Diffusion Feature Fusion for Novel Object Pose Estimation

A Thesis Presented

by

Vijayabharathi Murugan

to

The Graduate School

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in

Computer Science

Stony Brook University

May 2026

Stony Brook University

The Graduate School

Vijayabharathi Murugan

We, the thesis committee for the above candidate for the Master of Science degree, hereby recommend acceptance of this thesis.

Zhaozheng Yin – Thesis Advisor
SUNY Empire Innovation Associate Professor, Department of Computer Science, Stony Brook University

Paola Cascante-Bonilla – Second Committee Member
Assistant Professor, Department of Computer Science, Stony Brook University

Chenyu You – Additional Committee Member
Assistant Professor, Department of Applied Mathematics and Statistics and Department of Computer Science, Stony Brook University

This thesis is accepted by the Graduate School.

Celia Marshik
Dean of the Graduate School

Abstract of the Thesis

Diffusion Feature Fusion for Novel Object Pose Estimation

by

Vijayabharathi Murugan

Master of Science

in

Computer Science

Stony Brook University

2026

Novel object pose estimation requires estimating the full 6DoF pose of unseen objects using only a CAD model, without object-specific training. Discriminative features such as DINOv2 are effective but lack spatial smoothness and struggle on textureless objects. Diffusion models produce complementary spatially smooth representations, but their utility in the synthetic-to-real setting has not been studied. We propose a training-free pipeline that fuses diffusion and DINOv2 features into 3D point descriptors, and recovers pose via Kabsch+RANSAC. Evaluated on all seven core BOP datasets, the fused pipeline outperforms DINOv2 alone with a mean gain of +6.6 Average Recall (AR) across all datasets, with gains of up to +8.5 AR on TUD-L and YCB-V and +6.0 AR on LM-O. It also outperforms DZOP (an RGB-only method using diffusion features) on LM-O and YCB-V. Ablations show that low-noise features ($t \leq 50$) and all four decoder layers $\{\ell_2, \ell_5, \ell_8, \ell_{11}\}$ are preferable for object-level matching. This is a preliminary study; further work is needed to close the gap to state-of-the-art RGB-D methods.

Contents

List of Figures vi

List of Tables vii

Acknowledgements viii

1 Introduction 1

2 Related Work 3
2.1 Novel Object Pose Estimation 3
2.2 Diffusion Features for Visual Correspondence 4

3 Background 6
3.1 Kabsch Algorithm 6
3.2 BOP Benchmark and Average Recall 7

4 Methodology 8
4.1 Feature Extraction 8
4.2 Dimensionality Reduction via PCA 10
4.3 3D Point Lifting 10
4.4 Feature Fusion 11
4.5 Query Processing and Pose Estimation 12

5 Results and Discussion 13
5.1 Experimental Setup 13
5.1.1 Datasets and Evaluation Protocol 13
5.1.2 Implementation Details 14
5.2 Main Results 14
5.3 Ablation Study 15
5.3.1 Fusion Weight 15
5.3.2 Diffusion Timestep 16
5.3.3 Layer Selection 16

5.4 Qualitative Feature Visualization 17

6 Conclusion **19**
6.1 Limitations 19
6.2 Future Work 19

Bibliography **21**

List of Figures

4.1 Overview of our zero-shot 6DoF pose estimation pipeline. During onboarding, diffusion and DINOv2 features are extracted from multi-view CAD renders and aggregated into 3D point descriptors per object. At inference, query features are extracted from the detected object crop, fused with the corresponding modality weights, and matched against the precomputed object point descriptors via cosine nearest-neighbour search. The resulting 3D-3D correspondences are passed to Kabsch+RANSAC to recover the final 6DoF pose $(\mathbf{R}, \mathbf{t})$. 9

5.1 PCA visualization of RGB, DIFT, DINOv2, and fused per-point embeddings respectively for representative YCB-V objects. . . 18

List of Tables

5.1 Comparison with state-of-the-art zero-shot 6DoF pose estimation methods on the seven core BOP datasets. AR (%) is reported per dataset. ✓ denotes training-free methods. Results for rows 1–7 are taken from the respective published papers. Our method uses CNOS detections for all datasets. 14

5.2 Effect of fusion weight α on Average Recall (%). α=0 is DINOv2 only; α=1 is DIFT only. Results with $\mathcal{L} = \{\ell_2, \ell_5, \ell_8, \ell_{11}\}$ and t=50. 15

5.3 Effect of diffusion timestep t on Average Recall (%). Results with α=0.5 and $\mathcal{L} = \{\ell_2, \ell_5, \ell_8, \ell_{11}\}$. 16

5.4 Effect of decoder layer selection on Average Recall (%). Results with α=0.5 and t=50. 17

Acknowledgements

I would like to thank my thesis advisor, Professor Zhaozheng Yin, for his guidance and support throughout this project. I am also grateful to my family and friends for their encouragement and support.

Chapter 1

Introduction

Estimating the 6DoF pose of objects — their full 3D rotation and translation relative to a camera — is a core problem in computer vision with applications in robotic manipulation, augmented reality, and industrial inspection. Recent work has shifted focus toward the *zero-shot* setting, where pose must be estimated for objects not seen during training, using only a CAD model as reference. This setting is practically important: deploying a system that requires object-specific training data is expensive and does not scale to the long tail of objects encountered in the real world.

The dominant approach in zero-shot pose estimation is template matching: features are extracted from the query image and matched against features extracted from multi-view CAD renders. A standard novel object pose estimation pipeline consists of three stages: object detection and segmentation, coarse pose estimation, and pose refinement. This work focuses on the coarse pose estimation stage, which establishes an initial pose hypothesis from appearance-based feature matching. The central challenge is the *synthetic-to-real domain gap* — features computed from clean CAD renders must match features from real images with sensor noise, varied illumination, and occlusion. Vision foundation models such as DINOv2 [1] have proven effective in this role, providing view-invariant patch descriptors that transfer well across the domain gap. However, purely discriminative features lack spatial smoothness, which can limit correspondence density in textureless or uniformly coloured regions.

Generative diffusion models have recently been shown to produce internal representations — *diffusion features* — that encode rich spatial and semantic information suitable for visual correspondence [2]. Zhang et al. [3] showed that diffusion and DINOv2 features are complementary in the real-to-real setting: their normalized concatenation outperforms either feature alone by substantial margins on semantic correspondence benchmarks. Whether this complementarity extends to the synthetic-to-real setting of zero-shot pose estimation,

where the domain gap introduces an additional source of feature degradation, has not been studied.

In this work, we investigate the use of diffusion features for zero-shot 6DoF object pose estimation. We propose a pipeline that extracts diffusion and DINOv2 features from multi-view CAD renders, aggregates them into 3D point descriptors via a lifting procedure, and fuses the two feature types via normalized concatenation at inference. Pose is recovered from 3D-3D correspondences using Kabsch+RANSAC. We evaluate on the seven core BOP benchmark datasets and conduct ablations on fusion weight, diffusion timestep, and decoder layer selection.

Contributions.

- We present a training-free pipeline for zero-shot 6DoF object pose estimation that fuses diffusion and DINOv2 features at the level of 3D point descriptors, providing a preliminary investigation into whether the complementarity finding of Zhang et al. [3] extends to the synthetic-to-real setting.
- We show that the fused pipeline consistently outperforms DINOv2 alone across evaluated BOP datasets and outperforms the closest diffusion-feature baseline DZOP [4] on LM-O and YCB-V, though a gap to state-of-the-art RGB-D methods remains.
- Through ablations on timestep and decoder layer selection, we find that low-noise diffusion features ($t \leq 50$) and multi-layer concatenation are preferable for object-level synthetic-to-real matching.

Chapter 2

Related Work

2.1 Novel Object Pose Estimation

Novel object pose estimation requires estimating the full 3D rotation and translation of objects not encountered during training, typically using a CAD model as the sole reference. A common pipeline first detects and segments object instances in the query image, then establishes correspondences between the query and rendered CAD templates to recover the pose.

On the detection side, Nguyen et al. [5] propose CNOS, a strong baseline for CAD-based novel object segmentation that combines Segment Anything [6] for mask proposal generation with DINOv2 [1] CLS-token similarity for scoring proposals against multi-view CAD renders. CNOS won the Best Method Award for unseen object 2D detection at ICCV 2023 R6D workshop and serves as the detection front-end in most subsequent zero-shot pose estimation pipelines, including ours.

For pose estimation, methods diverge primarily in how correspondences are established and whether task-specific training is used. Labbé et al. [7] (MegaPose) demonstrated that training a coarse pose estimator and a learned refiner on two million synthetic renders of 20,000 objects enables strong generalization to novel objects at test time. Nguyen et al. [8] (GigaPose) accelerated this paradigm with a hybrid template-patch approach in which dense ViT features determine the best-matching template and a single 2D-2D correspondence from a lightweight patch network yields the full 6DoF pose, achieving 38× faster coarse pose estimation than MegaPose. Moon et al. [9] (GenFlow) approached the problem through iterative pose refinement, predicting dense optical flow between a rendered image at the current pose estimate and the observed image, constrained by the target object's 3D shape; GenFlow ranked first on the unseen object 6D localization task at BOP Challenge 2023 for both

RGB and RGB-D inputs.

Among training-free approaches, Sundermeyer et al. [10] (ZS6D) established the template matching paradigm using DINOv1 features for zero-shot pose estimation. Query patch descriptors are matched against rendered template features with nearest-neighbour search, followed by RANSAC-PnP for pose recovery. Lin et al. [11] (SAM-6D) extended this paradigm by using SAM for instance segmentation and treating pose estimation as a partial-to-partial 3D point matching problem. It constructs dense 3D-3D correspondences with background tokens to handle clutter, and outperforms prior methods on all seven core BOP datasets. Ornek et al. [12] (FoundPose) further strengthened DINOv2-based template matching with bag-of-words retrieval to shortlist templates. It then applies featuremetric alignment and achieves state-of-the-art performance among RGB-only, training-free methods on BOP benchmarks.

Caraffa et al. [13] (FreeZe) combines DINOv2 visual features with GeDi [14] geometric descriptors. GeDi processes local point neighborhoods transformed into canonical representations via a Local Reference Frame (LRF), achieving rotation and translation invariance that purely visual features lack. FreeZe and its successor FreeZeV2.1 achieved top results on the BOP 2024 model-based unseen object localization track [15], highlighting the complementary roles of visual and geometric features for zero-shot CAD-to-real matching.

2.2 Diffusion Features for Visual Correspondence

The internal representations of generative diffusion models have recently been shown to encode rich semantic and spatial information suitable for visual correspondence without any task-specific training. Tang et al. [2] first demonstrated this by proposing *DIffusion FeaTures* (DIFT): an input image is corrupted via the forward diffusion process at a chosen timestep t, and intermediate U-Net decoder activations are used directly as feature descriptors. Without any fine-tuning, DIFT outperforms DINOv1 and OpenCLIP by 19 and 14 PCK points respectively on the challenging SPair-71k benchmark [16]. A key insight is that the optimal timestep is task-dependent: higher t injects more noise and yields coarser, more semantic features, while lower t preserves finer local appearance detail.

Zhang et al. [3] conducted a systematic analysis of the complementary properties of Stable Diffusion and DINOv2 features for semantic correspondence, observing that SD features produce spatially smooth and dense correspondences while DINOv2 yields sparse but semantically precise matches. They

show that simple normalized concatenation of the two feature types, weighted by a scalar α, achieves gains of +13%, +15%, and +5% on SPair-71k, PF-Pascal, and TSS respectively over either feature alone, with equal weighting α=0.5 found optimal in the real-to-real correspondence setting. A case analysis reveals that in 20–30% of test cases one feature succeeds precisely where the other fails, quantifying the non-redundancy of the two representations. This finding of complementarity in the real-to-real setting is the primary motivation for our investigation of whether the same holds in the synthetic-to-real setting for 6DoF pose estimation.

Dutt et al. [17] (Diff3F) extended diffusion features to 3D shape correspondence by rendering untextured input shapes from multiple viewpoints using depth and normal maps as ControlNet [18] conditioning. Diff3F fuses the resulting diffusion features with DINOv2 features *per rendered view in 2D*—following the fusion strategy of [3]—before lifting and aggregating the combined descriptors onto the 3D surface, demonstrating that multi-view aggregation suppresses view-inconsistency of individual diffusion features to yield stable semantic 3D point descriptors. Their setting is entirely synthetic-to-synthetic, however; neither domain gap nor real sensor noise is encountered, which distinguishes it from the zero-shot pose estimation setting we address.

Wang et al. [19] (FreeReg) applied depth-conditioned ControlNet diffusion features to the problem of image-to-point-cloud registration, observing that the *early* U-Net decoder layers (indices 0, 4, and 6 in their notation) exhibit strong cross-modal consistency between RGB images and depth maps, whereas later layers (indices 5 and 8) encode texture-specific information that breaks this consistency.

Von Gimborn et al. [4] (DZOP) is the most directly related prior work. DZOP replaces DINOv1 features in the ZS6D pipeline with Stable Diffusion hyperfeatures, formed by concatenating activations from decoder layers 2, 5, 8, and 11. It also introduces *feature co-projection*, a joint PCA fitted on the union of query and matched single-view template patch features for each image pair. This aligns the two feature spaces before 2D-2D patch-level correspondence estimation. Co-projection improves AR from 0.149 to 0.427 on LM-O, a 3× gain demonstrating that feature-space misalignment is the dominant bottleneck for diffusion features in zero-shot pose estimation. Our work departs from DZOP in employing 3D point lifting with multi-view feature aggregation rather than single-view 2D patch matching. We show that this architectural choice, while beneficial for geometric stability, prevents co-projection from being effective: aggregation collapses the template feature distribution, eliminating the per-pair variance that co-projection relies upon to align the two feature spaces.

Chapter 3

Background

This chapter introduces two technical foundations used throughout this work: the Kabsch algorithm for optimal rigid body alignment, and the BOP benchmark evaluation protocol that defines the Average Recall metric.

3.1 Kabsch Algorithm

The Kabsch algorithm [20] finds the rotation matrix $\mathbf{R} \in SO(3)$ and translation vector $\mathbf{t} \in \mathbb{R}^3$ that minimize the weighted sum of squared distances between two sets of corresponding 3D points $\{\mathbf{p}_i\}$ and $\{\mathbf{q}_i\}$:

$$\mathcal{L}(\mathbf{R}, \mathbf{t}) = \sum_{i=1}^{N} w_i \left\| \mathbf{R}\,\mathbf{p}_i + \mathbf{t} - \mathbf{q}_i \right\|^2, \tag{3.1}$$

where $w_i \geq 0$ are per-point weights with $\sum_i w_i = 1$. The optimal translation aligns the weighted centroids, $\mathbf{t}^* = \bar{\mathbf{q}} - \mathbf{R}^*\bar{\mathbf{p}}$, and the optimal rotation is obtained via singular value decomposition of the cross-covariance matrix of the centred point sets, $\mathbf{H} = \sum_i w_i\, \tilde{\mathbf{p}}_i\, \tilde{\mathbf{q}}_i^\top = \mathbf{U}\mathbf{S}\mathbf{V}^\top$, yielding

$$\mathbf{R}^* = \mathbf{V}\,\mathbf{D}\,\mathbf{U}^\top, \tag{3.2}$$

where $\mathbf{D} = \mathrm{diag}(1, 1, \det(\mathbf{V}\mathbf{U}^\top))$ corrects for improper rotations that arise when the SVD solution is a reflection rather than a proper rotation.

Kabsch estimation is embedded within a RANSAC loop [21] to handle outlier correspondences. Minimal samples of three point pairs are drawn iteratively to generate pose hypotheses, each scored by counting inliers within a fixed residual threshold. The hypothesis with the most inliers is refined by re-running Kabsch on the full inlier set.

3.2 BOP Benchmark and Average Recall

The BOP benchmark [22, 23] is the standard evaluation framework for 6DoF object pose estimation, providing standardized datasets, ground truth annotations, and pose error functions. Three complementary error functions are defined, each assessing a different aspect of pose quality.

Visible Surface Discrepancy (VSD) measures the per-pixel depth discrepancy between renderings of the object at the estimated and ground truth poses, restricted to the object's visible surface. VSD is invariant to object symmetries since only the visible portion of the surface is evaluated.

Maximum Symmetry-Aware Surface Distance (MSSD) measures the maximum surface-point distance between the model at the estimated and ground truth poses, minimized over the set of indistinguishable views induced by the object's symmetries.

Maximum Symmetry-Aware Projection Distance (MSPD) is the 2D analogue of MSSD, measuring the maximum distance in the image plane between projections of corresponding model points at the estimated and ground truth poses.

For each error function $e \in \{\text{VSD, MSSD, MSPD}\}$, a pose estimate is declared correct if its error falls below a threshold θ. The *Average Recall* (AR) for a given error function is the mean recall across a range of thresholds Θ:

$$\text{AR}_e = \frac{1}{|\Theta|} \sum_{\theta \in \Theta} \frac{|\{(o,i) : e(\hat{\mathbf{P}}_{o,i}, \mathbf{P}^*_{o,i}) < \theta\}|}{|\{(o,i)\}|}, \tag{3.3}$$

where (o, i) indexes object instance o in image i. The final BOP score reported throughout this thesis is the mean of the three per-error AR values:

$$\text{AR} = \tfrac{1}{3}\big(\text{AR}_{\text{VSD}} + \text{AR}_{\text{MSSD}} + \text{AR}_{\text{MSPD}}\big). \tag{3.4}$$

Chapter 4

Methodology

This chapter describes our pipeline for zero-shot 6DoF object pose estimation using fused diffusion and discriminative visual features. The method operates in two stages: an offline onboarding stage that precomputes 3D point descriptors for each target object from its CAD model, and an online inference stage that extracts query features, finds 3D-3D correspondences, and recovers the object pose. Figure 4.1 provides an overview of the full pipeline.

Object detection and segmentation use CNOS [5] as a fixed front-end. CNOS generates mask proposals via the Segment Anything Model [6] and scores each proposal by comparing its DINOv2 CLS-token feature against multi-view CAD renders of the target object. The highest-scoring mask defines the query crop used for feature extraction.

4.1 Feature Extraction

We extract two feature representations from each image: diffusion features from a pretrained Stable Diffusion U-Net, and discriminative features from a pretrained DINOv2 model. The two feature types are complementary for visual correspondence [3], which motivates their fusion.

Diffusion features. Following Tang et al. [2], we extract diffusion features from each input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$. The image is first encoded into the Stable Diffusion VAE latent space [24]:

$$\mathbf{z}_0 = \mathcal{E}(\mathbf{I}), \tag{4.1}$$

where $\mathcal{E}$ denotes the VAE encoder and $\mathbf{z}_0 \in \mathbb{R}^{h \times w \times 4}$ is the clean latent with $h = H/8$, $w = W/8$. Noise is added at a chosen timestep t via the forward

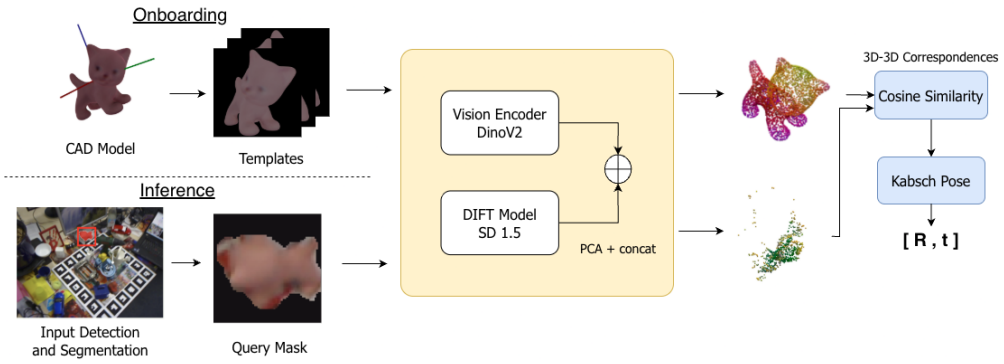


Figure 4.1: Overview of our zero-shot 6DoF pose estimation pipeline. During onboarding, diffusion and DINOv2 features are extracted from multi-view CAD renders and aggregated into 3D point descriptors per object. At inference, query features are extracted from the detected object crop, fused with the corresponding modality weights, and matched against the precomputed object point descriptors via cosine nearest-neighbour search. The resulting 3D-3D correspondences are passed to Kabsch+RANSAC to recover the final 6DoF pose $(\mathbf{R}, \mathbf{t})$.

diffusion process:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{4.2}$$

where $\bar{\alpha}_t$ is the cumulative noise schedule coefficient at timestep t. The noisy latent $\mathbf{z}_t$ is then passed through the denoising U-Net $\mathcal{U}$ conditioned on a null text embedding $\mathbf{c}_\varnothing$, and intermediate activations from selected decoder up-blocks are extracted as feature maps:

$$\mathcal{F}_\ell = \phi_\ell\big(\mathcal{U}(\mathbf{z}_t, t, \mathbf{c}_\varnothing)\big), \quad \ell \in \mathcal{L}, \tag{4.3}$$

where ϕ_ℓ denotes the hook extracting activations from the last ResNet block of decoder layer ℓ and $\mathcal{L} \subseteq \{\ell_2, \ell_5, \ell_8, \ell_{11}\}$ is the set of selected decoder layers. Each feature map $\mathcal{F}_\ell \in \mathbb{R}^{h_\ell \times w_\ell \times C_\ell}$ is bilinearly interpolated to a common spatial resolution $H_f \times W_f$ and the resulting maps are concatenated along the channel dimension to form the final diffusion feature map:

$$\mathcal{F}_{\text{DIFT}} = \big[\uparrow \mathcal{F}_\ell\big]_{\ell \in \mathcal{L}} \in \mathbb{R}^{H_f \times W_f \times D_{\text{DIFT}}}, \tag{4.4}$$

where $D_{\text{DIFT}} = \sum_{\ell \in \mathcal{L}} C_\ell$ and $[\cdot]$ denotes channel-wise concatenation. To reduce variance from the stochastic noise sampling, the process is repeated for K

independent noise realizations and the resulting feature maps are averaged:

$$\bar{\mathcal{F}}_{\text{DIFT}} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{F}_{\text{DIFT}}^{(k)}. \tag{4.5}$$

DINOv2 features. DINOv2 features are extracted by passing the input image through the pretrained DINOv2 ViT encoder [1]. The image is divided into a regular grid of non-overlapping patches of size $P \times P$ pixels, yielding $N_p = (H/P) \times (W/P)$ patch tokens. The last-layer patch token embeddings are used as the feature representation:

$$\mathcal{F}_{\text{DINO}} \in \mathbb{R}^{(H/P)\times(W/P)\times D_{\text{DINO}}}, \tag{4.6}$$

where D_{DINO} is the hidden dimension of the ViT encoder.

4.2 Dimensionality Reduction via PCA

The raw diffusion and DINOv2 feature maps are high-dimensional. We apply Principal Component Analysis (PCA) independently to each feature type to reduce the dimensionality to a common size d. For each model, PCA is fitted on the set of masked template patch features across all objects and views:

$$\mathbf{V}_* = \text{PCA}\big(\{\mathbf{f}_{i,j,p} : \text{mask}_{i,j,p} = 1\}\big), \tag{4.7}$$

where $\mathbf{f}_{i,j,p}$ is the feature vector at patch p of view j of object i, and $\text{mask}_{i,j,p}$ indicates whether the patch lies within the object foreground. The projection matrix $\mathbf{V}_* \in \mathbb{R}^{D\times d}$ comprises the d eigenvectors corresponding to the largest eigenvalues of the feature covariance matrix. At inference, features are projected as:

$$\hat{\mathbf{f}} = (\mathbf{f} - \boldsymbol{\mu})\,\mathbf{V}_*, \tag{4.8}$$

where $\boldsymbol{\mu}$ is the mean computed during fitting. This projection is applied identically to both template and query features.

4.3 3D Point Lifting

Template features are aggregated into 3D point descriptors on the object surface, formulating pose estimation as a 3D-3D correspondence problem.

For each target object with CAD model $\mathcal{M}$, we render T views from camera poses distributed uniformly on the viewing sphere using an icosphere tessel-

lation. Each view $j \in \{1, \ldots, T\}$ yields a rendered RGB image $\mathbf{I}_j^S$, together with the model-to-camera transformation $\mathbf{T}_j \in SE(3)$ and the camera intrinsic matrix $\mathbf{K}^S$.

A set of M surface points $\{\mathbf{x}_m\}_{m=1}^M$ is sampled uniformly from the mesh surface of $\mathcal{M}$. For each template view j, we determine the subset of points visible from that viewpoint by casting rays from the camera origin and checking for self-occlusion via mesh intersection. Let $\mathcal{V}_j \subseteq \{1, \ldots, M\}$ denote the set of visible point indices in view j.

For each visible point $m \in \mathcal{V}_j$, its 3D position $\mathbf{x}_m$ is projected into the image plane of view j:

$$\mathbf{u}_{m,j} = \pi\big(\mathbf{K}^S,\, \mathbf{T}_j,\, \mathbf{x}_m\big), \tag{4.9}$$

where π denotes perspective projection. The feature vector at the (non-integer) pixel location $\mathbf{u}_{m,j}$ is obtained by bilinear interpolation of the PCA-reduced feature map:

$$\mathbf{f}_{m,j} = \mathrm{Interp}\big(\hat{\mathcal{F}}_j,\, \mathbf{u}_{m,j}\big) \in \mathbb{R}^d, \tag{4.10}$$

where $\hat{\mathcal{F}}_j$ is the PCA-projected feature map of template view j. The final 3D point descriptor for point m is the mean of its per-view features, accumulated over all views in which it is visible:

$$\mathbf{e}_m = \frac{1}{|\mathcal{V}(m)|} \sum_{j\,:\,m \in \mathcal{V}_j} \mathbf{f}_{m,j}, \tag{4.11}$$

where $\mathcal{V}(m) = \{j : m \in \mathcal{V}_j\}$ is the set of views in which point m is visible.

4.4 Feature Fusion

Following Zhang et al. [3], we fuse diffusion and DINOv2 features at the level of 3D point descriptors. Let $\mathbf{e}_m^{\mathrm{DIFT}} \in \mathbb{R}^d$ and $\mathbf{e}_m^{\mathrm{DINO}} \in \mathbb{R}^d$ denote the PCA-reduced and aggregated point descriptors from the diffusion and DINOv2 models respectively. The fused descriptor is formed by normalizing each modality independently and concatenating with a scalar weighting parameter $\alpha \in [0, 1]$:

$$\mathbf{e}_m^{\mathrm{fused}} = \left(\alpha\, \frac{\mathbf{e}_m^{\mathrm{DIFT}}}{\|\mathbf{e}_m^{\mathrm{DIFT}}\|_2},\;\; (1-\alpha)\, \frac{\mathbf{e}_m^{\mathrm{DINO}}}{\|\mathbf{e}_m^{\mathrm{DINO}}\|_2} \right) \in \mathbb{R}^{2d}. \tag{4.12}$$

The same fusion is applied to query point descriptors at inference. The weight α controls the relative contribution of each feature type and is treated as a hyperparameter.

4.5 Query Processing and Pose Estimation

At inference time, for each detected object instance the query image crop is processed by both feature extractors to yield feature maps $\hat{\mathcal{F}}^{\text{DIFT}}_{\text{query}}$ and $\hat{\mathcal{F}}^{\text{DINO}}_{\text{query}}$. A set of query surface points is obtained by sampling N pixel locations from within the object mask and back-projecting them into 3D using the depth image and the known camera intrinsics:

$$\mathbf{y}_n = \mathbf{K}^{-1} \begin{pmatrix} u_n \\ v_n \\ 1 \end{pmatrix} d_n, \tag{4.13}$$

where (u_n, v_n) is the sampled pixel and d_n its depth value in millimetres. Feature vectors at each sampled location are obtained by bilinear interpolation of each PCA-reduced feature map, fused following Equation 4.12 to yield query point descriptors $\{\mathbf{q}^{\text{fused}}_n\}_{n=1}^{N}$.

Correspondences between query points and object model points are established by nearest-neighbour search in the fused descriptor space:

$$\hat{m}(n) = \underset{m \in \{1,\dots,M\}}{\arg\min} \left(1 - \langle \mathbf{q}^{\text{fused}}_n, \mathbf{e}^{\text{fused}}_m \rangle\right), \tag{4.14}$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity. The resulting 3D-3D correspondences are passed to the Kabsch+RANSAC estimator from Section 3.1 to recover the final object pose $(\mathbf{R}, \mathbf{t})$.

Chapter 5

Results and Discussion

5.1 Experimental Setup

5.1.1 Datasets and Evaluation Protocol

We evaluate on all seven core datasets of the BOP benchmark [22, 23]: LM-O [25], T-LESS [26], TUD-L [27], IC-BIN [28], ITODD [29], HB [30], and YCB-V [31]. Each dataset provides RGB-D test images and 3D object models, enabling zero-shot evaluation without any object-specific training.

The datasets vary across several axes relevant to appearance-based matching. In terms of object type, T-LESS and ITODD contain industrial objects (electrical and mechanical parts) with planar sides, sharp edges, and no surface texture. LM-O, TUD-L, and HB contain household objects with mostly uniform colours. IC-BIN and YCB-V contain objects with rich surface texture. In terms of scene difficulty, LM-O scenes are highly cluttered with significant inter-object occlusion. TUD-L contains a single object per scene but under challenging illumination. IC-BIN contains multiple instances of the same object per scene, introducing localization ambiguity. The remaining datasets contain mildly occluded objects. This variation across texture and occlusion conditions allows a broad assessment of where appearance-based features are effective and where they are not.

All evaluations use the standard BOP test splits with object detections provided by CNOS [5], following the established zero-shot evaluation protocol [23]. Performance is reported using the BOP Average Recall (AR) metric described in Section 3.2. Due to computational constraints, ablation studies are conducted on LM-O, YCB-V, and T-LESS, which together cover the key axes of texture variation and occlusion level present in the full benchmark.

5.1.2 Implementation Details

The pipeline follows Chapter 4. Diffusion features use Stable Diffusion v1.5 [24] with all four decoder layers, timestep $t = 50$, ensemble size $K = 4$, and 512×512 inputs. DINOv2 uses the ViT-L/14 model with registers [1, 32] at 224×224 resolution. Both feature types are reduced to $d = 256$ by PCA fitted on masked foreground template patches. CAD templates are rendered from $T = 162$ viewpoints, and $M = 5{,}000$ surface points are sampled per object. At inference, $N = 1{,}000$ query points are sampled from the CNOS mask and back-projected using the depth image. Features are fused with $\alpha = 0.5$ and pose is recovered by Kabsch+RANSAC with 1,000 iterations and an inlier threshold of 10% of the mean object diameter.

5.2 Main Results

Table 5.1 compares our method against state-of-the-art zero-shot pose estimation methods on the seven core BOP datasets. We report BOP Average Recall (AR) for each dataset for which results are available.

Table 5.1: Comparison with state-of-the-art zero-shot 6DoF pose estimation methods on the seven core BOP datasets. AR (%) is reported per dataset. ✓ denotes training-free methods. Results for rows 1–7 are taken from the respective published papers. Our method uses CNOS detections for all datasets.

	Method	Train-free	Input	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V	Mean AR
1	MegaPose [7]		RGB	22.9	17.7	25.8	15.2	10.8	25.1	28.1	20.8
2	ZS6D [10]	✓	RGB	29.8	21.0	–	–	–	–	32.4	-
3	GigaPose [8]		RGB	29.9	27.3	30.2	23.1	18.8	34.8	29.0	27.6
4	FoundPose [12]	✓	RGB	39.7	33.8	46.9	23.9	20.4	50.8	45.2	37.3
5	SAM-6D [11]		RGBD	57.0	38.2	69.8	41.5	41.4	66.9	73.2	55.4
6	FreeZe [13]	✓	RGBD	64.7	49.3	86.1	44.3	49.2	75.7	78.7	64.0
7	DZOP [4]	✓	RGB	32.8	26.7	–	–	–	–	36.5	-
8	DINOv2 only	✓	RGBD	36.3	20.7	46.4	24.9	23.5	49.1	54.2	36.4
9	DIFT only	✓	RGBD	27.4	11.9	32.1	19.6	19.4	39.2	48.2	28.3
10	**Fusion (ours)**	✓	RGBD	**42.3**	**23.1**	**54.9**	**31.1**	**27.6**	**61.3**	**62.7**	**43.2**

The fusion method outperforms the DINOv2 baseline across all evaluated datasets. Gains are largest on TUD-L and YCB-V (+8.5 AR), where objects are predominantly textured and occlusion is mild, and somewhat smaller on LM-O (+6.0 AR), which includes textureless objects under heavy occlusion. This is consistent with the finding of Zhang et al. [3] that diffusion

Table 5.2: Effect of fusion weight α on Average Recall (%). α=0 is DINOv2 only; α=1 is DIFT only. Results with $\mathcal{L} = \{\ell_2, \ell_5, \ell_8, \ell_{11}\}$ and t=50.

α	LM-O	YCB-V	T-LESS
0.0 (DINOv2 only)	36.3	54.2	20.7
0.1	37.2	56.1	21.3
0.2	37.5	57.9	23.6
0.3	39.6	59.3	**24.2**
0.4	42.1	62.6	22.8
0.5	**42.3**	**62.7**	23.1
0.6	38.6	59.7	19.7
0.7	34.4	55.2	16.5
0.8	30.8	51.5	13.6
0.9	27.7	49.3	13.2
1.0 (DIFT only)	27.4	48.2	11.9

and discriminative features are complementary, and suggests this holds in the synthetic-to-real setting as well.

Among training-free methods, fusion outperforms FoundPose [12] on LM-O (42.3 vs 39.7) and ZS6D [10] on both LM-O and YCB-V. Fusion also outperforms DZOP [4] on LM-O and YCB-V, the most direct diffusion-feature baseline, which uses diffusion features alone with 2D co-projection rather than 3D point lifting; however on T-LESS specifically, the fused pipeline also falls below (23.1 vs. 26.7 AR). The method remains below FreeZe [13], which uses specialized geometric descriptors with canonical alignment in addition to appearance features.

5.3 Ablation Study

We conduct ablations on three key hyperparameters of our pipeline: the fusion weight α, the diffusion timestep t, and the set of U-Net decoder layers used for feature extraction. All ablations are reported on LM-O, YCB-V, and T-LESS with the default configuration (α=0.5, t=50, $\mathcal{L} = \{\ell_2, \ell_5, \ell_8, \ell_{11}\}$) unless the parameter under study is varied.

5.3.1 Fusion Weight

Table 5.2 reports AR as a function of the fusion weight α, where α=0 corresponds to DINOv2 alone and α=1 to DIFT alone.

Table 5.3: Effect of diffusion timestep t on Average Recall (%). Results with α=0.5 and $\mathcal{L} = \{\ell_2, \ell_5, \ell_8, \ell_{11}\}$.

t	LM-O	YCB-V	T-LESS
0	**43.7**	61.3	20.4
50	42.3	**62.7**	**23.1**
100	40.7	62.6	22.3
150	39.5	62.7	22.3
200	39.7	62.5	22.1
250	39.3	61.9	21.7

Fusion outperforms both individual feature types across all datasets. The optimal weight is α=0.5 on two out of three datasets, giving equal contribution from both features. This matches the finding of Zhang et al. [3] in the real-to-real setting. Gains over DINOv2 alone are +6.0 AR on LM-O, +8.5 AR on YCB-V, and +2.4 AR on T-LESS. Performance decreases as α moves away from 0.5 in either direction, which indicates that both feature types contribute independently.

5.3.2 Diffusion Timestep

Table 5.3 reports AR as a function of the diffusion timestep t used for feature extraction, with α=0.5 and all four layers fixed.

Lower timesteps give better performance. At t=0, no noise is added and the U-Net operates on the clean image latent, giving the best result on LM-O (43.7 AR) but slightly lower performance than t=50 on YCB-V and T-LESS. We use t=50 as the default since it performs best overall. This differs from Tang et al. [2], who find t=261 optimal for semantic correspondence. For object-level synthetic-to-real matching, lower-noise features that retain fine-grained local detail appear more useful than the coarser representations produced at higher timesteps.

5.3.3 Layer Selection

Table 5.4 ablates the set of U-Net decoder up-blocks used for DIFT feature extraction, with α=0.5 and t=50 fixed.

Using all four decoder layers gives the best performance across all three datasets. Among individual layers, ℓ_{11} (finest spatial resolution at 64×64) gives the highest single-layer result, consistent with the finding of Von Gimborn et al. [4] that finer decoder layers are more useful for object-level pose

Table 5.4: Effect of decoder layer selection on Average Recall (%). Results with α=0.5 and t=50.

Layers	LM-O	YCB-V	T-LESS
ℓ_2 only	39.7	58.4	17.8
ℓ_5 only	39.9	60.4	18.2
ℓ_8 only	39.8	60.2	18.3
ℓ_{11} only	40.6	61.7	19.9
ℓ_2, ℓ_5	39.7	59.4	19.5
ℓ_5, ℓ_8	40.8	60.4	20.9
ℓ_8, ℓ_{11}	41.9	62.8	22.0
ℓ_2, ℓ_5, ℓ_8	40.3	60.8	20.8
$\ell_5, \ell_8, \ell_{11}$	42.2	61.9	22.3
$\ell_2, \ell_5, \ell_8, \ell_{11}$	**42.3**	**62.7**	**23.1**

estimation. Among two-layer combinations, ℓ_8 and ℓ_{11} perform best on YCB-V and T-LESS.

5.4 Qualitative Feature Visualization

Figure 5.1 provides a qualitative comparison of the three feature types on five YCB-V objects. For each object, we show: (col. 1) the RGB render; (col. 2) PCA-projected DIFT features; (col. 3) PCA-projected DINOv2 features; and (col. 4) PCA-projected fused features. Colours encode the top-3 PCA components and are computed independently per object.

DIFT features (col. 2) produce spatially smooth colour fields that vary gradually across broad surface regions. On the sugar box, the top, side, and base are clearly separated; on the soup can, a smooth gradient runs from lid to body. DINOv2 features (col. 3) show finer spatial variation: individual limbs of the cat receive distinct colours, and the head, torso, and legs of the minion figure are more clearly differentiated than in the DIFT map. The fused features (col. 4) tend to resemble DINOv2 in local detail while preserving some of the broader regional structure visible in the DIFT maps. On the rabbit figure, the fused map shows richer colour variation across the limbs and body compared to DIFT alone. This is broadly consistent with the quantitative gains in Table 5.1, where fusion improves over both individual feature types.



Figure 5.1: PCA visualization of RGB, DIFT, DINOv2, and fused per-point embeddings respectively for representative YCB-V objects.

Chapter 6

Conclusion

We investigated whether the complementarity between Stable Diffusion and DINOv2 features established for real-to-real correspondence [3] extends to the synthetic-to-real setting of zero-shot pose estimation. We built a training-free pipeline that aggregates both feature types into 3D point descriptors from multi-view CAD renders and fuses them via normalized concatenation. The fused pipeline outperforms DINOv2 alone with gains of up to +8.5 AR on YCB-V and TUD-L and +6.0 AR on LM-O, and outperforms DZOP [4] on LM-O and YCB-V. The optimal fusion weight α=0.5 matches the equal weighting found in the real-to-real setting. Ablations show that low timesteps ($t \leq 50$) and all four decoder layers $\{\ell_2, \ell_5, \ell_8, \ell_{11}\}$ are preferable for object-level matching. This work is a preliminary study into the use of diffusion features for synthetic-to-real pose estimation. The method uses RGB-D input but remains below state-of-the-art RGB-D methods; further investigation is needed to close this gap.

6.1 Limitations

The method relies entirely on appearance features. Textureless and rotationally symmetric objects are not reliably handled, and the gap to FreeZe [13] on T-LESS reflects this directly. The pipeline also does not include a pose refinement stage, which limits absolute performance relative to full systems.

6.2 Future Work

Two directions are particularly promising. First, adding geometric descriptors with canonical alignment such as GeDi [14] as a third feature stream would

directly address the textureless object limitation without modifying the visual feature pipeline.

Second, diffusion models define a natural probability distribution over poses via the denoising process, and this structure has not yet been exploited for pose refinement. One direction is to operate the denoising process directly in $SO(3)$ space: given a coarse pose estimate, a diffusion model conditioned on the observed image could iteratively denoise a noisy pose sample toward the true pose, analogous to how image diffusion models refine noisy latents. This would give a principled, generative approach to pose refinement that does not require task-specific discriminative training.

Bibliography

[1] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[2] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[3] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements DINO for zero-shot semantic correspondence. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[4] Bernd Von Gimborn, Philipp Ausserlechner, Markus Vincze, and Stefan Thalhammer. Diffusion features for zero-shot 6DoF object pose estimation. *arXiv preprint arXiv:2411.16668*, 2024.

[5] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Vincent Lepetit, and Tomas Hodan. CNOS: A strong baseline for CAD-based novel object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 2134–2140, 2023.

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[7] 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 and compare. In *Proceedings of the 6th Conference on Robot Learning*, 2022.

[8] 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*, 2024.

[9] Sungphill Moon, Hyeontae Son, Dongcheol Hur, and Sangwook Kim. GenFlow: Generalizable recurrent flow for 6d pose refinement of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2024.

[10] Martin Sundermeyer, Van Nguyen Nguyen, Markus Durner, and Rudolph Triebel. ZS6D: Zero-shot 6d object pose estimation using vision foundation models. In *IEEE International Conference on Robotics and Automation*, 2024.

[11] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui 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.

[12] Evin Pinar Ornek, Alexander Grabner, Timothy Patten, Rene Steinbrugger, Eric Brachmann, Vincent Lepetit, and Friedrich Fraundorfer. FoundPose: Unseen object pose estimation with foundation features. In *European Conference on Computer Vision*, 2024.

[13] Andrea Caraffa, Davide Boscaini, Amir Hamza, Federico Tombari, Pietro Morerio, and Alessio Del Bue. FreeZe: Training-free zero-shot 6d pose estimation with geometric and vision foundation models. In *European Conference on Computer Vision*, 2024.

[14] 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*, 2022.

[15] Van Nguyen Nguyen, Stephen Tyree, Andrew Guo, Mederic Fourmy, Anas Gouda, Taeyeop Lee, Sungphill Moon, Hyeontae Son, Lukas Ranftl, Jonathan Tremblay, Eric Brachmann, Bertram Drost, Vincent Lep-

etit, Carsten Rother, Stan Birchfield, Jiri Matas, Yann Labbe, Martin Sundermeyer, and Tomas Hodan. Bop challenge 2024 on model-based and model-free 6d object pose estimation, 2025. URL `https://arxiv.org/abs/2504.02812`.

[16] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. SPair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, 2019.

[17] Niladri Shekhar Dutt, Sanjeev Muralikrishnan, and Niloy J. Mitra. Diffusion 3d features (Diff3F): Decorating untextured shapes with distilled semantic features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4494–4504, 2024.

[18] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[19] Haiping Wang, Yuan Liu, Bing Wang, Yujing Sun, Zhen Dong, Wenping Wang, and Bisheng Yang. FreeReg: Image-to-point cloud registration leveraging pretrained diffusion models and monocular depth estimators. In *International Conference on Learning Representations*, 2024.

[20] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976.

[21] 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.

[22] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders Glent Buch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. BOP: Benchmark for 6D object pose estimation. In *European Conference on Computer Vision*, 2018.

[23] Tomas Hodan, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiri Matas. BOP challenge 2020 on 6D object localization. *European Conference on Computer Vision Workshops*, 2020.

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[25] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *European Conference on Computer Vision*, pages 536–551, 2014.

[26] Tomas Hodan, Pavel Haluza, Stepan Obdrzalek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-LESS: An RGB-D dataset for 6d pose estimation of texture-less objects. In *IEEE Winter Conference on Applications of Computer Vision*, 2017.

[27] Tomas Hodan, Vibhav Vineet, Ran Gal, Emanuel Shalev, Jon Hanzelka, Treb Connell, Pedro Urbina, Sudipta N. Sinha, and Juergen Gall. A photorealistic dataset and evaluation framework for 6d pose estimation. In *European Conference on Computer Vision Workshops*, 2018.

[28] Andreas Doumanoglou, Rigas Kouskouridas, Sotiris Malassiotis, and Tae-Kyun Kim. Recovering 6d object pose and predicting next-best-view in the crowd. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[29] Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Görtler, and Stephan Ihrke. Introducing MVTec ITODD: A dataset for 3d object recognition in industry. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017.

[30] Roman Kaskman, Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. HomebrewedDB: RGB-D dataset for 6d pose estimation of 3d objects. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.

[31] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Robotics: Science and Systems*, 2018.

[32] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024.