajda, P. M. Roth, and V. Lepetit. Geometric correspondence fields: Learned differentiable rendering for 3d pose refinement in the wild, 2020. 3
- [18] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4
- [19] X. He, J. Sun, Y. Wang, D. Huang, H. Bao, and X. Zhou. Onepose++: Keypoint-free one-shot object pose estimation without CAD models. In *Advances in Neural Information Processing Systems*, 2022. 3
- [20] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4
- [21] M. Z. Irshad, T. Kollar, M. Laskey, K. Stone, and Z. Kira. Centersnap: Single-shot multi-object 3d shape reconstruction and categorical 6d pose and size estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 10632–10640. IEEE, 2022. 3
- [22] M. Z. Irshad, S. Zakharov, R. Ambrus, T. Kollar, Z. Kira, and A. Gaidon. Shapo: Implicit representations for multi-object shape, appearance, and pose optimization. In *European Conference on Computer Vision*, pages 275–292. Springer, 2022. 3
- [23] V. Jampani, K.-K. Maninis, A. Engelhardt, A. Karpur, K. Truong, K. Sargent, S. Popov, A. Araujo, R. Martin-Brualla, K. Patel, D. Vlasic, V. Ferrari, A. Makadia, C. Liu, Y. Li, and H. Zhou. NAVI: Category-agnostic image collections with high-quality 3d shape and pose annotations. In *NeurIPS*, 2023. 7, 8, 9, 11
- [24] H. Jiang, Z. Jiang, K. Grauman, and Y. Zhu. Few-view object reconstruction with unknown categories and camera poses. *arXiv preprint arXiv:2212.04492*, 2022. 4
- [25] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 4
- [26] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5614–5623, 2019. 3
- [27] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 9
- [28] M. Kocabas, N. Athanasiou, and M. J. Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 3
- [29] F. Li, S. R. Vutukur, H. Yu, I. Shugurov, B. Busam, S. Yang, and S. Ilic. Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2115–2125, 2023. 3
- [30] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox. DeepIM: Deep iterative matching for 6D pose estimation. In *European Conference Computer Vision (ECCV)*, 2018. 3

- [31] A. Lin, J. Y. Zhang, D. Ramanan, and S. Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. *arXiv preprint arXiv:2305.04926*, 2023. 1, 3, 7, 9, 10, 11
- [32] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2010. 3, 10
- [33] J. Liu, Z. Cao, Y. Tang, X. Liu, and M. Tan. Category-level 6d object pose estimation with structure encoder and reasoning attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6728–6740, 2022. 3
- [34] M. Liu, C. Xu, H. Jin, L. Chen, Z. Xu, H. Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 6
- [35] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 1, 2, 3, 4, 5, 8, 9, 11
- [36] Y. Liu, C. Lin, Z. Zeng, X. Long, L. Liu, T. Komura, and W. Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2, 4
- [37] Y. Liu, Y. Wen, S. Peng, C. Lin, X. Long, T. Komura, and W. Wang. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. In *European Conference on Computer Vision*, pages 298–315. Springer, 2022. 1, 3, 6
- [38] X. Long, Y.-C. Guo, C. Lin, Y. Liu, Z. Dou, L. Liu, Y. Ma, S.-H. Zhang, M. Habermann, C. Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023. 2, 4
- [39] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293(5828):133–135, 1981. 3
- [40] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 3, 7, 8, 9
- [41] W.-C. Ma, A. J. Yang, S. Wang, R. Urtasun, and A. Torralba. Virtual correspondence: Humans as a cue for extreme-view geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15924–15934, 2022. 3
- [42] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *Acm transactions on graphics (tog)*, 36(4):1–14, 2017. 4
- [43] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu. Relative camera pose estimation using convolutional neural networks. In *Advanced Concepts for Intelligent Vision Systems: 18th International Conference, ACIVS 2017, Antwerp, Belgium, September 18-21, 2017, Proceedings 18*, pages 675–687. Springer, 2017. 3
- [44] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1
- [45] C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan, and X. Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 4
- [46] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 3
- [47] R. Mur-Artal and J. D. Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5):1255–1262, 2017. 3
- [48] V. N. Nguyen, T. Groueix, Y. Hu, M. Salzmann, and V. Lepetit. Nope: Novel object pose estimation from a single image. *arXiv preprint arXiv:2303.13612*, 2023. 3
- [49] V. N. Nguyen, T. Groueix, M. Salzmann, and V. Lepetit. Gigapose: Fast and robust novel object pose estimation via one correspondence. *arXiv preprint arXiv:2311.14155*, 2023. 3
- [50] V. N. Nguyen, Y. Hu, Y. Xiao, M. Salzmann, and V. Lepetit. Templates for 3d object pose estimation revisited: Generalization to new objects and robustness to occlusions. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3
- [51] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence*, 26(6):756–770, 2004. 3
- [52] B. Okorn, Q. Gu, M. Hebert, and D. Held. Zephyr: Zero-shot pose hypothesis rating. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14141–14148. IEEE, 2021. 3
- [53] C. Papaioannidis and I. Pitas. 3d object pose estimation using multi-objective quaternion learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2683–2693, 2020. 3
- [54] K. Park, A. Mousavian, Y. Xiang, and D. Fox. Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3
- [55] G. Pitteri, S. Ilic, and V. Lepetit. Cornet: Generic 3d corners for 6d pose estimation of new objects without retraining. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2807–2815, 2019. 3
- [56] J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordone, P. Labatut, and D. Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 2, 3, 7
- [57] C. Rockwell, J. Johnson, and D. F. Fouhey. The 8-point algorithm as an inductive bias for relative pose prediction by vits. In *2022 International Conference on 3D Vision (3DV)*, pages 1–11. IEEE, 2022. 3
- [58] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4, 5

- [59] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 3, 8
- [60] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 3
- [61] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016. 3
- [62] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2
- [63] R. Shi, H. Chen, Z. Zhang, M. Liu, C. Xu, X. Wei, L. Chen, C. Zeng, and H. Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 2
- [64] I. Shugurov, F. Li, B. Busam, and S. Ilic. Osop: A multi-stage one shot object pose estimation framework. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6825–6834, 2022. 3
- [65] S. Sinha, J. Y. Zhang, A. Tagliasacchi, I. Gilitschenski, and D. B. Lindell. Sparsepose: Sparse-view camera pose regression and refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21349–21359, 2023. 1, 3
- [66] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 3, 7, 9, 10
- [67] J. Sun, Z. Wang, S. Zhang, X. He, H. Zhao, G. Zhang, and X. Zhou. Onepose: One-shot object pose estimation without cad models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6825–6834, 2022. 3
- [68] Z. Teed and J. Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 3
- [69] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):815–830, 2009. 3
- [70] B. Usman, A. Tagliasacchi, K. Saenko, and A. Sud. Metapose: Fast 3d pose from multiple views without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6759–6770, 2022. 3
- [71] P. Wang, H. Tan, S. Bi, Y. Xu, F. Luan, K. Sunkavalli, W. Wang, Z. Xu, and K. Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023. 3
- [72] S. Wang, R. Clark, H. Wen, and N. Trigoni. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 2043–2050. IEEE, 2017. 3
- [73] B. Wen and K. E. Bekris. Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021. 3
- [74] B. Wen, J. Tremblay, V. Blukis, S. Tyree, T. Müller, A. Evans, D. Fox, J. Kautz, and S. Birchfield. BundleSDF: Neural 6-DoF tracking and 3D reconstruction of unknown objects. In *CVPR*, 2023. 3
- [75] Y. Wen, X. Li, H. Pan, L. Yang, Z. Wang, T. Komura, and W. Wang. Disp6d: Disentangled implicit shape and pose learning for scalable 6d pose estimation. In *European Conference on Computer Vision*, pages 404–421. Springer, 2022. 3
- [76] Y. Xiao, X. Qiu, P. Langlois, M. Aubry, and R. Marlet. Pose from shape: Deep pose estimation for arbitrary 3D objects. In *British Machine Vision Conference (BMVC)*, 2019. 3
- [77] N. Yang, L. v. Stumberg, R. Wang, and D. Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1281–1292, 2020. 3
- [78] L. Yen-Chen, P. Florence, J. T. Barron, A. Rodriguez, P. Isola, and T.-Y. Lin. Inerf: Inverting neural radiance fields for pose estimation, 2021. 3
- [79] H. Yisheng, W. Yao, F. Haoqiang, C. Qifeng, and S. Jian. Fs6d: Few-shot 6d pose estimation of novel objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 3
- [80] J. Zhang, D. Sun, Z. Luo, A. Yao, L. Zhou, T. Shen, Y. Chen, L. Quan, and H. Liao. Learning two-view correspondences and geometry using order-aware network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5845–5854, 2019. 8
- [81] J. Y. Zhang, D. Ramanan, and S. Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *European Conference on Computer Vision*, pages 592–611. Springer, 2022. 1, 3, 9
- [82] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 4
- [83] C. Zhao, Y. Hu, and M. Salzmann. Fusing local similarities for retrieval-based 3d orientation estimation of unseen objects, 2022. 3
- [84] C. Zhao, Y. Hu, and M. Salzmann. Locposenet: Robust location prior for unseen object pose estimation, 2023. 3
- [85] C. Zhao, T. Zhang, and M. Salzmann. 3d-aware hypothesis & verification for generalizable relative object pose estimation. *Proceedings of the International Conference on Learning Representations*, 2024. 7, 8, 9, 10