Object Pose Transformer: Unifying Unseen Object Pose Estimation

Weihang Li[1,2] Lorenzo Garattoni[3] Fabien Despinoy[3] Nassir Navab[1,2]
Benjamin Busam[1,2]
[1]Technical University of Munich [2]Munich Center for Machine Learning [3]Toyota Motor Europe

Abstract

Learning model-free object pose estimation for unseen instances remains a fundamental challenge in 3D vision. Existing methods typically fall into two disjoint paradigms: category-level approaches predict absolute poses in a canonical space but rely on predefined taxonomies, while relative pose methods estimate cross-view transformations but cannot recover single-view absolute pose. In this work, we propose Object Pose Transformer (OPT-Pose), a unified feed-forward framework that bridges these paradigms through task factorization within a single model. OPT-Pose jointly predicts depth, point maps, camera parameters, and normalized object coordinates (NOCS) from RGB inputs, enabling both category-level absolute SA(3) pose and unseen-object relative SE(3) pose. Our approach leverages contrastive object-centric latent embeddings for canonicalization without requiring semantic labels at inference time, and uses point maps as a camera-space representation to enable multi-view relative geometric reasoning. Through cross-frame feature interaction and shared object embeddings, our model leverages relative geometric consistency across views to improve absolute pose estimation, reducing ambiguity in single-view predictions. Furthermore, OPT-Pose is camera-agnostic, learning camera intrinsics on-the-fly and supporting optional depth input for metric-scale recovery, while remaining fully functional in RGB-only settings. Extensive experiments on diverse benchmarks (NOCS, HouseCat6D, Omni6DPose, Toyota-Light) demonstrate state-of-the-art performance in both absolute and relative pose estimation tasks within a single unified architecture.

1. Introduction

Object pose estimation for unseen object instances, without relying on prior CAD models, is fundamental in vision. It unlocks object understanding to enhance robotic manipulation, augmented reality, and autonomous systems. Existing model-free approaches follow two paradigms:

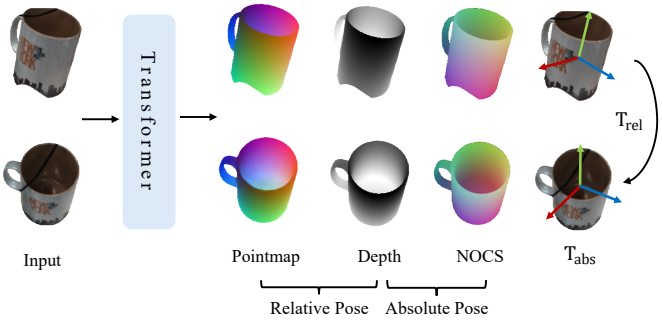

Figure 1. **Unified unseen object pose estimation.** OPT-Pose utilizes a feed-forward transformer to predict point map, depth, NOCS, and camera parameters. Existing category-level methods predict canonical absolute 9-DoF SA(3) poses (equivalent to Depth + NOCS), but require predefined category labels and calibrated cameras. Relative pose methods align unseen objects across views in 6-DoF SE(3) (equivalent to Pointmap + Depth), but do not support single-view absolute pose prediction. OPT-Pose enables the simultaneous recovery of both unseen-object relative and category-level absolute poses (right-most column) for flexible single or multi-view RGB or RGB-D input, without the need for CAD models or semantic labels.

category-level absolute pose estimation predicts canonical-space 9-DoF SA(3) transforms for instances within known categories [5, 12, 27–29, 31–33, 44, 49, 53] but relies on predefined taxonomies and category labels; *relative pose estimation* aligns unseen objects across views via 6-DoF SE(3) [6, 7, 15, 20, 21, 38, 40] but lacks canonicalization and cannot handle single-view absolute pose. However, both paradigms remain constrained. Category-level methods require explicit category names at inference [5, 17, 28, 29, 31, 34, 66, 67], limiting their generalization to open-vocabulary conditions. Additionally, relative methods typically require multiple views and cannot handle single-view absolute pose. To the best of our knowledge, no prior work unifies these complementary tasks in a single category-agnostic model while leveraging their interplay to improve pose estimation and generalization to unseen objects.

We propose Object Pose Transformer (OPT-Pose), a unified feed-forward framework for *model-free unseen object pose estimation with task factorization*. OPT-Pose unifies

Project Page: https://colin-de.github.io/OPT-Pose/

category-level absolute pose and unseen-object relative pose in a single model by predicting depth, point maps, camera parameters alongside NOCS from RGB images. The core design insight is a complementary geometric mechanism:

- **Canonical-space grounding.** Depth + NOCS align instances into a shared canonical space, enabling absolute SA(3) pose estimation without requiring category labels.
- **Relative geometric reasoning.** Depth + point maps represent objects in camera space, enabling multi-view SE(3) reasoning across frames that provides additional geometric constraints and improves absolute pose estimation.

This factorization bridges camera- and canonical-space reasoning without CAD models or predefined taxonomies.

OPT-Pose employs a multi-view transformer that aggregates image tokens and dispatches them to lightweight task heads. A keypoint-centric attention module builds soft correspondences over sampled pixels, while a visual–geometric fusion block integrates local 3D neighborhoods with global context to produce discriminative keypoint descriptors. These descriptors are pooled into an *object latent embedding* and used to FiLM-condition the NOCS head [41]. We train this latent representation with a contrastive InfoNCE objective across views [51], enabling a shared canonical space without requiring semantic labels at inference time. Unlike methods that scale to hundreds of categories but still depend on predefined taxonomies, OPT-Pose is category-agnostic and treats all objects uniformly. A dedicated camera head estimates intrinsic parameters, enabling camera-agnostic operation. A metric-recovery head aggregates keypoint-level depth evidence when measured depth is available, enabling metric-scale recovery in RGB-D mode while remaining fully functional in RGB-only settings.

Extensive experiments across diverse datasets and tasks, including category-level absolute pose (REAL275, HouseCat6D, Omni6DPose [18, 53, 66]) and unseen-object relative pose (REAL275, Toyota-Light)[16], show that OPT-Pose achieves state-of-the-art performance on both absolute and relative pose estimation, generalizing across object categories, camera types, and input modalities (RGB/-D).

We summarize our contributions as follows:

- **Unified Object Poses.** A unified model-free framework for category-level absolute SA(3) and unseen-object relative SE(3) pose via complementary geometric mechanisms. We leverage multi-view relative geometric reasoning to improve absolute pose estimation, without CAD model.
- **Category-Agnostic Canonicalization.** We learn a category-agnostic canonicalization for inference via a contrastive objective without class labels.
- **Flexible inputs and outputs.** OPT-Pose supports RGB and RGB-D inputs from single or multiple views, achieving state-of-the-art performance across both category-level absolute and relative unseen-object pose tasks.

2. Related Works

Category-level, Model-free Absolute Pose Estimation. Category-level methods lift objects into a canonical space using normalized object coordinate systems (NOCS) [53] and regress per-pixel correspondences or keypoints [28, 32] to recover 9D poses. While recent category-level approaches focus on improving accuracy in pose prediction using shape or semantic priors [3, 4, 8, 29, 32, 68, 71] or extending the existing category to large-vocabulary methods [2, 19, 66, 69], they still require predefined taxonomies and explicit category names at inference, limiting true open-vocabulary deployment. Many existing methods with high accuracy, like GCE-Pose[28] and AG-Pose[32], further design specialized networks with explicit categorical shape and semantic priors [5, 28, 30, 32, 55]. These constraints prevent generic, category-agnostic operations.

Feed-forward Geometry Transformers. Recent feed-forward vision transformers (e.g., DUSt3R [57], MASt3R [26], CUT3R [56], VGGT [56]) jointly predict geometric signals such as depth, camera parameters, and point maps in a single forward pass. They decouple representation learning from task-specific heads, enabling multi-view token aggregation and camera self-calibration. OPT-Pose extends this paradigm to object-centric perception by additionally learning keypoints, object-centric latent embeddings, and latent-conditioned NOCS, bridging camera- and canonical-space reasoning for joint absolute and relative pose estimation.

Relative Pose Estimation and Point Cloud Registration. Two-view relative pose estimation is commonly solved via feature matching and post optimization (e.g., essential matrix, PnP), or learned 2D/3D matching and registration [11, 15, 21, 25, 38, 47, 59]. For model-free methods, (H)Oryon [6, 7] establish cross-view correspondences via feature matching and point cloud registration. Any6D and OnePoseViaGen [13, 24] instead generate object meshes using image-to-3D diffusion [36, 60, 62, 63] and align them via render-and-compare. In contrast, we predict object-centric point maps and leverage measured depth to directly estimate SE(3) alignment with weighted Umeyama [50].

Metric Scale Recovery. Monocular predictions are inherently scale-ambiguous. Previous work leveraged category-level size information and regressed offsets to real sizes [9, 10, 17, 22, 58, 70]. In our setting, the optional sensor depth acts as an external signal. OPT-Pose predicts absolute translation/size from depth-derived points and derives the scale by comparing against normalized predictions. Additionally, in RGB-only mode, a lightweight head estimates per-frame log-scale.

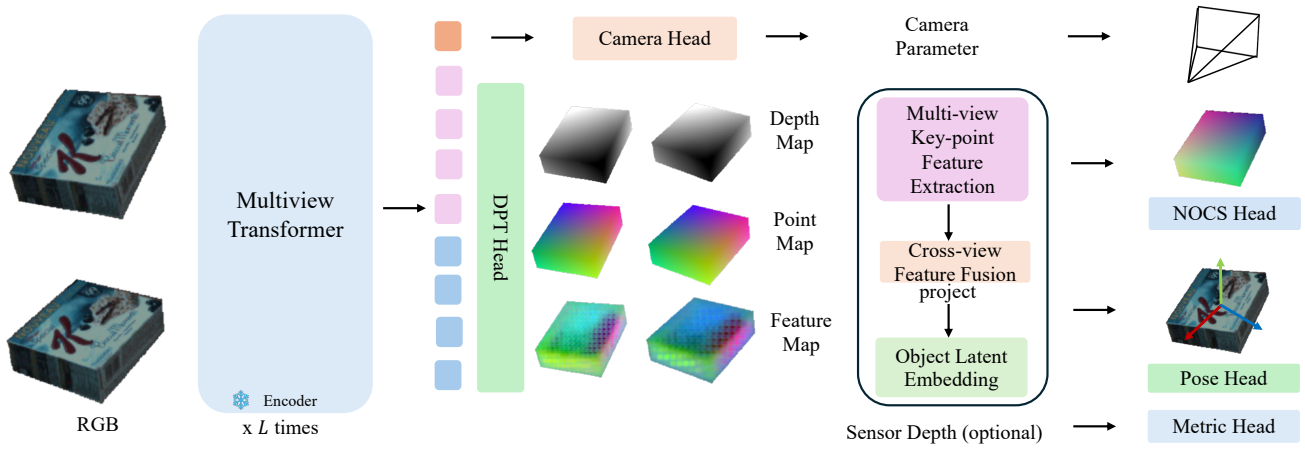

Figure 2. OPT-Pose overview. A multiview transformer aggregates image tokens and emits predictions from light heads: camera parameters, depth, and point maps for camera-space geometry; a multi-view-keypoint-centric module fuses RGB and 3D features to discover object keypoints, predict NOCS coordinates, and build an object latent embedding. Absolute pose (SA(3)) and relative pose (SE(3)) are recovered in a single forward pass. Optional sensor depth provides metric scale, while the system remains fully functional in RGB-only mode.

Semantic Priors and Category-agnostic Canonicalization. Vision and language foundation models [39, 43] provide rich semantic features, but existing category-level methods [5, 28, 32, 67] still require category labels at inference. In contrast, OPT-Pose projects object-centric latent embeddings, extracted from keypoint-level visual features [54] and geometric representations from Sonata [61], into a shared object-latent space. We train this space using a contrastive objective across views, enabling category-agnostic canonicalization without requiring category names at inference.

Taxonomy of Unseen Pose Estimation and OPT-Pose. Model-free object pose estimation falls into two paradigms: (i) *Category-level absolute pose estimation* (e.g., GCE-Pose, AG-Pose) [28, 32], which predicts SA(3) transforms in a canonical space but relies on predefined taxonomies; and (ii) *Unseen-object relative pose estimation* (e.g., OnePose++, MegaPose, OSOP) [21, 47, 48], which estimates SE(3) across views but cannot recover single-view absolute pose. OPT-Pose unifies these paradigms in a model-free framework via task factorization, jointly predicting canonical-space absolute pose and camera-space relative pose through complementary geometric mechanisms. This formulation removes the need for CAD models, category labels, and test-time optimization. In addition, OPT-Pose achieves category-agnostic generalization through contrastive latent learning, while remaining camera-agnostic and supporting flexible RGB(-D) inputs across single and multi-view settings.

3. Method

3.1. Design and Task Factorization

We address model-free, unseen-object pose estimation, covering category-level absolute and relative pose estimation between views, through a geometric observation: two complementary correspondence pairs are sufficient to link canonical and camera spaces.

1. **Depth + NOCS $\Rightarrow$ category-level absolute $\mathrm{SA}(3)$ pose.** Depth provides metric 3D observations in camera space, whereas NOCS yields canonical correspondences; aligning these recovers rotation, translation, and scales for category-level reasoning.
2. **Depth + Point Map $\Rightarrow$ unseen-object relative $\mathrm{SE}(3)$ pose.** The two 3D representations enable robust cross-frame alignment without explicit canonicalization.

This factorization naturally supports single- and multi-view cases with RGB-(D), and is category- and camera-agnostic.

3.2. Problem Formulation

Given a sequence of RGB frames $\{I_i\}_{i=1}^{S}$, we predict per-frame object geometry and poses, without CAD prior. Let $\mathbf{K}_i$ denote camera intrinsics, $\mathbf{T}_i \in \mathrm{SE}(3)$ the camera-to-world extrinsics (implicitly represented by our pose encoding), and $\mathbf{P}_i \in \mathbb{R}^{H\times W\times 3}$ the point map (i.e., the 3D camera-space coordinates). Let further $\mathcal{I}_i$ be a set of K sampled pixels and $\mathbf{X}_{i,k}^{\text{obj}} \in \mathbb{R}^3$ their camera-space 3D points.

In a canonical object space (NOCS), object coordinates lie in $[-0.5, 0.5]^3$. For each frame, we predict M keypoints with canonical coordinates $\mathbf{C}_{i,m} \in \mathbb{R}^3$, and corresponding 3D observations $\mathbf{X}_{i,m}^{\text{obj}} \in \mathbb{R}^3$ in the first (anchor) camera space. The (absolute) category-level object pose can be retrieved by aligning the canonical to the observed coordinates via transformation in the rigid anisotropic similarity transformation

$$\mathrm{SA}(3) := \mathbb{R}^3 \times SO(3) \times \mathrm{Diag}^+(3) \quad (1)$$

$$\text{with} \quad \mathbf{X}^{\text{obj}}_{i,m} = \mathbf{R}_i \mathbf{S} \mathbf{C}_{i,m} + \mathbf{t}_i \quad (2)$$

and $\mathbf{S} = \mathrm{diag}(s_i)_{i=1}^3$ with $\mathrm{SE}(3) \subset \mathrm{SA}(3)$ for $\mathbf{S} = \mathbf{I}$ and $\mathrm{Sim}(3) \subset \mathrm{SA}(3)$ for $\mathbf{S} = s\mathbf{I},\ s > 0$. We also estimate a relative pose $\Delta\mathbf{T}_i \in \mathrm{SE}(3)$ aligning the two geometry branches (i.e., depth-derived vs. point-map predictions).

3.3. Multiview Geometry and Feature Transformer

Our model is a single feed-forward multiview transformer that produces all geometric outputs jointly. (see. Fig. 2). The aggregator encodes each image into a sequence of tokens across several refinement iterations, using a visual backbone [39, 54] with frozen patch embedding for stable training. From these tokens, the camera head predicts per-frame intrinsics through a field-of-view representation and extrinsics as quaternions. The depth head estimates dense depth $\hat{D}_i$ and confidence, which we convert to camera-space points and normals; at sampled indices $\mathcal{I}_i$, we gather points $\mathbf{X}^{\text{obj}}_{i,k}$ together with colors and normals for keypoint reasoning. A parallel point-map head predicts a dense 3D point map $\hat{\mathbf{P}}_i$, and we extract $\mathbf{X}^{\text{pm}}_{i,k}$ as an independent structural cue. A canonicalization head predicts keypoint-level NOCS coordinates $\hat{\mathbf{C}}_{i,m}$ based on object-centric features fused with a global latent embedding $\mathbf{z}_{obj}$. Using these canonical coordinates and the observed keypoints, the pose head estimates $(\mathbf{R}, \mathbf{t}, s)$, while relative $\mathrm{SE}(3)$ is computed after inference through a weighted Umeyama solver [50]. To recover real-world scale, the model supports a relative-scale head that infers scale from RGB features and the object latent z_{obj}, and an absolute-scale head that uses sensor-depth point clouds to predict translation and object size in the camera frame if available. This unified design provides camera parameters, depth maps, point maps, canonical coordinates, and both absolute and relative pose within one coherent framework.

For the geometric supervision of camera extrinsics, point maps, and depth maps in normalized space, we follow [54]. The depth loss follows the aleatoric-uncertainty formulation and uses the predicted uncertainty map Σ^D_i to weight both the depth residual and the spatial gradient residual. The loss is $\mathcal{L}_{\text{depth}} = \sum_{i=1}^{N} \left(\|\Sigma^D_i \odot (\hat{D}_i - D_i)\| + \|\Sigma^D_i \odot (\nabla\hat{D}_i - \nabla D_i)\| - \alpha \log \Sigma^D_i \right)$ where $\odot$ denotes channel-broadcast element-wise multiplication. The point-map loss uses the same structure, but with the point-map uncertainty Σ^P_i: $\mathcal{L}_{\text{point}} = \sum_{i=1}^{N} \left(\|\Sigma^P_i \odot (\hat{P}_i - P_i)\| + \|\Sigma^P_i \odot (\nabla\hat{P}_i - \nabla P_i)\| - \alpha \log \Sigma^P_i \right)$.

3.4. Keypoint-level Multi-view Feature Fusion

Direct dense pixel-level NOCS regression with attention [52] is expensive and sensitive to noise. Following keypoint-based formulations [28, 32], we represent each object by a compact set of M latent keypoints that attend to joint visual and geometric evidence. Given a foreground RGB-D crop, we sample N pixels and lift them to camera-space points $\mathbf{X}_k$ with associated colors and normals. A transformer backbone extracts image features $\mathbf{f}^{\text{rgb}}_k$ at the sampled tokens, while a 3D backbone [61] processes $(\mathbf{X}_k, \mathbf{I}_k, \mathbf{n}_k)$ to produce geometric features $\mathbf{f}^{\text{geo}}_k$. Concatenation yields local descriptors $\mathbf{f}_k = [\mathbf{f}^{\text{rgb}}_k \| \mathbf{f}^{\text{geo}}_k]$. A learnable query performs cross-attention using cosine similarity, producing soft heatmaps $\mathbf{H}_{m,k}$. Each keypoint is $\mathbf{X}^{\text{obj}}_m = \mathbf{H}_{m,k}\mathbf{X}^{\text{obj}}_k$, keypoint feature is extracted as $\mathbf{F}^{\text{obj}}_m = \mathbf{H}_{m,k}[\mathbf{f}^{\text{rgb}}_k \| \mathbf{f}^{\text{geo}}_k]$. A visual-geometric fusion block [32] then refines $\mathbf{F}^{\text{obj}}_m$ by performing KNN grouping in 3D around $\mathbf{X}^{\text{obj}}_m$, encoding relative offsets and absolute coordinates, and applying cosine-similarity attention over the local neighborhoods. This aggregation yields enhanced keypoint descriptors $\tilde{\mathbf{F}}^{\text{obj}}_m$ that carry both local geometric context and global object cues.

We further aggregate keypoint features across frames using a cross-frame attention module. Concretely, we augment keypoint descriptors with frame-wise sinusoidal positional encodings and process them with a transformer encoder, allowing keypoints from different views to exchange information at the feature level. In parallel, we pool the object latent embedding across the input views and share it back to each frame. This design enables multi-view geometric reasoning, enforces cross-view consistency, and improves absolute pose estimation by reducing single-view ambiguity. To ensure geometric consistency, we constrain $\mathbf{X}^{\text{obj}}_m$ to lie on the object surface using a Chamfer distance loss $\mathcal{L}_{\text{cd}}$.

$$\mathcal{L}_{\text{cd}} = \frac{1}{|\mathbf{X}^{\text{obj}}_{\text{m}}|} \sum_{x \in \mathbf{X}^{\text{obj}}_{\text{m}}} \min_{y \in \mathbf{X}^{\text{obj}\star}_k} \|x - y\|_2^2. \quad (3)$$

To prevent keypoints from collapsing into a small region, we add a diversity regularization that balances surface adherence and spatial diversity

$$\mathcal{L}_{\text{div}} = \frac{1}{M(M-1)} \sum_{x \neq y \in \mathbf{X}^{\text{obj}}_{\text{m}}} \max\left(0, \tau_2 - \|x - y\|_2\right)^2, \quad (4)$$

where τ_2 controls the minimum separation between keypoints. To encourage keypoints to be representative of the depth-lifted point cloud, we employ a lightweight reconstruction head that takes keypoint positions and features $\hat{\mathbf{F}}^{\text{obj}}_m$ as input, applies positional encoding, and decodes per-point displacement deltas to recover the object geometry. The reconstruction loss is a one-sided Chamfer distance between the reconstructed point cloud $\hat{\mathbf{X}}^{\text{obj}}$ and the observed camera-space points $\mathbf{X}^{\text{obj}}_k$:

$$\mathcal{L}_{\text{rec}} = \frac{1}{\left|\hat{\mathbf{X}}^{\text{obj}}\right|} \sum_{x \in \hat{\mathbf{X}}^{\text{obj}}} \min_{y \in \mathbf{X}^{\text{obj}\star}_k} \|x - y\|_2 . \quad (5)$$

The keypoint regularization is $\mathcal{L}_{\text{kpt}} = \mathcal{L}_{\text{cd}} + \mathcal{L}_{\text{div}} + \mathcal{L}_{\text{rec}}$.

3.5. Canonical Correspondences & Absolute Poses

Given latent keypoint features, the NOCS head predicts canonical coordinates for each keypoint as $\hat{\mathbf{C}}_m = \mathrm{NOCS}(\tilde{\mathbf{F}}_m^{\text{obj}}, \mathbf{z}_{\text{obj}})$. NOCS regression is addtionally conditioned on FiLM based affine transformation with parameters (γ, β) to intermediate features, so that canonicalization is conditioned on the object code $\mathbf{z}_{\text{obj}}$ but remains category-agnostic. Absolute pose is estimated by an MLP-based pose and size head that takes $(\hat{\mathbf{C}}_m, \mathbf{X}_m^{\text{obj}})$ and the corresponding keypoint features $\tilde{\mathbf{F}}_m^{\text{obj}}$ as input. Rotation is represented in 6D [72] and mapped to $\mathrm{SO}(3)$ via orthogonalization, while translation is predicted as a residual with respect to the point-cloud center following [29, 31]. An isotropic scale $\hat{s}$ is obtained from a per-axis size vector $\hat{\mathbf{s}}$ by averaging its magnitudes. The resulting homogeneous transform $\hat{\mathbf{S}} = [\hat{s}\hat{\mathbf{R}}, \hat{\mathbf{t}}]$ maps canonical coordinates to the camera frame. To supervise the normalized scale prediction, we use

$$\mathcal{L}_{\text{pose}} = \|\mathbf{R}_{\text{gt}} - \mathbf{R}\|_{\text{F}} + \|\mathbf{t}_{\text{gt}} - \mathbf{t}\|_2 + \|\mathbf{s}_{\text{gt}} - \mathbf{s}\|_2. \quad (6)$$

For $\mathcal{L}_{\text{nocs}}$, we use the Smooth L_1 loss with

$$\mathcal{L}_{\text{nocs}} = \|\mathbf{C}_{\text{m}}^{\text{gt}} - \mathbf{C}_{\text{m}}^{\text{nocs}}\|_{\text{SL1}} \quad (7)$$

3.6. Relative Poses from Depth and Point Map

We estimate the metric relative pose by aligning two independently predicted 3D structures with a robust, weighted Procrustes/Umeyama procedure. Let (a, q) denote the anchor and query frames. Our model predicts two-view point maps $\mathbf{P}^{\mathrm{a}}$ and $\mathbf{P}^{\mathrm{q}}$ in the anchor coordinate system, while depth and intrinsics yield camera-space point clouds $\mathbf{X}_{\text{cam}}^{\mathrm{a}}$ and $\mathbf{X}_{\text{cam}}^{\mathrm{q}}$. We proceed in two steps:

1. **Anchor calibration (Sim(3)).** We compute a weighted Umeyama similarity transform $\mathbf{S}_{\mathrm{a}} \in \mathrm{Sim}(3)$ that aligns the predicted anchor point map to the depth-derived anchor camera points,

$$\mathbf{S}_{\mathrm{a}} = \underset{\mathbf{S}\in\mathrm{Sim}(3)}{\operatorname{argmin}} \sum_n w_n \left\| \mathbf{S}\,\mathbf{P}_n^{\mathrm{a}} - \mathbf{X}_{\text{cam},n}^{\mathrm{a}} \right\|_2^2 \quad (8)$$

   where weights w_n come from point map confidences. We then apply $\mathbf{S}_{\mathrm{a}}$ to both $\mathbf{P}^{\mathrm{a}}$ and $\mathbf{P}^{\mathrm{q}}$, removing global scale ambiguity due to projective geometry ambiguity for uncalibrated camera.
2. **Query alignment (SE(3)).** We then align the calibrated query point map $\mathbf{S}_{\mathrm{a}}\mathbf{P}^{\mathrm{q}}$ to the query camera-space points with a *fixed-scale* Umeyama (i.e., $\mathrm{SE}(3)$),

$$\mathbf{T}^{\mathrm{a}\to\mathrm{q}} = \underset{\mathbf{T}\in\mathrm{SE}(3)}{\operatorname{argmin}} \sum_n \tilde{w}_n \left\| \mathbf{T}\,(\mathbf{S}_{\mathrm{a}}\mathbf{P}_n^{\mathrm{q}}) - \mathbf{X}_{\text{cam},n}^{\mathrm{q}} \right\|_2^2. \quad (9)$$

The resulting $\mathbf{T}^{\mathrm{a}\to\mathrm{q}}$ is the relative pose from anchor to query. Given an absolute pose $\mathbf{T}^{\mathrm{a}}$ for the anchor, the query absolute pose is recovered as $\mathbf{T}^{\mathrm{q}} = \mathbf{T}^{\mathrm{a}\to\mathrm{q}}\mathbf{T}^{\mathrm{a}}$. This two-step formulation (Sim(3) followed by fixed-scale $\mathrm{SE}(3)$) is robust to monocular scale ambiguity.

3.7. Contrastive Learning for Object Latent

We train the object latent embedding $\mathbf{z}_{obj}$ using a supervised InfoNCE objective to enhance the alignment of samples belonging to the same instance or semantic category. Let $\{z_i\}_{i=1}^N$ denote the normalized latent features $\mathbf{z}_{obj}$ in a batch, the pairwise similarity is defined as $s_{ij} = \frac{z_i^\top z_j}{\tau}$, where τ is a fixed temperature. A binary mask $P \in \{0,1\}^{N\times N}$ specifies the positive pairs. The diagonal entries of P are zero. The denominator for each anchor i includes all non-diagonal terms $d_i = \log\left(\sum_{j\neq i}\exp(s_{ij})\right)$. Let w be a weight matrix aligned with P, and let ϵ be a small constant used for stability. The numerator in the log-sum mode is $n_i = \log(\sum_{j:P_{ij}=1}\exp(s_{ij} + \log(w_{ij} + \epsilon)))$. The final supervised InfoNCE loss is computed as

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}(n_i - d_i), \quad (10)$$

where $\mathcal{V} = \{i \mid \sum_j P_{ij} > 0\}$ denotes the anchors that contain at least one positive match. Anchors without any positive pairs are excluded from the average.

During distributed training, we gather all latent features across devices before computing the loss, which enlarges the set of negatives and keeps the loss consistent across ranks. The positive mask and weight matrix are expanded to match the aggregated latent features. This produces a stable latent space where samples of the same object or category become closer while unrelated samples remain separated.

3.8. Metric Scale Recovery & Camera Parameters

Monocular predictions are scale-ambiguous. With sensor depth D^{sens}, we sample K points at $\mathcal{I}_i$ and form a centered point cloud $\{\tilde{\mathbf{x}}_k\}$. An absolute-scale head encodes $\{\tilde{\mathbf{x}}_k\}$ with PointNet++ [42], fuses the features with a projected $\mathbf{z}_{\text{obj}}$, and predicts *absolute* camera-frame translation $\hat{\mathbf{t}}^{\text{abs}}$ and size $\hat{\mathbf{s}}^{\text{abs}}$ and supervise the translation ans size in absolute scale using L1 loss. When depth is absent, a lightweight head regresses a per-frame log-scale from global visual descriptors fused with $\mathbf{z}_{\text{obj}}$ and outputs a confidence. This branch is auxiliary during training and can be used at inference to provide a scale prior in pure RGB mode. We supervise these with an L1 loss $\mathcal{L}_{\text{scale}} = \|\hat{\mathbf{t}}^{\text{abs}} - \mathbf{t}^{\text{abs}}\|_1 + \|\hat{\mathbf{s}}^{\text{abs}} - \mathbf{s}^{\text{abs}}\|_1 + \|\log\hat{s} - \log s^*\|_1$.

The camera head predicts a compact pose encoding per frame that decodes to intrinsics and supports differentiable back-projection. Training uses an ℓ_1 loss on focal lengths and on the extrinsic represented by a quaternion, we supervise the pose encoding through a Huber distance between the predicted camera parameters $\hat{\mathbf{g}}_i$ and the ground truth $\mathbf{g}_i$: $\mathcal{L}_{\text{cam}} = \sum_{i=1}^N \|\hat{\mathbf{g}}_i - \mathbf{g}_i\|_\epsilon$.

Table 1. Category-level pose estimation (RGB) on REAL275 using scale-agnostic metrics.

Method	NIoU75	5°0.2d	5°0.5d	10°0.2d	10°0.5d	0.2d	5°	10°
MSOS [22]	0.7	-	-	3.3	15.3	10.6	-	17.0
OLD-Net [9]	0.4	0.9	3.0	5.0	16.0	12.4	4.2	20.9
DMSR [58]	9.5	15.1	23.7	25.6	45.2	35.0	27.4	52.0
LaPose [70]	15.8	15.7	21.3	37.4	57.4	46.9	23.4	60.7
GIVE-Pose [17]	**20.8**	-	-	**44.6**	64.8	46.9	-	67.8
UniDet [10]	19.2	**25.1**	31.8	43.7	66.1	53.5	32.1	68.8
OPT-Pose	11.1	24.3	**52.2**	44.1	**82.3**	**56.4**	**52.3**	**82.5**

3.9. Overall Loss Function

We supervise multi-head predictions with a combination of losses:

$$\mathcal{L} = \mathcal{L}_{\text{cam}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{point}} + \mathcal{L}_{\text{nocs}} + \mathcal{L}_{\text{kpt}} + \mathcal{L}_{\text{pose}} + \mathcal{L}_{\text{scale}} + \mathcal{L}_{\text{InfoNCE}} \tag{11}$$

We provide detailed settings in the supplementary material.

4. Experiment

4.1. Implementation Details

We follow the image configuration from [54]. Images are resized to 518×518 with a patch size 14. For each frame we sample K=1024 pixels to obtain indices $\mathcal{I}_i$ used by all heads. We predict M=128 keypoints. We set the softmax temperatures in keypoint attention and contrastive InfoNCE to 1.0, use a repulsion threshold 0.02 in the keypoint diversity loss, pose weight of ϵ=10^{-3} for absolute pose, Smooth-ℓ_1 with threshold 0.1 for sparse NOCS and relative-scale supervision (with β=0.1), and $k_n \in \{8, 16, 32\}$ neighbors per stage in the fusion block. The multiview transformer encoder is frozen, leaving the attention to fine-tune the features for canonicalization. The model is optimized with AdamW and parameter groups: NOCS/pose/fusion modules and projectors use a base learning rate of 5×10^{-4} with weight decay of 1×10^{-2}; the global optimizer uses a learning rate of 5×10^{-7} with weight decay of 0.05. We apply 5% linear warm-up then cosine decay, mixed precision, and per-module gradient clipping. We provide more parameter details in supplementary material.

4.2. Benchmarks and Protocols

We evaluate on three tasks to demonstrate unified categorical absolute and unseen-object relative pose in a single framework, with flexible RGB-(D) input and intrinsics:

- **Category-level absolute pose (RGB-D): HouseCat6D** [18]. We report metrics $(5°, 2\text{cm})$, $(5°, 5\text{cm})$, $(10°, 2\text{cm})$, $(10°, 5\text{cm})$ thresholds and 3D IoU. For **ROPE** [66], we report: **VUS** (Volume Under Surface) with rotation thresholds from $1°$ to $15°$ and translation thresholds from $1\,\text{cm}$ to $5\,\text{cm}$, and **AUC** (Area Under Curve), which evaluates Intersection over Union (IoU) of 3D bounding boxes over IoU thresholds from 0.25 to 0.95.
- **Category-level absolute pose (RGB; scale-agnostic on REAL275):** Following prior work, we report normalized IoU ($NIoU$) and distance thresholds on REAL275 in RGB-only settings for scale-agnostic pose estimation.
- **Unseen-object relative pose (RGB-D): NOCS-REAL, Toyota-Light (TOYL)**. Trained on SOPE [66], these benchmarks test SE(3) alignment across views for unseen objects. We evaluate using ADD(-S), AR, MSSD, MSPD, and VSD metrics; relative SE(3) is estimated post-hoc via weighted Umeyama alignment of depth and point-map structures (Sec. 3.6).

4.3. Comparison with the State of the Art

Category-level absolute pose (RGB; scale-agnostic REAL275). As shown in Tab. 1, our method substantially outperforms prior work on REAL275 in scale-agnostic evaluation. We achieve great performance on all reported metrics except $NIoU_{75}$, where we are comparable to UniDet[10] and surpass GIVEPose [17], DMSR[58], LaPose[70] by in-degree/normalized distance thresholds. Note that these methods explicitly use the predefined calculated size and inference with category priors.

Category-level absolute pose (RGB-D). Leveraging measured depth, OPT-Pose recovers metric scale via the absolute-scale head and achieves strong performance on HouseCat6D [18] (See Fig. 4). In Tab. 2, we obtain the best results under strict thresholds ($5°$2cm, $5°$5cm) and competitive accuracy at looser thresholds and IoU compared to GCE-Pose[28], while using no category, shape, semantic, or calibration priors. In addition, we evaluate a multi-view enhanced inference mode using S=2, 3, 4 frames without retraining. As shown in Tab. 2, multi-view inference consistently improves absolute pose accuracy across all metrics. This supports our central design: relative geometric reasoning across views provides additional constraints that reduce ambiguity in single-view predictions, leading to more stable and accurate absolute pose estimation. To validate the practical utility of our unified framework for real-world downstream tasks, we evaluate OPT-Pose on the large-vocabulary Omni6DPose benchmark, designed for robotic manipulation. As shown in Tab. 3, OPT-Pose surpasses recent state-of-the-art methods like GenPose++ [66] and CPPF++ [64] in strict IoU metrics and achieves highly competitive accuracy at $5°/10°$ thresholds while using less prior and performing inference with a category-agnostic setting. This demonstrates that our category-agnostic canonicalization effectively scales to large-vocabulary scenarios, which are critical for robotics.

Unseen-object relative pose. For unseen objects pose estimation, OPT-Pose aligns depth and point-map branches with a weighted Umeyama to yield robust SE(3) estimation

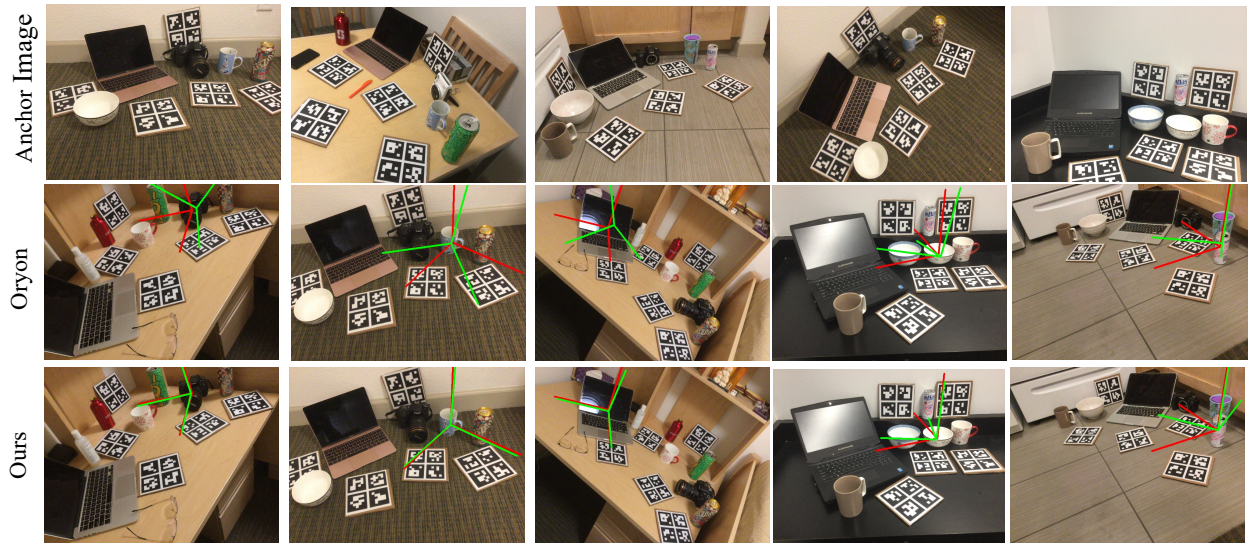

Figure 3. Qualitative result in relative pose estimation. We compare different object instances across different scenes with Oryon [7]. Visualization shows that our OPT-pose can estimate the relative object poses across different objects and scenes.

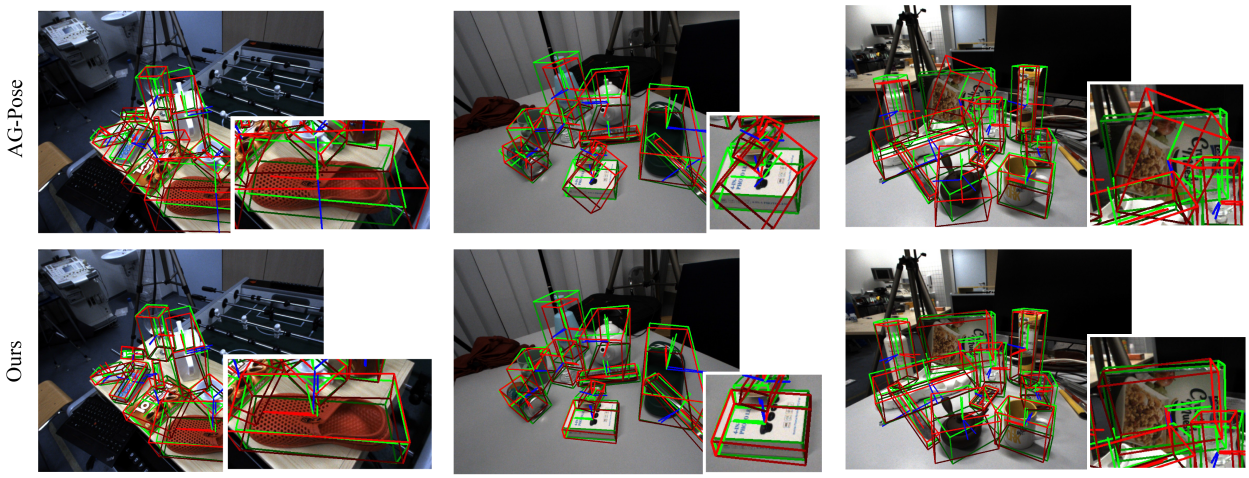

Figure 4. Qualitative result in absolute pose estimation with RGB-D input. We showcase some difficult instances in comparison with AG-Pose[32]. We zoom in on the difficult object categories, the shoe and the box, for better visualization.

across frames. As shown in Tab. 4, our method outperforms all existing methods by large margins across ADD(-S), AR, MSSD, MSPD, and VSD, especially on NOCS-REAL (Fig. 3), demonstrating that task factorization within a single model enables strong performance on both canonical-space (absolute) and camera-space (relative) reasoning. The slight performance gain in TOY-L is due to our method being more sensitive to illumination changes, and to geometric alignment being less accurate under these conditions.

4.4. Ablation Studies

We analyze key design choices on HouseCat6D to validate our unified model-free formulation with task factorization.

- **Object latent embedding.** Removing contrastive learning harms canonical correspondence stability and reduces HouseCat6D accuracy (Tab. 2), showing that object latent learning is key for category-agnostic behavior without predefined category names during inference.
- **Pointmap head.** Removing the pointmap head causes negligible change in absolute pose accuracy, which is expected: the pointmap branch serves relative SE(3) estimation, not absolute pose. This confirms that the two geometric pathways are decoupled by design — the absolute pose pathway (Depth + NOCS) is not degraded by the addition of the relative pose pathway, demonstrating that task factorization enables both capabilities within a single model at no accuracy cost to either task.
- **Metric head** Replacing the absolute metric head with a test-time Umeyama algorithm for scale calculation weakens pose accuracy under metric evaluation, indicating that

Table 2. **Category-level absolute pose estimation (RGB-D) on HouseCat6D.** Models are trained on respective training sets and evaluated on the test set. Prior columns indicate: **Category** (predefined category input), **Shape** (shape prior), **Semantic** (semantic prior), and **Calibration** (camera intrinsics). **MV** indicates whether the method supports multi-view pose reasoning. **Bold** and underlined denote the best and second-best results, respectively. We additionally report OPT-Pose with multi-view enhanced inference ($\mathbf{S}$=2, 3, 4), which is only applicable to our method and demonstrates its ability to leverage cross-view geometric cues for absolute pose estimation.

Dataset	Method	Category	Shape	Semantic	Calibration	MV	5°2cm	5°5cm	10°2cm	10°5cm	IoU50	IoU75
	Single-view Inference (standard setting)											
HouseCat6D	DPDN [30]	✓	✓		✓		6.4	6.9	22.2	25.8	56.2	26.0
	VI-Net [31]	✓			✓		8.4	10.3	20.5	29.1	56.4	20.4
	SecondPose [5]	✓		✓	✓		11.0	13.4	25.3	35.7	66.1	24.9
	AG-Pose [32]	✓			✓		11.5	12.0	32.7	35.8	66.0	45.0
	Sphere-Pose [45]	✓			✓		19.3	25.9	40.9	55.3	72.2	-
	Spot-Pose [46]	✓			✓		23.8	24.5	52.3	54.8	77.0	-
	GCE-Pose [28]	✓	✓	✓	✓		24.8	25.7	**55.4**	58.4	**79.2**	**60.6**
	OPT-Pose ($\mathbf{S}$=1)					✓	28.0	31.9	53.0	60.3	78.1	40.7
	OPT-Pose ($\mathbf{K}_{\text{gt}}$, $\mathbf{S}$=1)				✓	✓	**29.3**	**32.1**	55.1	**60.4**	79.0	52.2
	Multi-view Enhanced Inference (ours only)											
	OPT-Pose ($\mathbf{K}_{\text{gt}}$, $\mathbf{S}$=2)				✓	✓	34.1	37.0	55.8	61.2	79.6	54.3
	OPT-Pose ($\mathbf{K}_{\text{gt}}$, $\mathbf{S}$=3)				✓	✓	36.1	38.9	56.1	61.7	**79.7**	54.8
	OPT-Pose ($\mathbf{K}_{\text{gt}}$, $\mathbf{S}$=4)				✓	✓	**36.9**	**39.7**	**56.4**	**62.1**	79.6	**54.9**

Table 3. **Results on Omni6DPose (ICCV 2025 WLCOP Challenge** [1]). Our model is trained solely on Omni6DPose, whereas GenPose++ [67] and CPPF++ [64] further leverage categorical augmentation from PACE [65].

Dataset	Method	AUC ↑			VUS ↑			
		IoU25	IoU50	IoU75	5°2cm	5°5cm	10°2cm	10°5cm
ROPE	GenPose++ [66]	39.64	18.68	1.28	8.02	11.90	16.58	24.87
	CPPF++ [64]	38.19	17.18	1.07	**9.43**	**14.20**	**18.28**	**27.66**
	OPT-Pose ($\mathbf{K}_{\text{gt}}$)	**40.06**	**21.03**	**2.73**	9.08	13.14	17.62	25.46

Table 4. **Relative Object Pose Estimation Results.** AUC of ADD(-S), AR, MSSD, MSPD, and VSD on Real275 and Toyota-Light.

Dataset	Method	ADD(-S)	AR	MSSD	MSPD	VSD
REAL275	SIFT [37]	16.4	34.1	37.9	48.0	16.5
	Obj. Mat. [14]	13.4	26.0	31.7	30.8	15.5
	Oryon [7]	34.9	46.5	50.9	56.7	32.1
	Any6D [23]	53.5	51.0	56.5	65.3	31.1
	One2Any [35]	41.0	54.9	-	-	-
	OPT-Pose	**94.2**	**84.2**	**89.3**	**91.3**	**71.9**
Toyota-Light	SIFT [37]	14.1	30.3	39.6	44.1	7.3
	Obj. Mat. [14]	5.4	9.8	13.0	14.0	2.4
	Oryon [7]	22.9	34.1	42.9	45.5	13.9
	Any6D [23]	32.2	43.3	55.8	58.4	15.8
	One2Any [35]	34.6	42.0	-	-	-
	OPT-Pose	**47.0**	**57.1**	**59.4**	**62.8**	**49.0**

learning absolute translation/size from depth is necessary.

- **Keypoint extraction** Replacing the keypoint extraction with direct pixel-level DPT, with sensor depth aligned, significantly drops the performance, showcasing that the dense prediction may drop the performance due to noise(prediction, depth).
- **Multi-view reasoning (Tab. 2)** Increasing the number of input views at inference time consistently improves absolute pose estimation without retraining, indicating that relative geometric reasoning provides complementary constraints beyond single-view predictions.

Table 5. Ablations for single-view category-level absolute pose on HouseCat6D (RGB-D). Metrics follow Sec. 4.

HouseCat6D (RGB-D)	5°2cm	5°5cm	10°2cm	10°5cm
Full pipeline	28.0	31.9	53.0	60.3
w/o pointhead	27.9	30.6	54.5	58.8
w/o object latent embedding	24.5	26.7	48.9	52.6
w/o Abs. Metric Head	22.2	24.2	48.8	52.2
w/o Keypoint Extration (DPT)	20.0	22.9	47.0	54.8

5. Conclusions and Limitations

We presented Object Pose Transformer (OPT-Pose), the first unified model-free framework for task-factorized unseen object pose estimation. OPT-Pose bridges two previously fragmented paradigms: category-level absolute pose and unseen-object relative pose estimation. Our approach employs a single feed-forward model that predicts point maps, depth, NOCS, and camera parameters from RGB images. The complementary pairing of Depth+NOCS and Depth+Pointmap achieves canonical and relative pose reasoning. Evaluated on diverse datasets, OPT-Pose delivers state-of-the-art accuracy on both category-level absolute and unseen-object relative pose benchmarks within a unified architecture. Although OPT-Pose achieves strong performance on model-free tasks, the model shares limitations with other works in this domain: it requires object-centric crops, and performance degrades with large illumination changes. Moreover, model canonicalization can be ambiguous under symmetry or when the dataset convention changes. We believe this unified model-free framework with task factorization opens a promising direction for generic, category- and camera-agnostic object pose understanding, where future work may explore large-scale pretraining, multi-object reasoning, and tighter integration with robotic manipulation.

References

[1] Wclop 2025 challenge: Category-level object pose estimation for robotic manipulation. https://www.codabench.org/competitions/9742/, 2025. 8

[2] Junhao Cai, Yisheng He, Weihao Yuan, Siyu Zhu, Zilong Dong, Liefeng Bo, and Qifeng Chen. Ov9d: Open-vocabulary category-level 9d object pose and size estimation. *arXiv preprint arXiv:2403.12396*, 2024. 2

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

[4] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Shen Linlin, and Ales Leonardis. Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1581–1590, 2021. 2

[5] Yamei Chen, Yan Di, Guangyao Zhai, Fabian Manhardt, Chenyangguang Zhang, Ruida Zhang, Federico Tombari, Nassir Navab, and Benjamin Busam. Secondpose: Se (3)-consistent dual-stream feature fusion for category-level pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9959–9969, 2024. 1, 2, 3, 8

[6] Jaime Corsetti, Davide Boscaini, Francesco Giuliari, Changjae Oh, Andrea Cavallaro, and Fabio Poiesi. High-resolution open-vocabulary object 6d pose estimation. *arXiv preprint arXiv:2406.16384*, 2024. 1, 2

[7] Jaime Corsetti, Davide Boscaini, Changjae Oh, Andrea Cavallaro, and Fabio Poiesi. Open-vocabulary object 6d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18071–18080, 2024. 1, 2, 7, 8

[8] Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6781–6791, 2022. 2

[9] Zhaoxin Fan, Zhenbo Song, Jian Xu, Zhicheng Wang, Kejian Wu, Hongyan Liu, and Jun He. Object level depth reconstruction for category level 6d object pose estimation from monocular rgb image. In *European Conference on Computer Vision*, pages 220–236. Springer, 2022. 2, 6

[10] Tom Fischer, Xiaojie Zhang, and Eddy Ilg. Unified category-level object detection and pose estimation from rgb images using 3d prototypes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9790–9800, 2025. 2, 6

[11] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24:381–395, 1981. 2

[12] Yang Fu and Xiaolong Wang. Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. *Advances in Neural Information Processing Systems*, 35:27469–27483, 2022. 1

[13] Zheng Geng, Nan Wang, Shaocong Xu, Chongjie Ye, Bohan Li, Zhaoxi Chen, Sida Peng, and Hao Zhao. One view, many worlds: Single-image to 3d object meets generative domain randomization for one-shot 6d pose estimation. *arXiv preprint arXiv:2509.07978*, 2025. 2

[14] Can Gümeli, Angela Dai, and Matthias Nießner. Objectmatch: Robust registration using canonical object correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13082–13091, 2023. 8

[15] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without CAD models. In *NeurIPS*, 2022. 1, 2

[16] 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)*, pages 19–34, 2018. 2

[17] Zinqin Huang, Gu Wang, Chenyangguang Zhang, Ruida Zhang, Xiu Li, and Xiangyang Ji. Givepose: Gradual intra-class variation elimination for rgb-based category-level object pose estimation. In *CVPR*, 2025. 1, 2, 6

[18] HyunJun Jung, Shun-Cheng Wu, Patrick Ruhkamp, Guangyao Zhai, Hannah Schieber, Giulia Rizzoli, Pengyuan Wang, Hongcheng Zhao, Lorenzo Garattoni, Sven Meier, et al. Housecat6d-a large-scale multi-modal category level 6d object perception dataset with household objects in realistic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22498–22508, 2024. 2, 6

[19] Akshay Krishnan, Abhijit Kundu, Kevis-Kokitsi Maninis, James Hays, and Matthew Brown. Omninocs: A unified nocs dataset and model for 3d lifting of 2d objects. In *European Conference on Computer Vision*, pages 127–145. Springer, 2024. 2

[20] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *ECCV*, 2020. 1

[21] 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. *arXiv preprint arXiv:2212.06870*, 2022. 1, 2, 3

[22] Taeyeop Lee, Byeong-Uk Lee, Myungchul Kim, and In So Kweon. Category-level metric scale object shape and pose estimation. *IEEE Robotics and Automation Letters*, 6(4): 8575–8582, 2021. 2, 6

[23] Taeyeop Lee, Bowen Wen, Minjun Kang, Gyuree Kang, In So Kweon, and Kuk-Jin Yoon. Any6D: Model-free 6d pose estimation of novel objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025. 8

[24] Taeyeop Lee, Bowen Wen, Minjun Kang, Gyuree Kang, In So Kweon, and Kuk-Jin Yoon. Any6D: Model-free 6D Pose Estimation of Novel Objects, 2025. arXiv:2503.18673 [cs]. 2

[25] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o(n) solution to the pnp problem. *International Journal of Computer Vision*, 81, 2009. 2

[26] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. 2

[27] Weihang Li, Junwen Huang, HyunJun Jung, Guangyao Zhai, Pierluigi Zama Ramirez, Alex Costanzino, Fabio Tosi, Matteo Poggi, Luigi Di Stefano, Jean-Baptiste Weibel, Doris Antensteiner, Markus Vincze, Jing He, Yiqing Wang, Kexin Zhang, Licheng Jiao, Lingling Li, Fang Liu, Wenping Ma, and Benjamin Busam. Tricky 2025 housecat6d object pose estimation challenge with specular and transparent surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3323–3333, 2025. 1

[28] Weihang Li, Hongli XU, Junwen Huang, Hyunjun Jung, Peter KT Yu, Nassir Navab, and Benjamin Busam. Gce-pose: Global context enhancement for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27154–27165, 2025. 1, 2, 3, 4, 6, 8

[29] Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3560–3569, 2021. 1, 2, 5

[30] Jiehong Lin, Zewei Wei, Changxing Ding, and Kui Jia. Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks. In *European Conference on Computer Vision*, pages 19–34. Springer, 2022. 2, 8

[31] Jiehong Lin, Zewei Wei, Yabin Zhang, and Kui Jia. Vi-net: Boosting category-level 6d object pose estimation via learning decoupled rotations on the spherical representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14001–14011, 2023. 1, 5, 8

[32] Xiao Lin, Wenfei Yang, Yuan Gao, and Tianzhu Zhang. Instance-adaptive and geometric-aware keypoint learning for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21040–21049, 2024. 2, 3, 4, 7, 8

[33] Jianhui Liu, Yukang Chen, Xiaoqing Ye, and Xiaojuan Qi. Ist-net: Prior-free category-level pose estimation with implicit space transformation, 2023. 1

[34] Jian Liu, Wei Sun, Hui Yang, Pengchao Deng, Chongpei Liu, Nicu Sebe, Hossein Rahmani, and Ajmal Mian. Diff9d: Diffusion-based domain-generalized category-level 9-dof object pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[35] Mengya Liu, Siyuan Li, Ajad Chhatkuli, Prune Truong, Luc Van Gool, and Federico Tombari. One2any: One-reference 6d pose estimation for any object. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6457–6467, 2025. 8

[36] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9970–9980, 2024. 2

[37] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, pages 1150–1157. Ieee, 1999. 8

[38] 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*, pages 9903–9913, 2024. 1, 2

[39] 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. 3, 4

[40] 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*, pages 163–182. Springer, 2025. 1

[41] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. *arXiv preprint arXiv:1709.07871*, 2018. 2

[42] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017. 5

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[44] Alberto Remus, Salvatore D'Avella, Francesco Di Felice, Paolo Tripicchio, and Carlo Alberto Avizzano. i2c-net: Using instance-level neural networks for monocular category-level 6d pose estimation. *IEEE Robotics and Automation Letters*, 8(3):1515–1522, 2023. 1

[45] Huan Ren, Wenfei Yang, Xiang Liu, Shifeng Zhang, and Tianzhu Zhang. Learning shape-independent transformation via spherical representations for category-level object pose estimation. In *The Thirteenth International Conference on Learning Representations*, 2025. 8

[46] Huan Ren, Wenfei Yang, Shifeng Zhang, and Tianzhu Zhang. Rethinking correspondence-based category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 8

[47] Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic. Osop: A multi-stage one shot object pose estimation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6835–6844, 2022. 2, 3

[48] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou.

Onepose: One-shot object pose estimation without cad models. In *CVPR*, 2022. 3

[49] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 530–546. Springer, 2020. 1

[50] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04):376–380, 1991. 2, 4

[51] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[52] Ashish Vaswani. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 4

[53] 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 (CVPR)*, 2019. 1, 2

[54] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3, 4, 6

[55] Pengyuan Wang, Takuya Ikeda, Robert Lee, and Koichi Nishiwaki. Gs-pose: Category-level object pose estimation via geometric and semantic correspondence. In *European Conference on Computer Vision*, pages 108–126. Springer, 2025. 2

[56] Qianqian Wang*, Yifei Zhang*, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 2

[57] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2

[58] Jiaxin Wei, Xibin Song, Weizhe Liu, Laurent Kneip, Hongdong Li, and Pan Ji. Rgb-based category-level object pose estimation via decoupled metric scale recovery. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2036–2042. IEEE, 2024. 2, 6

[59] 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*, pages 17868–17879, 2024. 2

[60] Tianhao Wu, Chuanxia Zheng, Frank Guan, Andrea Vedaldi, and Tat-Jen Cham. Amodal3r: Amodal 3d reconstruction from occluded 2d images. *arXiv preprint arXiv:2503.13439*, 2025. 2

[61] Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *CVPR*, 2025. 3, 4

[62] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21469–21480, 2025. 2

[63] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 2

[64] Yang You, Wenhao He, Jin Liu, Hongkai Xiong, Weiming Wang, and Cewu Lu. Cppf++: Uncertainty-aware sim2real object pose estimation by vote aggregation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 9239–9254, 2024. 6, 8

[65] Yang You, Kai Xiong, Zhening Yang, Zhengxiang Huang, Junwei Zhou, Ruoxi Shi, Zhou Fang, Adam W. Harley, Leonidas Guibas, and Cewu Lu. Pace: Pose annotations in cluttered environments, 2024. 8

[66] Jiyao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In *European Conference on Computer Vision*. Springer, 2024. 1, 2, 6, 8

[67] Jiyao Zhang, Mingdong Wu, and Hao Dong. Generative category-level object pose estimation via diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3, 8

[68] Jiyao Zhang, Mingdong Wu, and Hao Dong. Generative category-level object pose estimation via diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[69] Mengchen Zhang, Tong Wu, Tai Wang, Tengfei Wang, Ziwei Liu, and Dahua Lin. Omni6d: Large-vocabulary 3d object dataset for category-level 6d object pose estimation, 2024. 2

[70] Ruida Zhang, Ziqin Huang, Gu Wang, Chenyangguang Zhang, Yan Di, Xingxing Zuo, Jiwen Tang, and Xiangyang Ji. Lapose: Laplacian mixture shape modeling for rgb-based category-level object pose estimation. In *European Conference on Computer Vision*, pages 467–484. Springer, 2024. 2, 6

[71] Linfang Zheng, Chen Wang, Yinghan Sun, Esha Dasgupta, Hua Chen, Aleš Leonardis, Wei Zhang, and Hyung Jin Chang. Hs-pose: Hybrid scope feature extraction for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17163–17173, 2023. 2

[72] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5