ALIGNPOSE: GENERALIZABLE 6D POSE ESTIMATION VIA MULTI-VIEW FEATURE-METRIC ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-view RGB model-based object pose estimation methods achieve strong generalization performance but are fundamentally limited by depth ambiguity, clutter, and occlusions. Multi-view pose estimation methods have the potential to solve these issues, but existing works rely on precise single-view pose estimates or lack generalization to unseen objects. To address these challenges, we introduce AlignPose, a 6D object pose estimation method that aggregates information from multiple extrinsically calibrated views and generalizes to unseen objects. The contributions of this work are threefold. First, leveraging powerful, frozen features from a foundation model, AlignPose iteratively minimizes the discrepancy between rendered and observed images across multiple viewpoints, enforcing geometric consistency without object-specific training. Second, robust handling of noisy inputs is achieved by aggregating pose candidates from an arbitrary single-view pose estimator via 3D non-maximum suppression. Third, extensive experiments on three BOP benchmarks (YCB-V, T-LESS, ITODD-MV) show AlignPose sets a new state of the art, especially on challenging industrial datasets where multiple views are readily available in practice.

1 INTRODUCTION

Model-based 6D object pose estimation is crucial for real-world applications like robotic grasping and scene understanding. Recent deep learning-based methods achieve strong generalization in this task, even for objects unseen during training (Labbé et al., 2022; Örnek et al., 2024). However, these approaches are single-view and hence inherently struggle in cluttered scenes due to occlusions or appearance ambiguities (e.g., a cup with a hidden handle (Hodaň et al., 2016)). RGB methods are also prone to errors in camera-object pose estimates due to the depth ambiguity. On the other hand, depth-based estimators typically achieve superior precision but may fail for certain types of objects (e.g., reflective, transparent Chen et al. (2022)), and industrial depth cameras are often expensive. Leveraging multiple RGB views thus offers a compelling alternative to address these issues.

However, existing methods either rely on a strict multi-view matching procedure (Labbé et al., 2020) that may discard valid candidates or require object-specific training (Shugurov et al., 2021b), which may require hours or days of training time. Provided with initial object poses in a common reference frame and camera extrinsics, consistency between views and the object model pose can be enforced by minimizing an image quantity, as is the case for reprojection error in bundle adjustment (Triggs et al., 1999) or photometric error in direct SLAM (Engel et al., 2014). Generalization to unseen objects, however, remains a major challenge. In this work, we formulate the RGB multi-view pose estimation problem as a *multi-view feature-metric* image alignment by using frozen features from a foundation model (Oquab et al., 2023) to achieve generalizability to new objects without training.

Our contributions are as follows:

- We design an approach for multi-view 6D pose estimation from RGB images that does not require any object-specific training or symmetry annotation. We use per-view pose candidates, obtained by any existing single-view pose estimation method, aggregate them in a common 3D coordinate frame, and refine them to obtain a new, improved object pose estimate.

- We propose a new formulation of multi-view object pose refinement based on multi-view feature-metric alignment to minimize the discrepancy between the view-dependent features extracted from the 3D model and the multiple observed views. We use features from a pre-trained vision foundation model, which unlocks generalizability to new objects without any additional training.

- We evaluate our method on the BOP challenge (Hodan et al., 2024) YCBV, T-LESS, and ITODD-MV benchmarks. We demonstrate substantial improvements in performance measured by average recall and average precision. Our approach surpasses single-view estimates by 11% on average and state-of-the-art multi-view RGB-based pose estimates by 5%.

2 Related Work

Single-view 6D pose estimation. Model-based object detection and 6D pose estimation from RGB(-D) images is a well-established field whose progress is tracked in several benchmarks and datasets (Hodan et al., 2018; Van Nguyen et al., 2025). Leading approaches are all deep learning-based, and can be broadly classified into direct regression (Xiang et al., 2018), template-matching (Nguyen et al., 2024; Örnek et al., 2024; Caraffa et al., 2024), 2D-3D correspondence matching (Liu et al., 2025), and render-and-compare (Li et al., 2018b; Labbé et al., 2020; Labbé et al., 2022). A recent focus has been to develop methods able to generalize to objects that are not seen at training time (Hodan et al., 2024) by training on large simulated datasets (Labbé et al., 2022; Nguyen et al., 2024; Wen et al., 2024). Visual (Oquab et al., 2023) and point-cloud (Poiesi & Boscaini, 2022) foundation models have also been shown to provide useful features for building training-free generalizable pose estimation approaches (Örnek et al., 2024; Caraffa et al., 2024). We build on this line of work by leveraging a vision foundation model for the task of multi-view object pose estimation.

Multi-view object pose estimation. Multiple viewpoints can be used to improve single-view RGB-based pose estimation methods by resolving depth/scale ambiguity and occlusions. Some methods estimate jointly the object pose and temporally linked camera poses, a problem known as object-SLAM (Salas-Moreno et al., 2013; Fu et al., 2021; Zorina et al., 2024), while other assume access to camera poses, e.g. from robot kinematics (Wada et al., 2020) or off-line localization (Li et al., 2018a). One approach is to first build 3D scene representation, using either volumetric representation obtained from RGBD frames (Wada et al., 2020; Kaskman et al., 2020) or implicit representations (Taher et al., 2024) and to then register the object model in the scene. Others fuse multiple single-view pose estimates, associated with confidence scores (Labbé et al., 2022) or even probability distributions (Yang et al., 2023). They are first expressed in a common reference frame using the camera poses and can then be aggregated using voting schemes (Sock et al., 2017; Li et al., 2018a) or Maximum Likelihood Estimation (Erkent et al., 2016; Yang et al., 2023). CosyPose (Labbé et al., 2020) uses pose-level RANSAC (Fischler & Bolles, 1981) aggregation to recover consistent camera and object poses, and refine them using bundle adjustment. DPODv2 (Shugurov et al., 2021a) proposes a refinement based differentiable rendering by training an model to predict object NOCS representation (Wang et al., 2019). We introduce a simple but effective 3D Non Maximum Suppression (NMS) aggregation scheme followed by a multi-view feature-metric refinement which does not require any object specific training and can therefore generalize to any object model.

Feature-metric refinement. Direct image alignment (Irani & Anandan, 1999; Baker & Matthews, 2004) methods localize a camera by minimizing a photometric loss and have been used to build highly precise visual SLAM systems (Engel et al., 2014). However, their performance drops significantly when the brightness consistency assumption is violated, which often appears in visual localization tasks. To improve robustness, several works propose a *feature-metric* refinement approach where images are encoded using deep features (Shu et al., 2020; Sarlin et al., 2021) including pre-trained features (Trivigno et al., 2024). Inspired by those approaches, Örnek et al. (2024) minimizes a feature-metric error between pre-rendered templates and query images using an image foundation model (Oquab et al., 2023) but only considers single views. We build on this line of work and introduce an approach to minimize a feature-metric loss across multiple camera viewpoints.

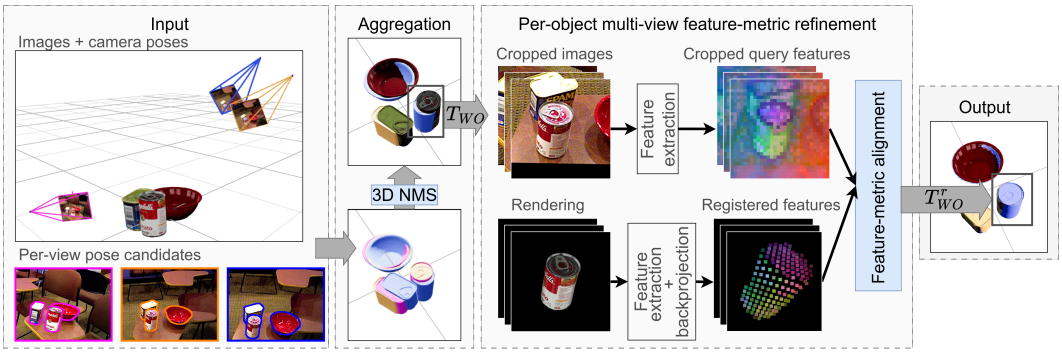

Figure 1: **Our feature-based multi-view pose estimation pipeline.** Single-view pose candidates are first generated independently for each view using state-of-the-art pose estimation methods (e.g. Labbé et al. (2022); Örnek et al. (2024)). During aggregation, candidates are transformed into a common coordinate frame, and non-maximum suppression (NMS) is applied to eliminate redundant detections of the same object. These filtered pose candidates $\boldsymbol{T}_{WO}$ are then refined using multi-view feature-metric refinement to obtain object poses that are consistent from all views.

3 Generalizable Multi-view object pose estimation

Problem formulation. The goal of multi-view 6D object pose estimation is to determine the 3D position and orientation of a previously unseen object, given its known 3D mesh and a set of RGB images from a calibrated multi-camera setup. The object's pose is represented by a single rigid transformation $\mathbf{T}_{WO} \in SE(3)$, which maps the object's local coordinate frame to a shared world coordinate frame. The pose observed by any individual camera, $\mathbf{T}_{CO} \in SE(3)$, is related to this world pose through the known camera extrinsics, $\mathbf{T}_{CW}$, via the kinematic chain: $\mathbf{T}_{CO} = \mathbf{T}_{CW}\mathbf{T}_{WO}$. Here $\mathbf{T}_{CW}$ is the transformation from the world frame to the camera's frame, which is assumed to be known from a one-time offline camera calibration procedure.

The primary challenges for multi-view pose estimation are threefold. First, the system must robustly aggregate multiple, often conflicting, pose candidates from different views to resolve per-view ambiguities and establish a single, globally consistent estimate. Second, this initial estimate must be refined by aligning it with subtle image features to achieve final accuracy. Finally, this entire aggregation-and-refinement pipeline must achieve zero-shot generalization, operating on novel objects using only their 3D mesh without any object-specific training. To address these challenges we propose multi-view pose estimation pipeline, as illustrated in Figure 1 and described next. Our approach leverages existing single-view pose estimators to produce per-view *candidate poses*, which are then aggregated, filtered, and refined to yield poses that are consistent in 3D and well-aligned with the input images. Individual components of the pipeline are described next.

3.1 Generation of single-view pose candidates

As an input to our method, we expect multiple RGB images of a scene, camera intrinsics, along with poses of cameras that captured these images relative to some world frame. For each view we let a single-view pose estimation method predict a set of *object pose candidates*. Each candidate is associated with a pose relative to the camera and a confidence score. We do not provide a specific implementation; rather, we assume the use of any off-the-shelf method for single-view pose estimation of unseen objects.

3.2 Multi-view Aggregation

The goal of the aggregation stage is to consolidate the noisy and redundant pose candidates from all single views into a clean, minimal set of unique 3D object pose candidates. The main challenge is that transforming these single-view estimates into a common world frame creates multiple, overlapping pose candidates for each physical object. To resolve this, we employ 3D Non-Maximum Suppression (NMS) to filter the duplicates. In this process, each candidate is represented by a 3D bounding box derived from its pose and 3D model, using the confidence score from the single-view

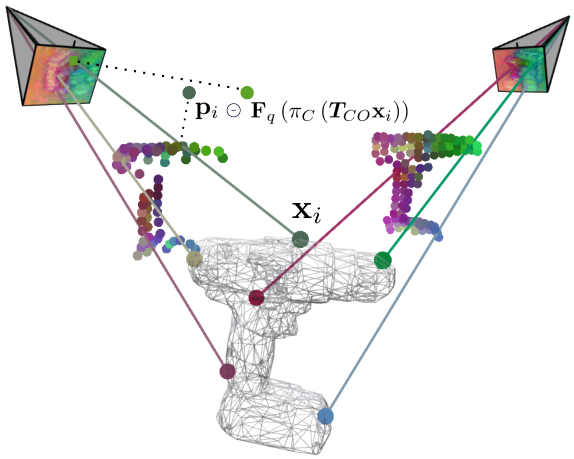

Figure 2: **Multi-view feature-metric refinement.** This figure illustrates a two-view feature-metric refinement. The two cameras show cropped query features (mapping the first three PCA components to RGB values). The partial multi-color point clouds represent the registered features $\mathcal{F}_{CO}$, shifted towards their corresponding camera for visualization purpose. The projection of registered feature 3D coordinates $\mathbf{x}_i$ is represented by a line colorcoded with a matching color of the feature value $\mathbf{p}_i$. During refinement, each registered feature $\mathbf{p}_i$ is compared to the feature value interpolated from the query feature image $\mathbf{F}_q$ at the location of the corresponding 3D point $\mathbf{x}_i$ projected into the image at $\pi_C\left(\boldsymbol{T}_{CO}\mathbf{x}_i\right)$, according to Equation 1. For clarity, only a few (3 per view) projected 3D points are shown.

estimator. Overlapping boxes are then suppressed based on their 3D Intersection over Union (IoU) if they exceed a predefined threshold. This stage concludes with a set of unique pose candidates, each with a coarse but consolidated pose with respect to the world coordinate frame, ready for refinement.

3.3 Multi-view Refinement

In the refinement stage, each pose candidate is refined independently via *multi-view feature-metric alignment*. The goal is to maximize the alignment between the projection of 3D registered features extracted from the 3D object model and 2D feature maps extracted from input images by optimizing the object pose using a multi-view feature-metric loss. In the following, we explain how the 3D registered features and 2D feature maps are obtained, and then describe our loss formulation.

Feature extraction. Before the multi-view alignment begins, we prepare two fixed representations for each input view: a *query 2D feature map* derived from the observed image and *3D registered features* derived from the object's coarse pose candidate. The query is the observed appearance of the scene that acts as the evidence we want to match. For each view, it is derived from the input image by cropping the region that (presumably) contains the object to a standardized size and extracting its feature map with a feature extractor. The 3D registered features represent view-dependent appearance of the object based on the coarse pose. This view-dependent appearance is represented by a set of *3D registered features* $\mathcal{F}_{CO} = \{\mathbf{p}_i, \mathbf{x}_i\}$. This is a point cloud where each 3D point $\mathbf{x}_i$ on the object's surface has an associated feature descriptor $\mathbf{p}_i$. We obtain this representation by first rendering color and depth frames for each view using the coarse object pose. We then extract features from the color render and register them in 3D by lifting them to the object's coordinate system using the rendered depth map. Note that while Örnek et al. (2024) pre-computes a large collection of registered features from template views, we only compute one set of registered features per view, which reduces the memory requirements of the method. More details can be found in the appendix A.

Per-view loss function. For each view, we define the *feature-metric loss* of a pose estimate $\boldsymbol{T}_{CO}$ as:

$$\mathcal{L}_{\text{FE}}^{C}(\boldsymbol{T}_{CO}) = \sum_{(\mathbf{p}_i, \mathbf{x}_i) \in \mathcal{F}_{CO}} \rho\left(\mathbf{p}_i - \mathbf{F}_q\left(\pi_C\left(\boldsymbol{T}_{CO}\mathbf{x}_i\right)\right)\right), \tag{1}$$

where $\mathcal{F}_{CO}$ is the set of registered features of the view and $\mathbf{F}_q$ are the 2D query features. Each registered object feature $\mathbf{p}_i$ is compared with its corresponding feature from the query feature map $\mathbf{F}_q\left(\pi_C\left(\boldsymbol{T}_{CO}\mathbf{x}_i\right)\right)$. The correspondence is obtained by transforming the 3D object point $\mathbf{x}_i$ from the object frame into the camera frame (by transformation $\mathbf{T}_{CO}\mathbf{x}_i$), projecting it into the query image (with camera projection π_C), and then sampling the query feature map $\mathbf{F}_q$ with bilinear interpolation at the given image coordinates. The loss for each registered feature is then calculated using a robust cost function $\rho(\cdot)$ by Barron (2019) with parameters $\alpha = -5$, $c = 0.5$.

Multi-view alignment. The goal of the multi-view alignment is to find a consistent object pose in the world frame $\boldsymbol{T}_{WO}^{r}$ that minimizes the feature-metric loss across all camera views. This is achieved by minimizing the sum of per-view feature-metric losses:

$$\boldsymbol{T}_{WO}^{r} = \underset{\boldsymbol{T}_{WO}}{\arg\min} \quad \sum_{C \in \mathcal{C}} \mathcal{L}_{\text{FE}}^{C}(\boldsymbol{T}_{CW}\boldsymbol{T}_{WO}), \tag{2}$$

where $\mathcal{C}$ is the set of all cameras. The optimized object pose $\boldsymbol{T}_{CO}$ is expressed in the camera frame following the kinematic chain $\boldsymbol{T}_{CO} = \boldsymbol{T}_{CW}\boldsymbol{T}_{WO}$. We visualise the multi-view alignment and the feature-metric loss in Figure 2.

Optimization. The pose is refined iteratively by the Levenberg-Marquardt(Levenberg, 1944; Marquardt, 1963) optimization algorithm until convergence or a maximum of 30 iterations. After the optimization, the refined pose $\mathbf{T}_{W,O}^{r}$ is assigned a score based on the average per-view error:

$$s(\boldsymbol{T}_{WO}^{r}) = 1 - \frac{1}{|\mathcal{C}|} \sum_{C \in \mathcal{C}} \mathcal{L}_{\text{FE}}^{C}(\mathbf{T}_{CW}\mathbf{T}_{WO}^{r}). \tag{3}$$

The score ranges from 0 to 1, with higher values indicating better poses, reaching its maximum when the pose best aligns the registered features with the query features. This score serves as a final confidence measure, quantifying how consistently the refined pose aligns with the visual evidence across all available camera views.

4 Experiments

Evaluation. We follow the BOP evaluation protocol (Hodaň et al., 2020) and use three pose-error functions: Maximum Symmetry-Aware Surface Distance (MSSD), which measures the maximum 3D distance between corresponding object vertices accounting for symmetries; Maximum Symmetry-Aware Projection Distance (MSPD), which computes the maximum 2D distance between the projected vertices of the predicted and ground-truth poses in the image plane; and Visible Surface Discrepancy (VSD), which measures the difference between visible object surfaces by comparing rendered depth maps, considering occlusions.

We evaluate our method on two tasks: 6D object localization, measured by Average Recall (AR) and 6D object detection, measured by Average Precision (AP) (see Hodaň et al. (2020)[A.1]). The AR metric measures the fraction of object instances for which a correct pose is found, averaged across several error thresholds. The AP metric follows an evaluation methodology similar to the COCO challenge (Lin et al., 2014) but it replaces the standard Intersection over Union (IoU) with the MSSD and MSPD pose errors. AP is generally a stricter metric than AR. See Van Nguyen et al. (2025) for more details. For each method, we report the average recalls for each error function ($\text{AR}_{\text{VSD}}, \text{AR}_{\text{MSSD}}, \text{AR}_{\text{MSPD}}$), their average $\text{AR} = (\text{AR}_{\text{VSD}} + \text{AR}_{\text{MSSD}} + \text{AR}_{\text{MSPD}})/3$, as well as the average precisions for error functions ($\text{AP}_{\text{MSSD}}, \text{AP}_{\text{MSPD}}$), and their average $\text{AP} = (\text{AP}_{\text{MSSD}} + \text{AP}_{\text{MSPD}})/2$.

Datasets. We evaluate our method on three benchmark datasets from the BOP benchmark: YCB-V (Xiang et al., 2018), T-LESS (Hodaň et al., 2017), and ITODD (Drost et al., 2017). YCB-V contains household objects with texture. T-LESS includes industry-relevant, texture-less objects, often arranged in heavily cluttered environments. ITODD dataset is composed of small, metallic,

Table 1: **Multi-view pose estimation of unseen objects on YCB-V, T-LESS, and ITODD.** Both our method and CosyPose (Labbé et al., 2020) refine candidates from FoundPose (Örnek et al., 2024), GigaPose (Nguyen et al., 2024), MegaPose (Labbé et al., 2022), and Co-Op (Moon et al., 2025). Our approach achieves higher performance across datasets. On ITODD, we omit the aggregation stage and refine candidates from one view only; CosyPose is unable to refine such candidates.

Dataset	Method	AR	AR_{VSD}	AR_{MSSD}	AR_{MSPD}	AP	AP_{MSSD}	AP_{MSPD}
YCB-V	FoundPose	69.0	60.2	67.0	79.7	63.0	54.5	71.4
	+ CosyPose MV	79.2	73.6	81.4	82.7	76.1	74.3	77.8
	+ Ours	**84.2**	**79.0**	**89.0**	**84.5**	**83.0**	**85.2**	**80.8**
	GigaPose	66.6	57.4	64.2	78.2	63.1	54.2	72.0
	+ CosyPose MV	76.5	70.3	77.8	81.2	70.9	68.3	73.4
	+ Ours	**82.7**	**77.9**	**87.4**	**82.3**	**80.2**	**82.4**	**78.0**
	MegaPose	62.0	53.5	59.7	72.8	56.1	47.8	64.5
	+ CosyPose MV	71.1	65.1	72.3	75.8	64.5	61.8	67.2
	+ Ours	**80.2**	**75.4**	**84.7**	**80.5**	**77.1**	**79.0**	**75.3**
	Co-op	69.7	58.3	66.3	84.6	69.5	57.8	81.2
	+ CosyPose MV	81.0	73.9	83.0	**86.1**	79.2	76.3	**82.1**
	+ Ours	**83.2**	**78.0**	**88.1**	83.5	**82.1**	**84.2**	80.0
T-LESS	FoundPose	57.0	53.6	54.9	62.3	57.0	52.9	61.1
	+ CosyPose MV	66.4	62.4	63.9	67.0	63.0	62.1	64.0
	+ Ours	**82.0**	**77.7**	**83.7**	**84.5**	**86.9**	**86.9**	**86.9**
	GigaPose	58.2	54.9	56.0	63.7	54.3	50.8	57.8
	+ CosyPose MV	61.9	59.5	60.6	65.6	55.9	55.5	57.3
	+ Ours	**80.7**	**76.8**	**82.4**	**83.1**	**83.8**	**83.8**	**83.8**
	Megapose	50.8	48.1	48.5	55.9	50.5	46.6	54.4
	+ CosyPose MV	57.6	55.8	56.7	60.3	56.1	54.9	57.3
	+ Ours	**76.9**	**73.0**	**78.6**	**79.2**	**79.8**	**79.8**	**79.8**
	Co-op	68.2	64.0	65.8	74.8	68.9	64.3	73.4
	+ CosyPose MV	78.7	76.9	78.5	80.7	78.9	78.3	79.4
	+ Ours	**86.0**	**81.6**	**87.9**	**88.6**	**89.4**	**89.4**	**89.3**
ITODD	Co-op	50.6	42.7	47.5	**61.7**	50.4	44.1	**56.7**
	+ Ours	**54.2**	**47.6**	**56.2**	58.8	**54.5**	**54.4**	54.7

and reflective objects. It is captured in grayscale and represents industrial settings. We use the ITODD-MV version provided by BOP, which includes multi-view data.

For YCB-V and T-LESS, we sample view groups out of the test targets following the same procedure as (Labbé et al., 2020) for a fair comparison. In contrast, ITODD provides exactly four predefined views for each scene, so no sampling is necessary. All datasets include camera calibration.

4.1 Unseen object pose estimation task

Our work introduces a generalizable multi-view pose estimation method, capable of operating on novel objects without object-specific training. To validate this generalization capability, we evaluate our approach in the unseen object pose estimation setting defined by (Hodan et al., 2024). In this scenario, the pose estimation method is not permitted to train on the 3D object models; only a brief onboarding stage is allowed. To obtain input single-view pose candidates, we use state-of-the-art single-view pose estimation methods that operate on unseen objects[1]: FoundPose (Örnek et al., 2024), GigaPose (Nguyen et al., 2024), MegaPose (Labbé et al., 2022) and Co-Op (Moon et al., 2025).

[1]We use single-view pose candidates downloaded from BOP Leaderboard, specifically the following versions: FoundPose+FeatRef+Megapose-5hyp, GigaPose+GenFlow (5 hypotheses), MegaPose-CNOS_fastSAM+MultiHyp, Co-op (F3DT2D, 5 Hypo)

Table 2: **Seen object pose estimation task on the T-LESS dataset.** We compare our multi-view pose estimation method with multi-view pose estimation methods CenDerNet(Haugaard & Iversen, 2023), DPODv2(Shugurov et al., 2021b) and CosyPose(Labbé et al., 2020). Our method gives the best results across all metrics. For baseline results we rely on publicly available results and therefore not all metrics are available by the baseline methods.

Data	**Method**	#views	AR	AR_{VSD}	AR_{MSSD}	AR_{MSPD}	AP	AP_{MSSD}	AP_{MSPD}
synt	CenDerNet	5	71.3	70.7	71.7	71.5	*not available*		
	DPODv2	1	63.6	56.1	60.2	74.4	*not available*		
	+ DPODv2 MV	4	68.9	64.5	71.0	71.2	*not available*		
	CosyPose	1	64.0	57.1	58.9	76.1	63.0	54.7	71.4
	+ CosyPose MV	4	72.8	68.2	71.5	78.9	70.6	68.6	72.7
	+ Our refinement	4	**85.5**	**80.3**	**87.6**	**88.7**	**89.2**	**88.9**	**89.4**
synt+real	DPODv2	1	65.5	57.9	62.1	76.4	*not available*		
	DPODv2 MV	4	72.0	67.9	74.2	74.0	*not available*		
	CosyPose	1	72.8	66.9	69.5	82.1	74.1	67.8	80.3
	+ CosyPose MV	4	81.5	77.9	81.4	85.4	81.7	80.6	82.7
	+ Our refinement	4	**86.8**	**81.9**	**88.9**	**89.7**	**91.0**	**91.0**	**91.0**

YCBV and T-LESS evaluation. We compare our multi-view pose estimation method against the multi-view aggregation-and-refinement strategy proposed in CosyPose (Labbé et al., 2020). To our knowledge, this is the only available method capable of multi-view pose refinement of unseen objects. Unlike our method, CosyPose jointly refines object and camera poses and does not require known camera extrinsics. To ensure a fair comparison, we evaluate CosyPose in a setting where camera extrinsics are known and the refinement is applied only to object poses. Additionally, CosyPose relies on annotated object symmetries, while our method is symmetry-agnostic. The results are shown in Table 1. The table shows that our multi-view pose estimation method significantly outperforms CosyPose multi-view on both YCB-V and T-LESS datasets.

An exception to the overall trend appears in the MSPD error when refining candidates from Co-Op (Moon et al., 2025). In this case, our method does not yield an improvement in this single reprojection error metric. A likely explanation is that our refinement adjusts the object pose to better align in spatial/depth direction, thereby reducing MSSD. However, this stricter alignment in 3D can sometimes lead to a slight increase in the 2D projection error (MSPD), for instance when the refined pose corrects depth or orientation errors in ways that shift the projected silhouette. This highlights a trade-off: optimizing for geometric consistency in 3D may not always translate to lower reprojection error in 2D, especially when the input candidates have a very low reprojection error.

ITODD evaluation. For ITODD dataset, we use pose candidates from one of the four views as input and refine them with respect to features from all of the views as single-view pose candidates for all views were not publicly available. The results (see Table 1) show that our refinement improves performance even without the aggregation stage, though the improvement compared to the single-view input is not as substantial. In contrast, CosyPose multi-view refinement cannot be applied in this setting.

Qualitative results. Qualitative results for YCB-V dataset are shown in Figure 4.1 and for T-LESS in Figure 4.1. Our qualitative results demonstrate the superior performance of our multi-view method over both the initial single-view estimates and the CosyPose refinement. For each scene, we compare the initial single-view pose candidates against the refined multi-view outputs from both CosyPose and our approach. A key advantage of our method is its robustness. For instance, CosyPose often fails to produce a final estimate when its RANSAC-based aggregation step discards all initial pose candidates for an object. In contrast, our method successfully processes and refines any object pose detected in at least one valid view, showcasing a significant improvement in reliability.

4.2 Seen Object Pose Estimation Task

Although our method is designed for the primary challenge of generalizable pose estimation for unseen objects, we also benchmark its performance in the seen object setting. This allows for a

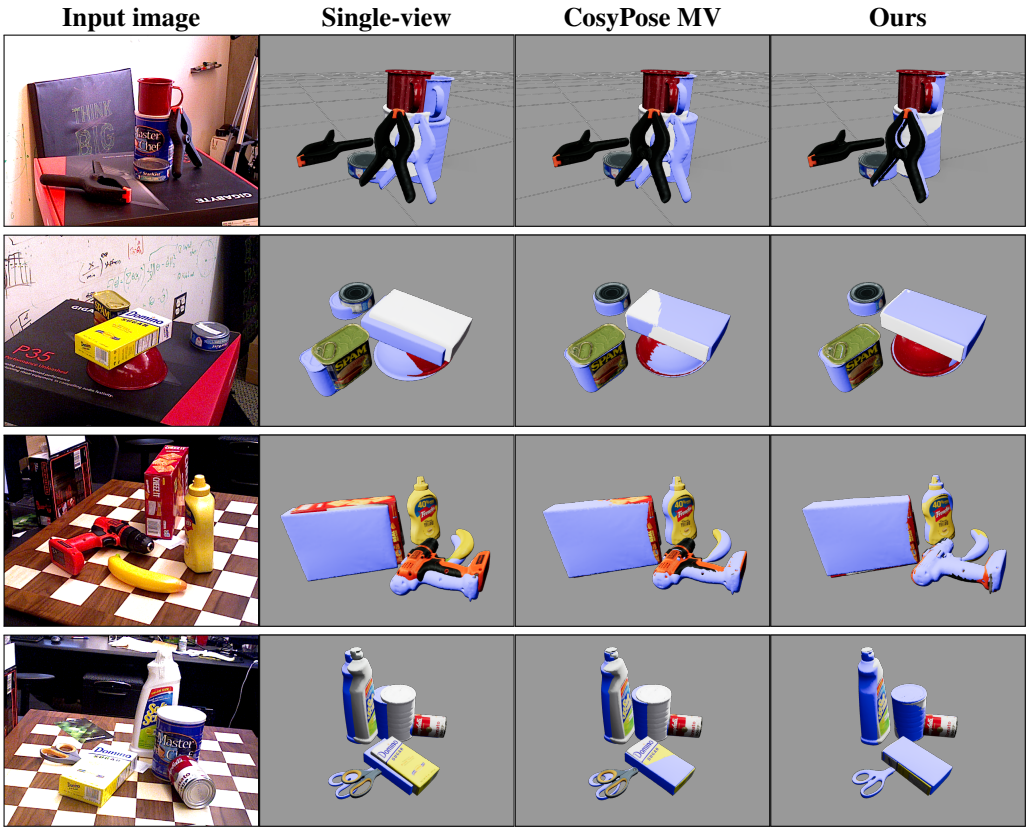

Figure 3: **Qualitative results of multi-view refinement on the YCB-V dataset.** We show one input image in the first column and its corresponding single-view pose estimates from the Co-op method (Moon et al., 2025) shown in blue in the second column. The ground-truth object poses are shown with textures. The third column presents the results of CosyPose multi-view pose estimation and the forth column shows the results of our multi-view pose estimation method. CosyPose and our method both used Co-op pose candidates from four input images/views. Our approach refines these single-view candidates more effectively than CosyPose, producing more accurate 6D poses.

direct comparison against a wider range of established multi-view baselines that require object-specific training on the exact test objects. For this evaluation, we focus on the industry-relevant T-LESS dataset for which existing multi-view pose estimation methods provide results publicly.

We show the results for seen object pose estimation in Table 2. We compare with methods trained on synthetic data (synt) as well as those trained on a mix of synthetic and real data (synth+real), namely CenDerNet (Haugaard & Iversen, 2023), DPODv2 (Shugurov et al., 2021b), and CosyPose (Labbé et al., 2020). DPODv2 and CosyPose both have a single-view version and a multi-view (MV) version that combines and refines these single-view inputs. Results for CosyPose MV were obtained by running the original pipeline with known camera extrinsics. As shown in Table 2, our method outperforms all baselines across all metrics. This indicates that, despite requiring no training, our approach is competitive even against methods explicitly trained on the test objects. For fairness, we refine pose candidates generated by CosyPose's single-view estimator, which was trained on T-LESS, ensuring the inputs to our refinement are comparable to those used by the baselines.

4.3 Ablations

To understand and quantify the contribution of each component within our pipeline, we performed an ablation study. The results are shown in Table 3. Aggregating single-view candidates from multiple views slightly increases AR, as more objects become visible, but substantially reduces AP due to redundant or inconsistent pose estimates. Applying NMS restores precision by keeping only the best candidates. Finally, the full pipeline, including refinement, demonstrates that visual-feature-based refinement is crucial for achieving high-quality 6D pose estimates.

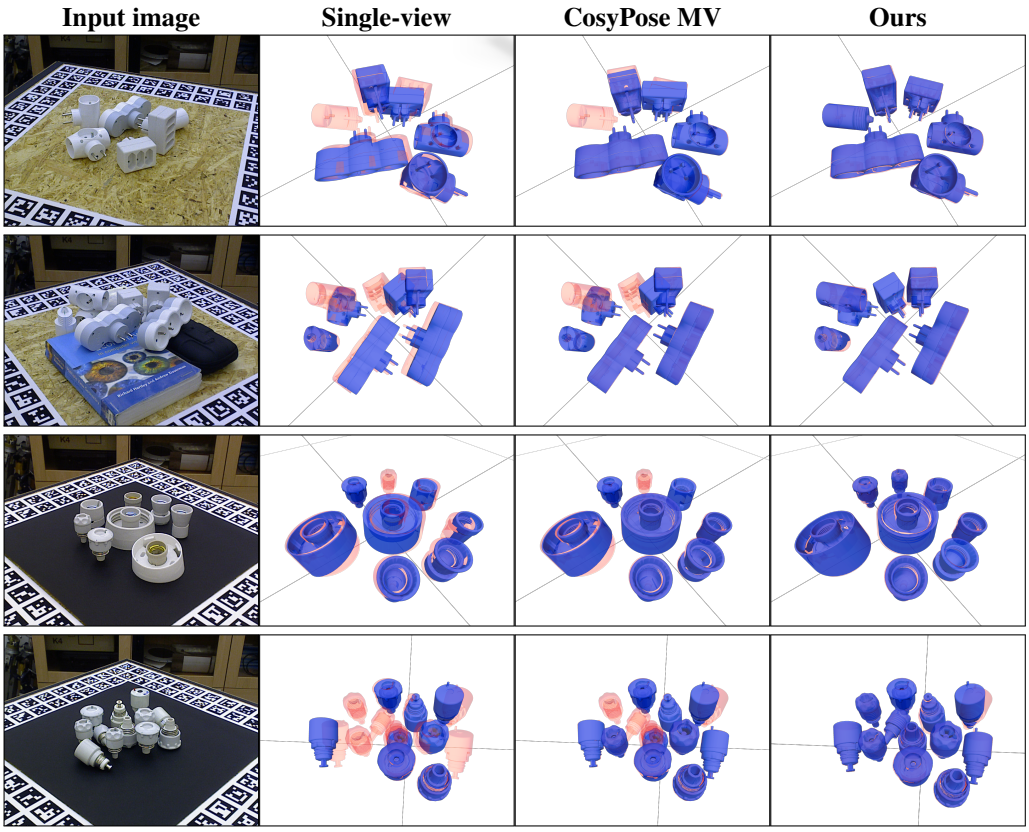

Figure 4: **Qualitative results of multi-view refinement on the T-LESS dataset.** We show one input image in the first column. Single-view pose estimates obtained by MegaPose (Labbé et al., 2022) are shown in blue in the second column. Ground-truth poses are highlighted in red. The third column presents the results of CosyPose multi-view estimation and the fourth column shows the results of our multi-view pose estimation method. CosyPose and our method both use MegaPose single-view pose candidates from four input views/images. Our approach refines the poses better than CosyPose (see rows 1-2) and is able to correctly retrieve poses for more objects (see rows 1-4) in this challenging setting with multiple textureless objects.

5 Conclusion

In this work, we introduce AlignPose, a novel and effective method for multi-view generalizable 6D object pose estimation. The method efficiently aggregates information from multiple views and generates high-quality pose estimates. We show that a frozen vision foundation model can be leveraged for this task as a powerful feature extractor. This allows AlignPose to achieve remarkable generalization to new, unseen objects without any additional training. Our strategy sets a new state-of-the-art on key BOP benchmarks and provides a powerful solution that is well-suited for industrial setups with challenging conditions, where multi-view information is most useful.

Table 3: **Ablation of the key components of our method on the T-LESS dataset.** Performance of the single-view input to our method (1v (Co-op)), followed by successive stages of our approach: multi-view aggregation (4v aggregate), 3D non-maximum suppression (4v aggregate + NMS), and multi-view refinement (4v aggregate + NMS + refinement).

Method	AR	AR_{VSD}	AR_{MSSD}	AR_{MSPD}	AP	AP_{MSSD}	AP_{MSPD}
1v (Co-op)	68.2	64.0	65.8	74.8	68.9	64.3	73.4
4v aggregate	69.3	56.4	73.8	77.7	48.6	47.4	49.8
4v aggregate + NMS	73.1	58.4	78.3	82.7	80.3	79.0	81.5
4v aggregate + NMS + refinement	**86.0**	**81.6**	**87.9**	**88.6**	**89.4**	**89.4**	**89.3**

REFERENCES

Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *International journal of computer vision*, 56(3):221–255, 2004.

Jonathan T Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4331–4339, 2019.

Andrea Caraffa, Davide Boscaini, Amir Hamza, and Fabio Poiesi. Freeze: Training-free zero-shot 6d pose estimation with geometric and vision foundation models. In *European Conference on Computer Vision*, pp. 414–431. Springer, 2024.

Xiaotong Chen, Huijie Zhang, Zeren Yu, Anthony Opipari, and Odest Chadwicke Jenkins. Clearpose: Large-scale transparent object dataset and benchmark. In *European conference on computer vision*, pp. 381–396. Springer, 2022.

Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Härtinger, and Carsten Steger. Introducing mvtec itodd — a dataset for 3d object recognition in industry. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 2200–2208, 2017. doi: 10.1109/ICCVW.2017.257.

Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pp. 834–849. Springer, 2014.

Özgür Erkent, Dadhichi Shukla, and Justus Piater. Integration of probabilistic pose estimates from multiple views. In *European Conference on Computer Vision*, pp. 154–170. Springer, 2016.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Jiahui Fu, Qiangqiang Huang, Kevin Doherty, Yue Wang, and John J Leonard. A multi-hypothesis approach to pose ambiguity in object-based slam. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7639–7646. IEEE, 2021.

Rasmus Laurvig Haugaard and Thorbjorn Mosekjaer Iversen. Multi-view object pose estimation from correspondence distributions and epipolar geometry. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1786–1792, 2023. doi: 10.1109/ICRA48891.2023.10161514.

Tomáš Hodaň, Jiří Matas, and Štěpán Obdržálek. On evaluation of 6d object pose estimation. In *European conference on computer vision*, pp. 606–619. Springer, 2016.

Tomáš Hodaň, Pavel Haluza, Štěpán Obdržálek, Jiří Matas, Manolis Lourakis, and Xenophon Zabulis. T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017.

Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 19–34, 2018.

Tomáš Hodaň, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiří Matas. Bop challenge 2020 on 6d object localization. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 577–594. Springer, 2020.

Tomas Hodan, Martin Sundermeyer, Yann Labbe, Van Nguyen Nguyen, Gu Wang, Eric Brachmann, Bertram Drost, Vincent Lepetit, Carsten Rother, and Jiri Matas. Bop challenge 2023 on detection segmentation and pose estimation of seen and unseen rigid objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 5610–5619, 6 2024.

Michal Irani and Prabu Anandan. About direct methods. In *International Workshop on Vision Algorithms*, pp. 267–277. Springer, 1999.

Roman Kaskman, Ivan Shugurov, Sergey Zakharov, and Slobodan Ilic. 6 dof pose estimation of textureless objects from multiple rgb frames. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 612–630. Springer, 2020.

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

Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.

Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944.

Chi Li, Jin Bai, and Gregory D Hager. A unified framework for multi-view multi-class object pose estimation. In *Proceedings of the european conference on computer vision (eccv)*, pp. 254–269, 2018a.

Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 683–698, 2018b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Xingyu Liu, Ruida Zhang, Chenyangguang Zhang, Gu Wang, Jiwen Tang, Zhigang Li, and Xiangyang Ji. Gdrnpp: A geometry-guided and fully learning-based object pose estimator. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

Sungphill Moon, Hyeontae Son, Dongcheol Hur, and Sangwook Kim. Co-op: Correspondence-based novel object pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11622–11632, 2025.

Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. Gigapose: Fast and robust novel object pose estimation via one correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9903–9913, 2024.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomas Hodan. Foundpose: Unseen object pose estimation with foundation features. In *European Conference on Computer Vision*, pp. 163–182. Springer, 2024.

Fabio Poiesi and Davide Boscaini. Learning general and distinctive 3d local deep descriptors for point cloud registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3979–3985, 2022.

Renato F Salas-Moreno, Richard A Newcombe, Hauke Strasdat, Paul HJ Kelly, and Andrew J Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1352–1359, 2013.

Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, et al. Back to the feature: Learning robust camera localization from pixels to pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3247–3257, 2021.

Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *European Conference on Computer Vision*, pp. 572–588. Springer, 2020.

Ivan Shugurov, Ivan Pavlov, Sergey Zakharov, and Slobodan Ilic. Multi-view object pose refinement with differentiable renderer. *IEEE Robotics and Automation Letters*, 6(2):2579–2586, 2021a.

Ivan Shugurov, Sergey Zakharov, and Slobodan Ilic. Dpodv2: Dense correspondence-based 6 dof pose estimation. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7417–7435, 2021b.

Juil Sock, S Hamidreza Kasaei, Luis Seabra Lopes, and Tae-Kyun Kim. Multi-view 6d object pose estimation and camera motion planning using rgbd images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 2228–2235, 2017.

Marwan Taher, Ignacio Alzugaray, and Andrew J Davison. Fit-ngp: Fitting object models to neural graphics primitives. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 18186–18192. IEEE, 2024.

Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*, pp. 298–372. Springer, 1999.

Gabriele Trivigno, Carlo Masone, Barbara Caputo, and Torsten Sattler. The unreasonable effectiveness of pre-trained features for camera pose refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12786–12798, 2024.

Nguyen Van Nguyen, Stephen Tyree, Andrew Guo, Mederic Fourmy, Anas Gouda, Taeyeop Lee, Sungphill Moon, Hyeontae Son, Lukas Ranftl, Jonathan Tremblay, Eric Brachmann, et al. Bop challenge 2024 on model-based and model-free 6d object pose estimation. *CoRR*, 2025.

Kentaro Wada, Edgar Sucar, Stephen James, Daniel Lenton, and Andrew J Davison. Morefusion: Multi-object reasoning for 6d pose estimation from volumetric fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14540–14549, 2020.

He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2642–2651, 2019.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17868–17879, 2024.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Proceedings of Robotics: Science and Systems (RSS)*, 2018. URL `https://www.roboticsproceedings.org/rss14/p19.pdf`.

Jun Yang, Wenjie Xue, Sahar Ghavidel, and Steven L Waslander. 6d pose estimation for textureless objects on rgb frames using multi-view optimization. In *2023 IEEE international conference on robotics and automation (ICRA)*, pp. 2905–2912. IEEE, 2023.

Kateryna Zorina, Vojtech Priban, Mederic Fourmy, Josef Sivic, and Vladimir Petrik. Temporally consistent object 6d pose estimation for robot control. *IEEE Robotics and Automation Letters*, 2024.

A APPENDIX

A.1 IMPLEMENTATION DETAILS

As a preparation for multi-view alignment, we prepare two distinct, fixed representations for each view: a **2D query feature map** a set of **3D registered features**. In this section we describe the generation process based on the work of Örnek et al. (2024). We now describe how the query features and registered features are created for some view associated to camera C and for an object O with a coarse pose estimate $\mathbf{T}_{WO}$.

2D query features. Given an 2D bounding box obtained from the coarse pose $\mathbf{T}_{WO}$, we crop the image using the perspective cropping method implemented by Örnek et al. (2024). For each view, this process yields a cropped image of size 420×420 and a corresponding crop camera C' with pose $\boldsymbol{T}_{WC'}$ in the world coordinates. We then extract a 2D feature maps using DINOv2: the cropped image is decomposed into a grid of non-overlapping 14×14 patches and each of these patch i, is embeded with a feature descriptor. We use the hidden state of layer 18 of the DINOv2 backbone which was empirically found to provide a good balance between positional information and semantic abstraction (Örnek et al., 2024). The resulting feature map is then upsampled to the crop resolution via bilinear interpolation.

3D registered features. To generate the registered feature, we render an RGB-D image of the object as it would appear from the crop camera C' given the coarse pose $\boldsymbol{T}_{WO}$. After rendering, we extract a feature descriptor for each patch, analogously to the process used for the query features, but we keep only the descriptors of patches whose center belong to the object and not the background. Additionally, we lift the 2D descriptors into 3D object space. Each descriptor $\mathbf{p}_i$ corresponding to a patch centered at pixel c_i is assigned the 3D point $\mathbf{x}_i$ in object coordinates that projects to c_i. This yields a set of *registered features* $\mathcal{F}_{CO} = \{\mathbf{p}_i, \mathbf{x}_i\}$ where $\mathbf{p}_i$ is a patch descriptor and $\mathbf{x}_i$ its corresponding 3D location in object space.

Dimensionality reduction. Feature extractors such as DINOv2 produce high-dimensional descriptors (e.g., 1024 dimensions), which can be computationally expensive for optimization. To make the process tractable, we reduce the dimensionality of these descriptors using Principal Component Analysis (PCA). The principal components are pre-computed for each object type during an offline onboarding stage and then applied to all subsequent patch descriptors of that object. Specifically, the components are estimated from feature descriptors extracted from renders of the object observed under a diverse set of viewpoints.