UnPose: *Un*certainty-Guided Diffusion Priors for Zero-Shot *Pose* Estimation

Zhaodong Jiang[1,2*] **Ashish Sinha**[1] **Tongtong Cao**[1] **Yuan Ren**[1] **Bingbing Liu**[1] **Binbin Xu**[1†]
[1]Huawei Noah's Ark Lab, Canada [2]University of Toronto, Canada
{firstname.lastname}@huawei.com zhaodong.jiang@mail.utoronto.ca

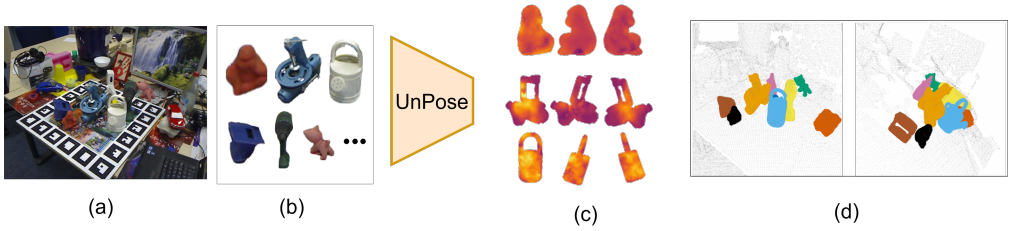


Figure 1: We present *UnPose* for zero-shot model-free pose estimation, that takes (a) RGB-D frames and (b) object segments in the scene, to output (c) 3D priors and uncertainty from a diffusion model to (d) reconstruct and continually refine the geometry of the object represented as 3DGS [1] for 6D pose estimation. (d) shows the resulting object poses with globally consistent front and back views from the input viewpoint.

Abstract: Estimating the 6D pose of novel objects is a fundamental yet challenging problem in robotics, often relying on access to object CAD models. However, acquiring such models can be costly and impractical. Recent approaches aim to bypass this requirement by leveraging strong priors from foundation models to reconstruct objects from single or multi-view images, but typically require additional training or produce hallucinated geometry. To this end, we propose *UnPose*, a novel framework for zero-shot, model-free 6D object pose estimation and reconstruction that exploits 3D priors and uncertainty estimates from a pre-trained diffusion model. Specifically, starting from a single-view RGB-D frame, *UnPose* uses a multi-view diffusion model to estimate an initial 3D model using 3D Gaussian Splatting (3DGS) representation, along with pixel-wise epistemic uncertainty estimates. As additional observations become available, we incrementally refine the 3DGS model by fusing new views guided by the diffusion model's uncertainty, thereby, continuously improving the pose estimation accuracy and 3D reconstruction quality. To ensure global consistency, the diffusion prior-generated views and subsequent observations are further integrated in a pose graph and jointly optimized into a coherent 3DGS field. Extensive experiments demonstrate that *UnPose* significantly outperforms existing approaches in both 6D pose estimation accuracy and 3D reconstruction quality. We further showcase its practical applicability in real-world robotic manipulation tasks. Video demos can be found at our project page: https://frankzhaodong.github.io/UnPose.

Keywords: 6D Pose Estimation, Diffusion Model, Object Reconstruction, Uncertainty Estimation

[*]Work done during an internship at Huawei Noah's Ark Lab
[†]Corresponding author

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

1 Introduction

6D object pose estimation is a fundamental challenge in robotic manipulation, enabling tasks from precision grasping to complex assembly [2, 3]. While deep learning (DL) has driven notable progress, prior approaches have predominantly relied on object-specific textured CAD models [4, 5, 6, 7] or category-level training data [8, 9, 10, 11, 12], severely limiting their application in dynamic, open-world environments. Even recent category-agnostic approaches [13, 14, 15] still depend on ground-truth CAD models during inference, essentially approximating the model-based paradigm rather than transcending it. As robots increasingly operate in open-world environments with novel objects, the need for methods that can estimate poses without prior object models has become increasingly urgent. Model-free pose estimation methods offer a promising alternative by eliminating dependence on pre-existing object models, instead leveraging reference views [14, 16, 17, 18], or image-to-3D reconstruction models [19, 20, 21, 22]. However, despite their "model-free" designation, these methods still implicitly rely on reference object models generated in advance from either diffusion models or multiple reference views of the target objects. Therefore, a key issue arises when 3D representations are generated from limited observations: unobserved regions introduce variable confidence *i.e.*, epistemic uncertainty during reconstruction. Existing approaches typically fail to quantify this epistemic uncertainty [23], treating all generated regions equally regardless of observability. This limitation becomes particularly problematic when integrating new observations, as the system cannot appropriately weigh the reliability of the prior model against newly observed data [24, 25].

To this end, we introduce *UnPose*, a principled framework for zero-shot, model-free 6D object pose estimation that addresses these limitations through uncertainty-aware multi-view integration and pose graph optimization. Starting from a single RGB-D observation, *UnPose* generates an initial 3D representation using an image-to-3D diffusion model with pixel-level uncertainty estimates that quantify confidence in the diffusion priors and balance them with sensor measurements. This 3D representation is encoded using 3D Gaussian Splatting (3DGS) [1], enabling efficient and high-fidelity textured models. As more observations arrive, *UnPose* incrementally refines the geometry and appearance via pose graph optimization, with uncertainty estimates guiding the fusion process to prioritize reliable measurements while refining uncertain regions.

In summary, our contributions are: (1) we present a model-free approach that eliminates dependencies on object-specific CAD models, category-level training, or multi-view requirements with known camera poses; (2) we introduce an uncertainty-guided refinement that adaptively integrates observed data into diffusion priors; (3) We formulate pose estimation as an incremental factor graph optimization problem that incorporates both diffusion priors and observations, and ensures global consistency; and (4) We demonstrate significant improvements in both pose estimation accuracy and reconstruction quality compared to existing SOTA methods.

2 Related Works

2.1 Model-based Object Pose Estimation

Instance-level pose estimation techniques [6, 4, 5, 7] typically rely on textured CAD models for specific objects, where both training and testing are conducted on the same set of instances. In contrast, category-level methods [9, 10, 8, 11, 12] aim to generalize to *unseen* instances within *known* categories, though they require costly category-specific annotations and are limited to predefined object classes. To address these limitations, category-agnostic approaches [13, 14, 15] aim to estimate the poses of novel objects without relying on predefined categories. However, many of these methods still depend on ground-truth CAD models to render reference views during inference, which restricts their practical use in open-world applications where such models are often unavailable or costly to obtain. More recent works [26, 27] attempt to address this gap by retrieving CAD models from large repositories, though retrieval accuracy and high quality datasets continue to pose challenges.

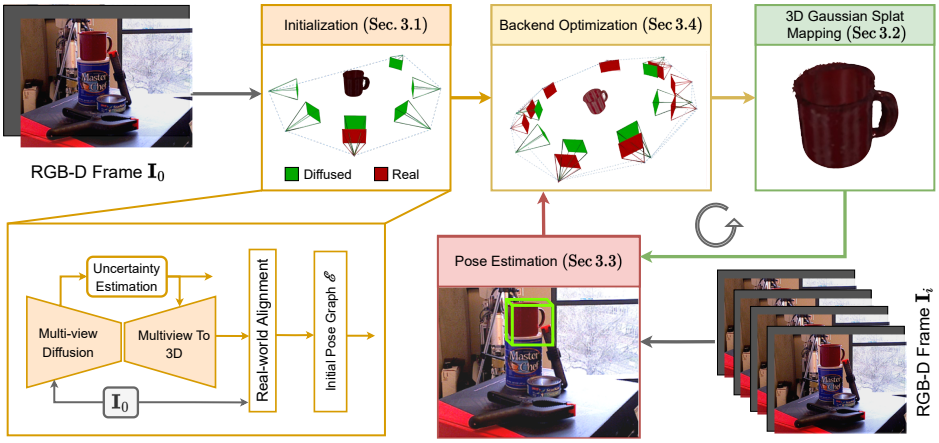


Figure 2: We present an overview of our proposed pipeline. Starting from a single-view RGB-D frame, *UnPose* uses a multi-view diffusion model to estimate an initial 3D model using 3D Gaussian Splatting (3DGS) representation, along with pixel-wise epistemic uncertainty estimates. As additional observations become available, we incrementally refine the 3DGS model by fusing new views guided by the diffusion model's uncertainty, thereby, continuously improving the pose estimation accuracy and 3D reconstruction quality.

2.2 Model-free Object Pose Estimation

To reduce reliance on textured CAD models, recent research has shifted toward *model-free* object pose estimation methods using either reference images [28, 14, 20] or single/multi-view images or videos [29, 27]. However, current SOTA approaches still rely on posed reference views [14] or reconstructed object models [30] and, thus, are not *strictly model free*. Several works attempt to relax this requirement by estimating relative orientation from a single anchor image [18, 20], using feature-matching [17] or language guidance [31]. Despite these advances, performance degrades when query views have limited overlap or severe occlusions. To tackle occlusion, recent works [21, 20, 22] leverage image-to-3D diffusion priors to reconstruct object's geometry from limited views and estimate the 6D pose, alleviating the need for object-specific CAD models. Yet, these methods do not refine the generated 3D models to correct diffusion-induced hallucinations [32], often producing inaccurate geometry and appearance for unobserved regions. Another direction draws on object-level SLAM [33, 34, 35], which incrementally refines object's geometry over time and handles occlusions. For example, GOM [25] combines diffusion priors with multi-view sensor data in an alternative optimization scheme. However, the alternative optimizations between diffusion and NeRF [36] yields suboptimal reconstruction quality. To address these challenges, we propose a *model-free* approach that combines diffusion priors and observations into an incrementally refined 3DGS [1] representation. We incorporate them into a factor graph [37] to ensure a globally consistent object representation that leads to robust and accurate object pose estimations.

2.3 Uncertainty Estimation

Uncertainty has been widely studied in robotics and computer vision communities [24, 23] to improve the robustness and accuracy under noisy (*aleatoric uncertainty*) or incomplete (*epistemic uncertainty*) observations [38, 39]. We focus on epistemic uncertainty due to occlusions, ambiguities and limited viewpoints. A common estimation approach is Deep Ensembles, which captures uncertainty via prediction variance across independently initialized models [40]. Alternatively, Bayesian Neural Networks [41] model a posterior distribution over weights, typically using variational inference with KL regularization to approximate the true posterior [42]. While effective, these methods are costly, which recent works circumvent through post-hoc Laplace Approximations (LA) [43]. However, applying Bayesian uncertainty to image-to-3D diffusion models (DMs) is challenging due to their high-dimensional parameter space and iterative reverse process, which complicates uncer-

Figure 3: Visualization of ground-truth rendering, and its corresponding diffused view and pixel-level uncertainties extracted from Wonder3D [46] using LLLA [48]. The diffused images show larger variance at unseen perspective.

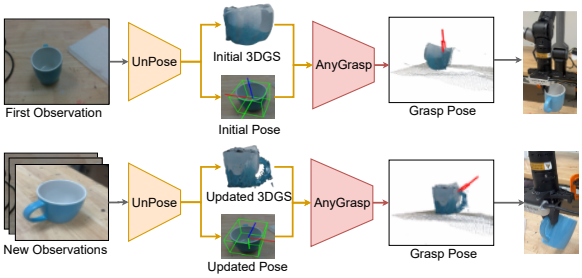


Figure 4: Illustration of *UnPose* for a real-world robotic manipulation task.

tainty propagation. Additionally, DMs are known to hallucinate [32], potentially compromising modeling accuracy. To address this, we model epistemic uncertainty in image-to-3D DMs using efficient Bayesian inference [44], and further refined through future observations.

3 Method

Our proposed pipeline, visualized in Fig. 2, consists of four modules: initialization (Sec. 3.1), pose estimation (Sec. 3.3), Backend Optimization (Sec. 3.4), 3DGS mapping (Sec. 3.2). For each object $\mathcal{O}$ in a scene, we continuously refine its 3D model, represented as a 3DGS field $\boldsymbol{Q}_o$ [1], in its own object canonical coordinate frame $\underrightarrow{\mathcal{F}}_O$, along with a relative pose, $\boldsymbol{T}_{OC}$, from the camera coordinate frame, $\underrightarrow{\mathcal{F}}_C$, to the object coordinate frame $\underrightarrow{\mathcal{F}}_O$. Camera poses, $\boldsymbol{T}_{WC} \in \mathrm{SE}(3)$, are known using an off-the-shelf SLAM method [45].

3.1 Initialization

Without relying on any CAD models or pre-reconstructed models, our system initializes from an arbitrary RGB-D frame $\boldsymbol{\mathcal{I}}_0 = \{\mathbf{I}_0 \in \mathbb{R}^{w\times h\times 3}, \mathbf{D}_0 \in \mathbb{R}^{w\times h}, \mathbf{M} \in \{0,1\}^{w\times h\times m}\}$ by synthesizing multiview diffusion images $\hat{\mathbf{I}}_{1:k}$ using the Wonder3D [46] where k is set to be 6. We assume that object masks $\mathbf{M}$ are provided to us beforehand for m objects in the scene using a SOTA instance segmentation model. We further estimate the corresponding pixel-wise uncertainty $\mathrm{Var}(\hat{\mathbf{I}})_{1:k} \in \mathbb{R}^{w\times h}$ for the diffused views and align them together with the initial frame $\boldsymbol{\mathcal{I}}_0$ following VGGT [47]. The aligned relative poses $\boldsymbol{T}_{C_0C_k}$ together with the estimated uncertainties are refined using pose graph optimization to reconstruct an initial 3DGS object field $\boldsymbol{Q}_o$ [1].

Uncertainty Estimation. To estimate uncertainty from a pre-trained diffusion model (DM) without additional training, we adapt the approach proposed in BayesDiff [44] into the Wonder3D framework [46]. Specifically, we adopt last-layer Laplace approximation (LLLA) [48] to approximate the predictive distribution of the DM's noise prediction for efficient Bayesian inference, enabling the DM to produce pixel-wise uncertainty alongside synthesized multi-view images.

Following LLLA [48], we approximate the distribution of the noise prediction ϵ_t of a pre-trained DM given the noisy state $\mathbf{x}_t$ at diffusion timestep t as a Gaussian distribution:

$$p(\epsilon_t \mid \mathbf{x}_t, t) \approx \mathcal{N}\left(\epsilon_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}\epsilon_t\right), \tag{1}$$

where θ represents the parameters of DM, $\epsilon_\theta(\mathbf{x}_t, t)$ is the standard prediction of the pre-trained model, and $\boldsymbol{\Sigma}\epsilon_t$ is the predictive covariance derived from the LLLA appplied to the model's last layer parameters (see [44, 48] for more details). For pixel-wise uncertainty, we are primarily interested in the diagonal elements of the covariance, $\mathrm{Var}\left(\epsilon_t\right) = \mathrm{diag}\left(\boldsymbol{\Sigma}\epsilon_t\right)$, visualized in Fig. 3.

Wonder3D [46] employs DDIM [49] for its deterministic sampling process. The standard DDIM update step to obtain $\mathbf{x}_{t-1}$ from $\mathbf{x}_t$ is given by

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\,\epsilon_t(\mathbf{x}_t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_t(\mathbf{x}_t), \tag{2}$$

where α_t and σ_t are noise schedule parameters.

To propagate uncertainty through the denoising steps, we iteratively update the variance $\mathrm{Var}(\mathbf{x}_{t-1})$ based on the variance from the previous step $\mathrm{Var}(\mathbf{x}_t)$ and the uncertainty from the noise prediction $\mathrm{Var}\,(\epsilon_t)$. The variance update corresponding to the DDIM step, Eq. (2), is:

$$\begin{aligned}\mathrm{Var}(\mathbf{x}_{t-1}) =& \frac{\alpha_{t-1}}{\alpha_t}\,\mathrm{Var}(\mathbf{x}_t) - 2\frac{\sqrt{\alpha_{t-1}}}{\sqrt{\alpha_t}}(\sqrt{1-\alpha_{t-1}-\sigma_t^2} - \frac{\sqrt{\alpha_{t-1}}}{\alpha_t}\sqrt{1-\alpha_t})\,\mathrm{Cov}\,(\mathbf{x}_t, \epsilon_t) \\ &+ (\sqrt{1-\alpha_{t-1}-\sigma_t^2} - \frac{\sqrt{\alpha_{t-1}}}{\alpha_t}\sqrt{1-\alpha_t})^2\,\mathrm{Var}\,(\epsilon_t) \end{aligned} \tag{3}$$

The key challenge is estimating the covariance term $\mathrm{Cov}\,(\mathbf{x}_t, \epsilon_t)$, which captures the correlation between the noisy state and the model's prediction. As this is analytically intractable, we estimate it using Monte Carlo sampling [50]:

$$\mathrm{Cov}\,(\mathbf{x}_t, \epsilon_\theta\,(\mathbf{x}_t, t)) \approx \frac{1}{S}\sum_{i=1}^{S}\big(\mathbf{x}_{t,i} \odot \epsilon_\theta(\mathbf{x}_{t,i}, t)\big) - \mathbb{E}[\mathbf{x}_t] \odot \frac{1}{S}\sum_{i=1}^{S}\epsilon_\theta(\mathbf{x}_{t,i}, t), \tag{4}$$

where $\odot$ denotes element-wise multiplication, S is the number of Monte Carlo samples, and $\mathbf{x}_{t,i}$ are samples drawn around the current state $\mathbf{x}_t$.

The entire uncertainty estimation pipeline operates during inference without any model retraining. The final $\mathrm{Var}(\mathbf{x}_0)$ provides the pixel-wise uncertainty estimates $\mathrm{Var}(\hat{\mathbf{I}})$ for the generated images $\hat{\mathbf{I}}$. The accuracy of the uncertainty estimates depends on the number of samples and diffusion steps. In practice, we use $S = 20$ samples per step across 50 denoising steps to balance the quality of the uncertainty estimates and computational cost.

Real-world Alignment. The DM-generated multi-view images $\hat{\mathbf{I}}_{1:k}$ typically lack the correct scale relative to the real-world observations $\boldsymbol{\mathcal{I}}$. To recover the true scale and align the views to the real-world measurements, we feed the diffusion images to the VGGT network [47], which outputs point-cloud $\hat{\mathbf{P}}_{1:k}$, associated confidence maps $\hat{\mathbf{C}}_{1:k}$, and the relative transformations $\boldsymbol{T}_{C_0 C_i}$ between input views. The predicted confidence maps are modulated by the estimated diffusion uncertainties $\mathrm{Var}(\hat{\mathbf{I}})$ through weighted exponential operation $\mathbf{C}_{1:k} = \exp\left(\frac{\hat{\mathbf{C}}_{1:k}}{\mathrm{Var}(\hat{\mathbf{I}}_{1:k})}\right)$.

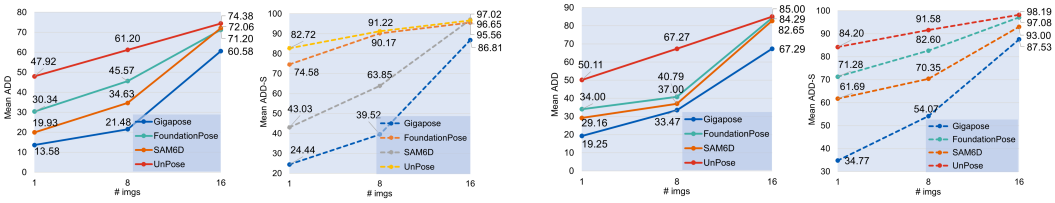
To align the diffusion pointcloud $\hat{\mathbf{P}}_{1:k}$ with the real-world pointcloud $\mathbf{P}_0$ extracted from $\boldsymbol{\mathcal{I}}_0$, we first match the eigenvalues of their covariance matrices via Principal Component Analysis (PCA). After recovering the scale s, we further refine the rigid transformation between them using Iterative Closest Point (ICP) [51]. The refined diffusion multiview frames together with the first real frame creates an initial pose graph, yielding a metrically consistent 3D reconstruction, aligned with the real-world coordinate frame $\underset{\rightarrow}{\mathcal{F}}_W$ while preserving uncertainty estimates.

3.2 3D Gaussian Splatting Mapping

With the poses optimized, we proceed to generate a compact 3DGS field [1] for each object. Inspired by SplaTAM [52], we model each 3D Gaussian as isotropic, encoding only RGB, 3D position, a scalar radius, and opacity. Importantly, we incorporate pixel-wise uncertainty to guide the updates of the 3DGS field when more observations are received. The overall mapping loss is defined as:

$$L_t = \sum_{\mathbf{p}} \mathbf{C}(\mathbf{p})\,(\mathbf{S}(\mathbf{p}) > 0.99)\,(\mathcal{L}_1(D(\mathbf{p})) + 0.5\,\mathcal{L}_1(C(\mathbf{p})))\,, \tag{5}$$

where $\mathbf{C}(\mathbf{p})$ denotes the uncertainty at pixel $\mathbf{p}$, $\mathbf{S}(\mathbf{p})$ represents the visibility score, and $D(\mathbf{p})$, $C(\mathbf{p})$ are the rendered depth and color residual, respectively. By incorporating certainties from diffusion priors, our mapping is able to always maintain a complete shape while continuously improving he geometric accuracy and photometric consistency when more information are observed.



(a) Performance on YCB-Video [53] dataset.

(b) Performance on LM-O [54] dataset.

Figure 5: We show quantitative comparison on YCB-V [53] and LM-O [54] using ADD and ADD-S metric w.r.t the number of reference-views. *UnPose* significantly outperforms the baselines. Note the consistent improvement in performance of our method as the number of reference views increases.

3.3 6D Pose Estimation

Once we have the optimized 3DGS representation of the object, we employ the Pose Refinement network from FoundationPose [30] for 6D pose estimation due to its SOTA performance on speed and accuracy. The network takes two inputs: (1) a rendered RGB-D view of the object from the current pose estimate (2) a cropped RGB-D observation from the camera. In our implementation, we modify the first input to be rendered directly from the 3DGS field [1], rather than a pre-reconstructed 3D model, improving applicability in open-world deployments. Both inputs are processed through a Siamese encoder with shared weights to extract feature maps, which are concatenated and tokenized into patches with positional embeddings. The transformer iteratively refines the pose estimates, minimizing the discrepancy between rendered and observed views in feature space.

3.4 Backend Optimization

After estimating the relative transformation $\boldsymbol{T}_{OC_i}$ for a new frame $\mathcal{F}_{C_i}$, we evaluate its correspondence to the closest existing keyframe, including both real and virtual diffusion frames, using Mast3R [55]. If the number of matching inliers is below a threshold, the frame is added as a new keyframe $\mathcal{K}_i$ and a bidirectional edge is created between $\mathcal{K}_i$ and its closest keyframe $\mathcal{K}_{i-1}$, updating the pose graph edge set $\mathcal{E}$. To close loops for local and global pose graphs, we encode keyframe features using Aggregated Selective Match Kernels (ASMK) [56, 57].

Geometric Pose Graph Optimization. Once loop closures are identified, we optimize the geometric pose graph by minimizing a weighted geometric residual over keyframe correspondences:

$$E_g = \sum_{i,j \in \mathcal{E}} \sum_{m,n \in \mathbf{m}_{i,j}} \left\| \frac{\tilde{\mathbf{X}}^i_{i,m} - \mathbf{T}_{ij} \tilde{\mathbf{X}}^j_{j,n}}{w(\mathbf{q}_{m,n}, \mathbf{C}_{m,n}, \sigma_g^2)} \right\|_\rho, \; w(\mathbf{q}, \mathbf{C}, \sigma^2) = \begin{cases} \sigma^2/\mathbf{q} & \text{otherwise} \\ \infty & \mathbf{q} < \mathbf{q}_{min} \| \mathbf{C} < c_{min} \end{cases} \tag{6}$$

where $\mathbf{m}$ denotes set of matches between keyframes $\mathcal{K}_i$ and $\mathcal{K}_j$ predicted by Mast3R [55], and $\tilde{\mathbf{X}}^i_{i,m}, \tilde{\mathbf{X}}^j_{j,n}$ are the corresponding 3D points. $\mathbf{q}_{m,n}$ is the match confidence, $\|\|_\rho$ is Huber loss and $w(\mathbf{q}, \mathbf{C}, \sigma^2)$ is per-match weighting [57] which we extend to include diffusion uncertainties. This nonlinear least squares problem is solved efficiently via Gauss-Newton optimization with sparse Cholesky decomposition in CUDA. To avoid ambiguity between virtual and real frames, we fix the poses of all diffusion-rendered frames. The refined poses of real keyframes are then sent back to the 3DGS mapping module (see Sec. 3.2) to ensure a multi-view consistent object representation.

Relocalization. If object tracking fails due to occlusion, fast motion or other disruptions, relocalization is triggered. Failure is detected through two signals: (a) a sharp drop in frame-to-keyframe correspondence, and (b) a large geometric discrepancy between the observed point cloud and the expected model rendered from the current pose estimate. Our system benefits from the completeness of the 3D model, allowing even unobserved parts to serve as weak priors. Upon failure detection, the relocalization module retrieves top-k candidate frames from the retrieval database using ASMK features and ranks them based on Euclidean distance. The selected frames are jointly optimized with the current frame to recover its pose, allowing tracking to resume reliably.

4 Experiments

We evaluate our proposed method, *UnPose*, on two benchmarks: YCB-Video [53] and LM-O [54]. For pose estimation, we compare against three SOTA baselines: GigaPose [21] (model-based), SAM6D [14] (zero-shot reference view-based) and FoundationPose [30] (model-free reconstruction-based). As mentioned in Sec. 2, these methods require estimation of the complete geometry prior to estimating the 6D pose. To quantitatively evaluate how the view number of observations impact the pose estimation results, we report the performance given 1, 8, and 16 reference images surrounding the target object, representing pose estimation performance at single view, partial observations, and complete observations cases. For reconstruction quality, we compare *UnPose* against three SOTA baselines: BundleSDF [29] (neural implicit reconstruction), GOM [25] (3D diffusion priors-based multi-view optimization), and Wonder3D [46] (single-view diffusion-based 3D generation), on several objects from YCB-Video [53] dataset. Similarly, we test the reconstruction performance under 1, 8, 16 reference views. Pose estimation accuracy is evaluated using Area Under the Curve (AUC) of Average Distance of Model Points (ADD) and its symmetric variant (ADD-S). Reconstruction quality is measured by Chamfer Distance (CD) to ground truth. Furthermore, we report the computation time of *UnPose*.

4.1 Results

Pose Estimation. As visible in Figs. 5a and 5b, our method consistently outperforms all baselines in both sparse and dense observation settings, with performance continuously improving as the number of reference views increase. In the single view case, *UnPose* surpasses the model-based GigaPose [21] by 71.66%, and the model-free FoundationPose (also the second-best method) [30] by 36.6% on average, benefiting from strong 3D diffusion priors in the initial pose graph. With additional views, our geometric pose graph optimization enforces multi-view consistency and leverages uncertainty estimates to refine unreliable regions, further enhancing performance. For the sake of page limit, we show per-object quantitative and qualitative comparisons in the appendix.

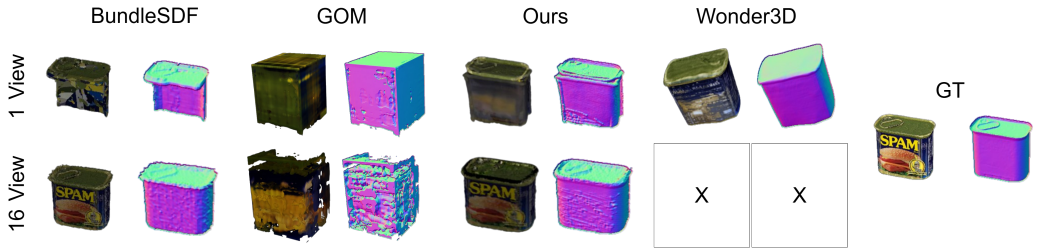
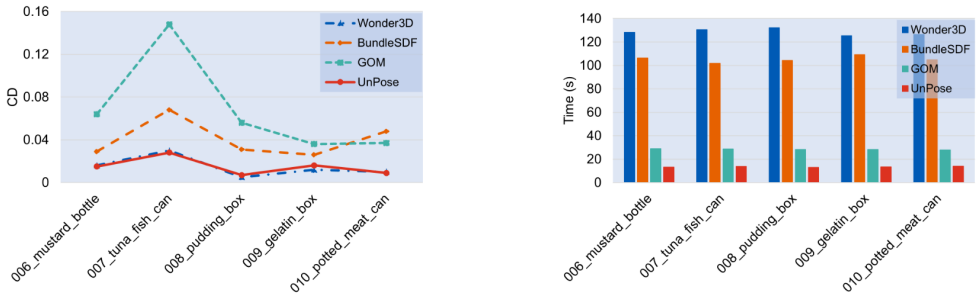


Figure 6: We show qualitative results using our uncertainty-aware 3D reconstruction when 1 and 16 reference views are used for reconstruction. Notice the improvement in geometry and appearance of the object with our method as the number of reference views increases. Since Wonder3D [46] does not support multi-view reconstruction, we do not show it.

Reconstruction. Fig. 7 shows the comparison of quantitative results for single-view object reconstruction against baselines [29, 25, 46] on a subset of objects from YCB-Video [53] dataset. As visible in Figs. 6 and 7a, *UnPose* significantly outperforms the baselines. Compared to the multiview-to-3D diffusion model [46, 58], *UnPose* is significantly faster while yielding comparable reconstruction. In comparison to other optimization-based methods [29, 25], *UnPose* is $7\times$ and $2\times$ faster while being $2.3\times$ and $3.8\times$ accurate on average, respectively. These performance gains are owing to the proposed uncertainty-guided reconstruction using 3D diffusion priors and global consistency brought by pose graph optimization.

Runtime Analysis. For a fair comparison, all quantitative comparisons were performed on the same hardware platform. Further, we report the time for reconstruction compared to prior works



(a) Chamfer Distance

(b) Reconstruction Time (s)

Figure 7: Quantitative comparison of object reconstruction for fidelity (a) and efficiency (b) using a single reference view, on objects from YCB-Video [53] dataset.

in Fig. 7b, showcasing the efficiency and efficacy of our uncertainty-guided diffusion priors. We also provide a comprehensive analysis of the runtime of each component of our framework in Tab. 2.

	16 images		8 images	
	PSNR (↑)	CD (↑)	PSNR (↑)	CD (↑)
Ours	**35.70**	**0.023**	**32.90**	**0.032**
↪ w/o Uncertainty Estimation	29.80	0.027	29.3	0.034
↪ w/o Bundle Adjustment	26.11	0.035	30.1	0.038
↪ w/o Diffused frame in the pose graph	35.5	0.024	28.5	0.046

Table 1: Ablation study demonstrating the impact of uncertainty estimation and incremental refinement on appearance of real views (PSNR) and reconstructed geometry (CD). ↑ denotes higher is better.

Process	Time (s)
Uncertainty Estimation	12.08
Remaining Initialization (Sec. 3.1)	1.07
3DGS Mapping (Sec. 3.2)	0.22
Pose Estimation (Sec. 3.3)	0.12
Backend Optimization (Sec. 3.4)	1.70
Total	13.49

Table 2: Runtime of each component of our pipeline.

4.2 Ablation Study

We tested our method performance without uncertainty information, without bundle adjustment, and without including diffused frames in pose graph optimization in two scenarios: 16 images and 8 images, using PSNR and CD metrics. With 16 images, uncertainty guidance and bundle adjustment improved the performance by 20%, while incorporating diffused frames had minimum effect when the density of observations is sufficient. In contrast, with only 8 images, incorporating diffused frames into the pose graph yielded a 15% performance gain, highlighting their importance in sparser observation scenarios.

4.3 Robotic Application

We further demonstrate the effectiveness of *UnPose* in robot arm manipulation tasks, where it is used to estimate an object's 6D pose and completed geometry, followed by grasp pose estimation using AnyGrasp [59]. An example can be seen in Fig. 4, the in-hand camera initially captures the mug with its handle occluded, leading to a grasp near the cup's wall. This is not ideal for tasks such as water pouring. As new views reveal the handle, *UnPose* incrementally updates the geometry, resulting in a new handle-aligned grasp pose, thereby allowing the robot arm to successfully grasp the mug by its handle. More details are provided in Appendix C.

5 Conclusion

We presented *UnPose*, a novel framework for model-free 6D pose estimation that exploits 3D priors and uncertainty estimates from a diffusion model. Our method incrementally refines object geometry by fusing diffused views with their uncertainty estimates and incorporating them into a factor graph. We demonstrated qualitatively and quantitatively the advantages of our proposed method for pose estimation and reconstruction tasks.

6 Limitations

Our current backend optimization relies on correspondence computed from Mast3R[55], which limits performance on textureless objects. A promising direction is to integrate direct matching losses from the direct SLAM literature [60]. Additionally, despite the significant performance gains, real-time deployment remains a challenge due to the computational overhead of Monte Carlo sampling for diffusion-based uncertainty estimation. This could be addressed by training a diffusion model that jointly predicts both mean and uncertainty. In addition, our current focus is on individual objects, incorporating scene-level priors and cross-object relationships may further improve global consistency. We believe these limitation analysis and future directions offer valuable opportunities for the community.

Acknowledgments

The authors would like to thanks the reviewers for their constructive feedback on improving this work. We would also like to the thank the members of Huawei Noah's Ark Lab for insightful discussions during the development of this project.

References

[1] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 2, 3, 4, 5, 6

[2] J. Zhou, Y.-S. Liu, and Z. Han. Zero-shot scene reconstruction from single images with deep prior assembly. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[3] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox. Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. *IEEE International Conference on Robotics and Automation (ICRA)*, 2021. 2

[4] K. Park, T. Patten, and M. Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7668–7677, 2019. 2

[5] Y. Labbé, J. Carpentier, M. Aubry, and J. 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. 2

[6] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11632–11641, 2020. 2

[7] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3003–3013, 2021. 2

[8] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2642–2651, 2019. 2

[9] J. Zhang, W. Huang, B. Peng, M. Wu, F. Hu, Z. Chen, B. Zhao, and H. Dong. Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In *European Conference on Computer Vision*, pages 199–216. Springer, 2024. 2

[10] D. Chen, J. Li, Z. Wang, and K. Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11973–11982, 2020. 2

[11] X. Lin, W. Yang, Y. Gao, and T. Zhang. Instance-adaptive and geometric-aware keypoint learning for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21040–21049, 2024. 2

[12] Y. Chen, Y. Di, G. Zhai, F. Manhardt, C. Zhang, R. Zhang, F. Tombari, N. Navab, and B. 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. 2

[13] A. Caraffa, D. Boscaini, A. Hamza, and F. Poiesi. Freeze: Training-free zero-shot 6d pose estimation with geometric and vision foundation models. In *European Conference on Computer Vision*, pages 414–431. Springer, 2024. 2

[14] J. Lin, L. Liu, D. Lu, and K. Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27906–27916, 2024. 2, 3, 7, 13, 14, 15, 16

[15] J. Chen, Z. Zhou, M. Sun, R. Zhao, L. Wu, T. Bao, and Z. He. Zeropose: Cad-prompted zero-shot object 6d pose estimation in cluttered scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[16] 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. 2

[17] 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. 2, 3

[18] V. N. Nguyen, T. Groueix, G. Ponimatkin, Y. Hu, R. Marlet, M. Salzmann, and V. Lepetit. Nope: Novel object pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17923–17932, 2024. 2, 3

[19] 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. 2

[20] T. Lee, B. Wen, M. Kang, G. Kang, I. S. Kweon, and K.-J. Yoon. Any6d: Model-free 6d pose estimation of novel objects. *CVPR*, 2025. 2, 3, 14

[21] V. N. Nguyen, T. Groueix, M. Salzmann, and V. 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*, pages 9903–9913, 2024. 2, 3, 7, 13, 14, 15, 16

[22] Y. Liu, Z. Jiang, B. Xu, G. Wu, Y. Ren, T. Cao, B. Liu, R. H. Yang, A. Rasouli, and J. Shan. Hippo: Harnessing image-to-3d priors for model-free zero-shot 6d pose estimation. *IEEE Robotics and Automation Letters*, 10(8):8284–8291, 2025. 2, 3, 14

[23] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 2, 3

[24] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. Cambridge: MIT Press, 2005. 2, 3

[25] Z. Liao, B. Xu, and S. L. Waslander. Toward general object-level mapping from sparse views with 3d diffusion priors. In *8th Annual Conference on Robot Learning*, 2024. 2, 3, 7, 14

[26] D. Cai, J. Heikkilä, and E. Rahtu. Gs-pose: Cascaded framework for generalizable segmentation-based 6d object pose estimation. *arXiv e-prints*, pages arXiv–2403, 2024. 2

[27] G. Ponimatkin, M. Cífka, T. Souček, M. Fourmy, Y. Labbé, V. Petrik, and J. Sivic. 6d object pose tracking in internet videos for robotic manipulation. *arXiv preprint arXiv:2503.10307*, 2025. 2, 3

[28] X. Liu, G. Wang, R. Zhang, C. Zhang, F. Tombari, and X. Ji. Unopose: Unseen object pose estimation with an unposed rgb-d reference image. *arXiv preprint arXiv:2411.16106*, 2024. 3

[29] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 606–617, 2023. 3, 7, 14

[30] B. Wen, W. Yang, J. Kautz, and S. Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024. 3, 6, 7, 13, 14, 15, 16

[31] J. Corsetti, D. Boscaini, C. Oh, A. Cavallaro, and F. Poiesi. Open-vocabulary object 6d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18071–18080, 2024. 3

[32] S. K. Aithal, P. Maini, Z. Lipton, and J. Z. Kolter. Understanding hallucinations in diffusion models through mode interpolation. *Advances in Neural Information Processing Systems*, 37:134614–134644, 2024. 3, 4

[33] J. McCormac, R. Clark, M. Bloesch, A. Davison, and S. Leutenegger. Fusion++: Volumetric object-level slam. In *2018 international conference on 3D vision (3DV)*, pages 32–41, 2018. 3

[34] B. Xu, W. Li, D. Tzoumanikas, M. Bloesch, A. Davison, and S. Leutenegger. Mid-fusion: Octree-based object-level multi-instance dynamic slam. In *2019 International Conference on Robotics and Automation (ICRA)*, 2019. 3

[35] L. Nicholson, M. Milford, and N. Sünderhauf. Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam. *IEEE Robotics and Automation Letters*, 4(1):1–8, 2018. 3

[36] 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. 3

[37] F. Dellaert, M. Kaess, et al. Factor graphs for robot perception. *Foundations and Trends® in Robotics*, 6 (1-2):1–139, 2017. 3

[38] Z. Liao, J. Yang, J. Qian, A. P. Schoellig, and S. L. Waslander. Uncertainty-aware 3d object-level mapping with deep shape priors. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024. 3

[39] Z. Tan, X. Chen, J. Zhang, L. Feng, and D. Hu. Uncertainty-aware normal-guided gaussian splatting for surface reconstruction from sparse image sequences. *arXiv preprint arXiv:2503.11172*, 2025. 3

[40] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 3

[41] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022. 3

[42] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015. 3

[43] H. Ritter, A. Botev, and D. Barber. A scalable laplace approximation for neural networks. In *6th international conference on learning representations, ICLR 2018-conference track proceedings*, volume 6. International Conference on Representation Learning, 2018. 3

[44] S. Kou, L. Gan, D. Wang, C. Li, and Z. Deng. Bayesdiff: Estimating pixel-wise uncertainty in diffusion via bayesian inference. *ICLR*, 2024. 4

[45] 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. 4

[46] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 4, 7, 13, 14

[47] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny. Vggt: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025. 4, 5

[48] A. Kristiadi, M. Hein, and P. Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pages 5436–5446. PMLR, 2020. 4

[49] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *ICLR*, 2021. 4

[50] W. K. Hastings. *Monte Carlo sampling methods using Markov chains and their applications*. Oxford University Press, 1970. 5

[51] J. Zhang, Y. Yao, and B. Deng. Fast and robust iterative closest point. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3450–3466, 2021. 5

[52] N. Keetha, J. Karhade, K. M. Jatavallabhula, G. Yang, S. Scherer, D. Ramanan, and J. Luiten. Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21357–21366, 2024. 5

[53] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 6, 7, 8, 13, 14, 16

[54] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother. Learning 6d object pose estimation using 3d object coordinates. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 536–551. Springer, 2014. 6, 7, 13, 14, 15

[55] V. Leroy, Y. Cabon, and J. Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91, 2024. 6, 9

[56] G. Tolias, Y. Avrithis, and H. Jégou. To aggregate or not to aggregate: Selective match kernels for image search. In *IEEE International Conference on Computer Vision*, 2013. 6

[57] R. Murai, E. Dexheimer, and A. J. Davison. MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors. *CVPR*, 2025. 6

[58] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 7

[59] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics (T-RO)*, 2023. 8, 14

[60] T. Whelan, R. F. Salas-Moreno, B. Glocker, A. J. Davison, and S. Leutenegger. Elasticfusion: Real-time dense slam and light source estimation. *The International Journal of Robotics Research*, 35(14): 1697–1716, 2016. 9

[61] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 13

[62] T. Hodan, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *WACV*, 2017. 13, 14

[63] T. Hodan, F. Michel, E. Brachmann, W. Kehl, A. GlentBuch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *ECCV*, pages 19–34, 2018. 13, 14

A Additional Qualitative Results

Scene-level Reconstructions. To demonstrate the generalization capablities of *UnPose*, we test its performance beyond tabletop objects (YCB-V, LM-O) on large-scale furniture objects from the ScanNet [61]. As shown in Fig. 8, *UnPose* reconstruct multiple diverse objects in a scene, including four chairs, one sofa, and one table, all from the same pipeline. We showcase the reconstruction of one object (chair) in the scene, and show it from an input camera viewpoint (Front View) and two novel views (Back View 1 and 2). The smaller floater images are real observations and diffusion frames provided to backend optimization, offering context on the input data.

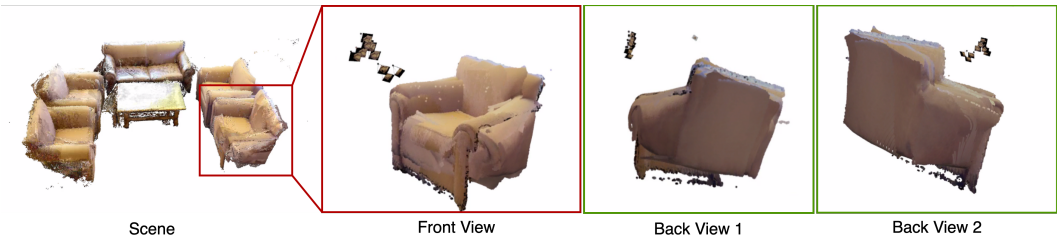


Figure 8: Visualization of a scene-level reconstruction on a ScanNet scene [61] by *UnPose*. We show the reconstruction of one of the chairs from one input camera viewpoint (Front View) as well as diffused novel views (Back View 1 and 2). Real and diffusion images of the scene used in backend optimization are shown as floaters.

Fig. 9 visualizes the pixel-level uncertainty estimations associated with the novel views generated by the diffusion model [46] used in our pipeline. For a better understanding, we also show the ground truth renderings of the object alongside each diffused view. It can be seen that the diffusion model exhibits higher variance (greater uncertainties) for novel views that significantly deviates from the input. This implies the model is less confident when synthesizing unseen regions of the object.

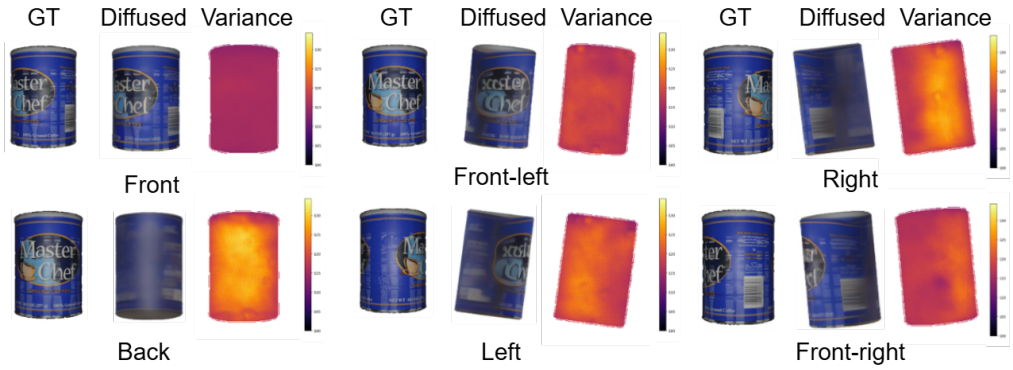


Figure 9: Visualization of diffused views, the associated pixel-level uncertainty estimates, and the corresponding ground-truth rendering perspectives. The diffused images show larger variance at unseen angles.

We further provide qualitative comparisons with SOTA 6D pose estimation methods [21, 14, 30], including GigaPose [21], SAM-6D [14], and FoundationPose [30] on standard 6D pose estimation benchmark datasets: YCB-Video [53] and LM-O [54]. Fig. 10 clearly shows that compared to prior methods where the estimated translation and rotation show high error, our proposed method, *UnPose*, accurately estimates the 6DOF of the object.

B Additional Quantitative Results

In this section, we present additional quantitative comparison details on YCB-Video [53] and LM-O [54] datasets, and further experiments on T-Less [62] and TYO-L [63]. We report object reconstruction results on YCB-V [53] dataset in Tab. 3, evaluating performance with 1, 8 and 16 images. Our proposed method consistently achieves lower reconstruction error (measured as Chamfer Distance) and is significantly faster than prior

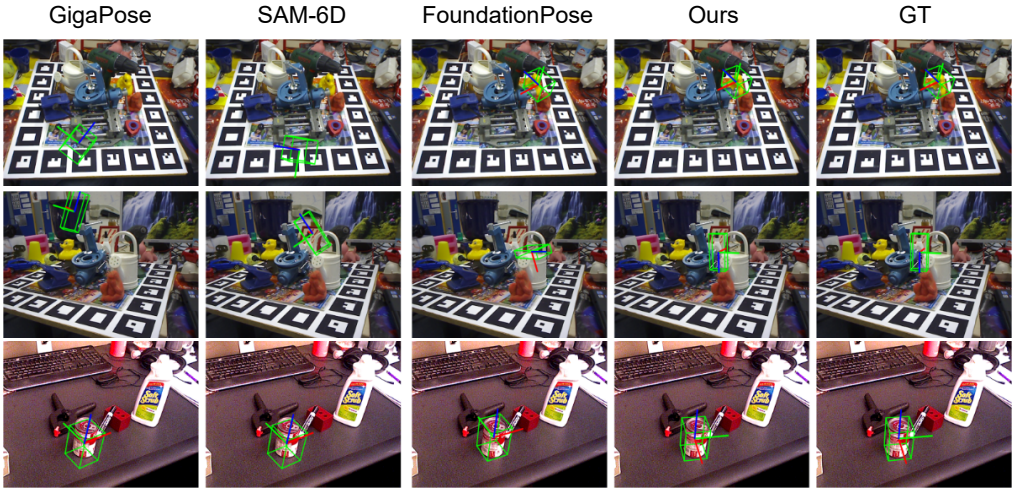


Figure 10: Qualitative results of *UnPose* compared to prior pose estimation methods on YCB-Video [53] and LM-O [54] datasets when 8 reference views are used.

methods [25, 29, 46]. We tabulate the 6D pose estimation results on YCB-Video [53], LM-O [54], T-Less [62], and TYO-L [63] in Tabs. 4 to 6 respectively. Following prior works [30, 22, 20], we evaluate the performance on pose estimation using ADD and ADD-S. Furthermore, for each method we report the results when 1, 8 and 16 images are used for reconstruction [21, 30] or as reference views [14]. It can be seen in Tabs. 4 to 6 that our method consistently outperforms prior SOTA methods (visible as gradual increase in the brightness of the gradient) in all settings *i.e.*, both metric and the number of images.

Method	Metric	$ycbv_5$			$ycbv_6$			$ycbv_7$			$ycbv_8$			$ycbv_9$			$ycbv_{10}$		
		1	8	16	1	8	16	1	8	16	1	8	16	1	8	16	1	8	16
Wonder3d [46]	CD	0.016	0.00	0.00	0.030	0.00	0.00	0.005	0.00	0.00	0.012	0.00	0.00	0.010	0.00	0.00	0.021	0.00	0.00
	Time (s)	128.20	0.00	0.00	130.63	0.00	0.00	132.22	0.00	0.00	125.30	0.00	0.00	126.40	0.00	0.00	130.20	0.00	0.00
BundleSDF [29]	CD	0.029	0.020	0.016	0.068	0.044	0.015	0.031	0.017	0.011	0.026	0.018	0.011	0.048	0.024	0.014	0.005	0.004	0.001
	Time (s)	106.35	135.27	172.96	101.74	112.90	126.31	104.23	118.74	131.49	109.20	114.46	130.76	104.97	129.13	147.39	102.11	115.84	132.76
GOM [25]	CD	0.064	0.117	0.080	0.148	0.121	0.100	0.056	0.066	0.079	0.036	0.169	0.092	0.037	0.039	0.046	0.153	0.268	0.170
	Time (s)	28.88	53.64	121.19	28.63	54.92	128.65	28.38	53.53	120.24	28.39	52.47	116.98	27.97	53.11	115.53	28.27	52.58	117.18
Ours	CD	0.015	0.005	0.004	0.028	0.023	0.020	0.007	0.006	0.005	0.016	0.008	0.007	0.009	0.009	0.004	0.021	0.005	0.003
	Time (s)	13.20	26.20	41.20	13.80	27.20	41.70	12.90	26.30	43.22	13.50	25.30	40.25	14.10	22.21	41.30	11.30	23.20	42.10

Table 3: Quantitative comparison for reconstruction with $1, 8$ and 16 images using *UnPose* in terms of accuracy (reported using chamfer distance (CD) as 10^{-3}) and time (sec) on objects from YCB-V [53]. For both CD and time, brighter gradients denotes better performance.

(a) T-LESS [62]

# imgs	Metric	GigaPose [21]	SAM6D [14]	FoundationPose [30]	Ours
1	ADD	10.0	14.9	<u>16.6</u>	**32.5**
	ADD-S	71.3	78.5	<u>82.2</u>	**89.2**
8	ADD	17.2	28.9	<u>33.2</u>	**40.1**
	ADD-S	73.0	81.8	<u>85.4</u>	**93.6**
16	ADD	44.0	50.1	<u>53.5</u>	**54.2**
	ADD-S	76.7	89.5	<u>90.7</u>	**94.7**

(b) TYO-L [63]

# imgs	Metric	GigaPose [21]	SAM6D [14]	FoundationPose [30]	Ours
1	ADD	14.2	37.5	<u>42.2</u>	**55.3**
	ADD-S	62.2	75.8	<u>82.7</u>	**88.2**
8	ADD	23.3	42.7	<u>45.5</u>	**67.2**
	ADD-S	72.2	82.6	<u>86.2</u>	**93.6**
16	ADD	59.2	68.3	<u>72.2</u>	**75.3**
	ADD-S	87.3	93.3	<u>97.3</u>	**98.1**

Table 4: Quantitative comparison of pose estimation on T-LESS [62] and TYO-L [63]. Best in **bold**; second best is <u>underlined</u>.

C Deployments on a Real-world Robotic Platform

We further demonstrate the effectiveness and the deployment capablity of *UnPose* in a real-world robotic manipulation pipeline using a PiPER robot arm. In this setup, *UnPose* estimates the complete 3D goemetry and 6D pose of a target object from RGB-D stream inputs captured by a wrist-mounted Intel RealSense D435 camera. This geometric and pose information is then passed to a subsequent grasp planning module, AnyGrasp [59], to determine a suitable grasp pose for the robot. Fig. 11 provides a visual snapshot of this system in action, showcasing *UnPose*'s reconstruction and pose estimation for a target mug on PiPER. Further details, includ-

Method	# imgs	Metric	Mean	bath duck	cat toy	hole puncher	power drill	water can
Gigapose [21]	1	ADD	19.25	1.07	1.15	2.53	48.65	42.86
		ADD-S	34.77	6.82	6.90	8.63	83.78	67.72
	8	ADD	33.46	15.56	4.87	10.84	66.67	69.39
		ADD-S	54.07	33.37	12.43	50.21	90.90	83.45
	16	ADD	67.29	63.82	57.47	29.36	89.47	96.36
		ADD-S	87.53	97.28	97.70	47.37	97.37	97.96
FoundationPose [30]	1	ADD	34.00	8.70	26.12	22.00	82.23	30.95
		ADD-S	71.28	34.78	78.10	70.10	90.13	83.33
	8	ADD	40.79	19.57	22.78	18.26	88.23	55.14
		ADD-S	82.60	78.27	86.83	66.26	96.33	85.33
	16	ADD	84.29	76.26	78.30	94.26	81.20	91.46
		ADD-S	97.08	95.65	94.13	98.12	100.00	97.52
SAM6D [14]	1	ADD	29.16	6.13	16.49	33.23	50.47	39.50
		ADD-S	61.69	14.55	60.82	94.10	67.20	71.81
	8	ADD	37.00	11.35	31.96	42.13	52.12	47.47
		ADD-S	70.35	33.67	64.95	95.10	76.25	81.81
	16	ADD	82.65	69.21	69.16	88.36	96.12	90.40
		ADD-S	93.00	90.22	76.29	99.93	99.59	98.99
Ours	1	ADD	50.11	40.30	34.22	40.10	83.63	52.31
		ADD-S	84.20	66.67	85.23	88.30	92.10	88.70
	8	ADD	67.27	52.21	52.10	71.10	86.67	74.27
		ADD-S	91.58	87.70	90.12	90.20	97.30	92.62
	16	ADD	85.00	77.21	81.10	88.20	90.21	88.30
		ADD-S	98.19	97.34	95.33	98.30	100.00	100.00

Table 5: Quantitative performance of *UnPose* compared to SOTA baselines on LM-O [54] for 6D Pose estimation. For both ADD and ADD-S metrics, we show higher values with brighter gradient and vice-versa.

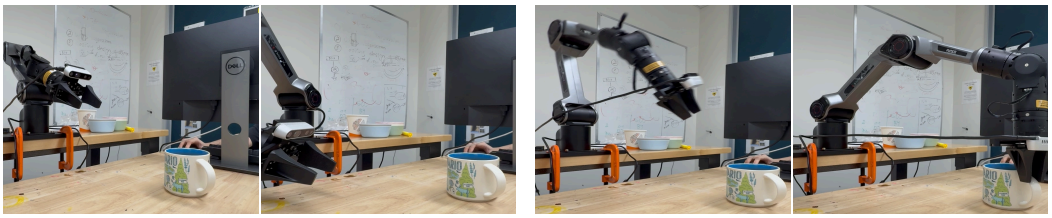


Figure 11: Visualization of *UnPose* for a real-world robotic manipulation task where it estimates the complete geometry and 6D pose of the cup.

ing the dynamic operation of the system capturing multiple views, the resulting reconstruction, and the robot successfully grasping the object, are presented in the accompanying video.

Method	# img	Metric	Mean	005_tomato_soup_can	006_mustard_bottle	007_tuna_fish_can	009_gelatin_box	010_potted_meat_can	011_banana
GigaPose [21]	1	ADD	13.58	8.20	4.15	1.77	45.92	4.72	16.74
		ADD-S	24.44	26.23	10.66	5.42	66.69	14.17	23.48
	8	ADD	21.48	17.74	14.03	3.33	75.00	6.62	12.16
		ADD-S	39.52	58.06	28.87	14.35	79.62	22.47	33.78
	16	ADD	60.58	88.80	62.38	13.67	93.75	62.35	42.56
		ADD-S	86.8	96.82	100.00	49.33	100.00	92.35	82.33
FoundationPose [30]	1	ADD	30.34	1.69	40.00	7.69	76.00	16.67	40.00
		ADD-S	74.58	42.01	86.90	71.10	100.00	72.10	75.40
	8	ADD	45.57	44.07	55.00	41.00	80.00	13.33	40.00
		ADD-S	90.17	76.20	87.33	91.20	100.00	89.10	97.20
	16	ADD	71.2	62.10	66.00	72.30	100.00	60.60	66.20
		ADD-S	95.56	98.37	95.00	100.00	100.00	80.00	100.00
SAM6D [14]	1	ADD	19.93	11.66	37.50	3.75	31.60	10.00	25.10
		ADD-S	43.04	13.33	86.83	16.76	52.10	50.00	39.20
	8	ADD	34.63	16.67	44.67	6.87	76.25	38.33	25.00
		ADD-S	63.86	36.67	82.33	31.00	96.70	65.17	71.28
	16	ADD	72.06	60.27	80.20	77.10	100.00	56.36	58.44
		ADD-S	97.02	98.42	100.00	100.00	100.00	83.73	100.00
Ours	1	ADD	47.92	20.34	55.00	38.20	80.00	44.00	50.00
		ADD-S	82.72	82.12	90.00	69.23	100.00	80.00	75.00
	8	ADD	61.2	56.60	70.00	53.60	87.00	40.00	60.00
		ADD-S	91.22	88.22	100.00	89.13	100.00	90.00	80.00
	16	ADD	74.39	64.98	80.00	66.67	91.00	76.67	67.00
		ADD-S	96.65	89.83	100.00	94.87	100.00	100.00	95.20

Table 6: Quantitative performance of *UnPose* compared to SOTA baselines on YCB-Video [53] for 6D Pose estimation. For both ADD and ADD-S metrics, we show higher values with brighter gradient and vice-versa.