GCE-Pose: Global Context Enhancement for Category-level Object Pose Estimation

Weihang Li[1,2*]   Hongli Xu[1*]   Junwen Huang[1,2*†]   Hyunjun Jung[1*]

Peter KT Yu[3]   Nassir Navab[1,2]   Benjamin Busam[1,2]

[1]Technical University of Munich   [2]Munich Center for Machine Learning   [3]XYZ Robotics

Abstract

A key challenge in model-free category-level pose estimation is the extraction of contextual object features that generalize across varying instances within a specific category. Recent approaches leverage foundational features to capture semantic and geometry cues from data. However, these approaches fail under partial visibility. We overcome this with a first-complete-then-aggregate strategy for feature extraction utilizing class priors. In this paper, we present GCE-Pose, a method that enhances pose estimation for novel instances by integrating category-level global context prior. GCE-Pose performs semantic shape reconstruction with a proposed Semantic Shape Reconstruction (SSR) module. Given an unseen partial RGB-D object instance, our SSR module reconstructs the instance's global geometry and semantics by deforming category-specific 3D semantic prototypes through a learned deep Linear Shape Model. We further introduce a Global Context Enhanced (GCE) feature fusion module that effectively fuses features from partial RGB-D observations and the reconstructed global context. Extensive experiments validate the impact of our global context prior and the effectiveness of the GCE fusion module, demonstrating that GCE-Pose significantly outperforms existing methods on challenging real-world datasets HouseCat6D and NOCS-REAL275. Our project page is available at https://colin-de.github.io/GCE-Pose/.

1. Introduction

The task of object pose estimation varies according to generalization level and input modality. *Instance-level pose estimation* methods [23, 56, 63] focus on specific object instances that does not generalize to other objects, while *fully unseen object pose estimation* methods [12, 24, 54] are designed to handle novel objects, however requires object model as prior. Unlike aforementioned methods, *category-*
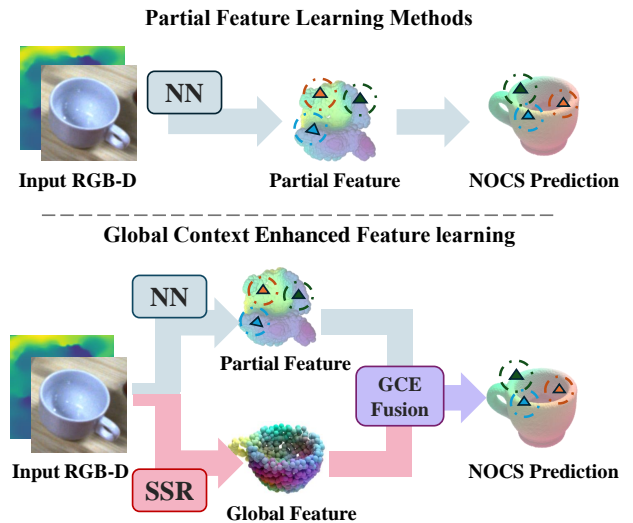

Figure 1. Overview of the category-level Pose Estimation Pipeline: (A) Previous methods, *i.g.* AG-Pose [39] and Second Pose [7], rely on partial features extracted by a neural network (NN) to regress object poses. (B) We introduce a novel approach that leverages a semantic shape reconstruction (SSR) module for global feature extraction. This global context enhances (GCE) the mapping from partial features to NOCS features.

level pose estimation methods [7, 36, 64] aim to generalize across unseen instances within a defined category that requires only an RGB(-D) image of a new instance during the inference, making the method model-free that does not require predefined object models.

Current category-level approaches primarily estimate the Normalized Object Coordinate Space (NOCS) [64] and employ a pose solver, such as the Umeyama algorithm [60], to obtain the object pose [36]. To effectively extract category-level features from RGB (and/or depth) inputs, researchers have developed various neural network architectures to capture features from partial RGB(-D) observations. Some recent methods [1, 45, 47] leverage foundation models like DINOv2 [46] for improved performance. Additionally, research [3, 7, 24, 38] highlights the importance of combining semantic and geometric information to enhance feature ro-

bustness and distinguishability, aiding in better correspondence and pose estimation. However, category-level pose estimation being model-free and having only partially observed RGB(-D) inputs limits the extraction of global context information. Some methods [5, 36, 59, 70, 78] have introduced categorical geometric shape priors to reconstruct instance models from partial input points, solving for object pose by establishing dense correspondence between partial input points and reconstructed models. However, these methods solely introduce shape priors neglecting the semantic context of the category. More recently, GS-Pose [69] selects one instance as a reference prototype within a category and applies semantic feature matching between partial points and the reference instance. However, this design struggles with intra-class shape variations and is particularly vulnerable to noise in partial point cloud observations.

In this work, we propose GCE-Pose, a novel approach that integrates global context incorporating both geometric and semantic cues to enhance category-level object pose estimation. We propose two major modules named Semantic Shape Reconstruction (SSR) and Global Context Enhanced (GCE) feature fusion modules to facilitate pose estimation. The SSR module is a first-complete-then-aggregate strategy that reconstructs the input partial points into a complete shape and smoothly aggregates the semantic prototype to the instance. The GCE feature fusion model is proposed to effectively fuse the reconstructed global context with local cues. The efficacy of our proposed method is confirmed by extensive evaluation on the challenging real-world datasets, achieving SOTA performance against the existing approaches. Our main contributions are as follows:

- We propose GCE-Pose, a Global Context Enhancement (GCE) approach that integrates global context with both geometric and semantic cues for category-level object pose estimation.
- We introduce a Semantic Shape Reconstruction (SSR) strategy that addresses partially observed inputs by reconstructing both object geometry and semantics through learned categorical deformation prototypes.
- Extensive experiments demonstrate that our method achieves robust pose estimation even under significant shape variations and occlusions improving the generalization to unseen instances.

2. Related Works

2.1. Object Reconstruction for Pose Estimation

Object reconstruction is essential for object pose estimation when CAD models are unavailable, as it captures object geometry and appearance while establishing a canonical space. Methods like OnePose [57], OnePose++ [20], and CosyPose [31] employ Structure from Motion (SfM) to match features across views, while approaches such as

NeRFPose [33], GS-Pose [2], and FoundationPose [71] utilize Neural Radiance Fields or 3D Gaussian Splatting [29] for flexible reconstruction.

For known object categories, semantic information can enhance instance-level reconstructions. Some methods build semantic representations directly: Goodwin *et al.* [18] align 3D views to a query view, and Zero123-6D [11] synthesizes views via diffusion models to reduce reference views. I2cNet [51] expands such techniques to categories by integrating a 3D mesh reconstruction module. Other methods use shape priors, such as SPD [59], RePoNet [16], and Wang *et al.* [65], which learn instance reconstructions from category-specific priors. SGPA [5] and RBP [78] dynamically adapt priors based on observed structures, while SAR-Net [34] and ACR-Pose [14] further incorporate geometric and adversarial strategies. GS-Pose [69] projects DINOv2 [46] features onto a 3D reference shape, aiding feature alignment and pose prediction. From here, we propose a novel method that further improves the aforementioned methods by integrating semantics to the reconstructions to provide global contextual information for pose estimation.

2.2. Representation Learning for Pose Estimation

Learning effective feature representations from input modalities is crucial to pose estimation, evolving alongside advancements in vision neural networks. Early visual feature extractors relied on CNN backbones [28, 32, 48, 58, 72, 76] to predict or refine object poses from single RGB images. Recently, foundational models like DINOv2 have been widely adopted [1, 7, 38, 45, 47] to enhance robustness and contextual understanding.

Beyond 2D-only approaches, many methods now combine 2D image and 3D point cloud networks to jointly extract semantic and geometric features, constructing robust embeddings for tasks such as direct pose regression [7, 21, 22, 62] or feature matching [3, 24, 38]. RGB-D methods address feature fusion at multiple levels, such as 2D-3D and local-global fusion. Methods like DenseFusion [62] and PVN3D [21] concatenate per-pixel geometric and RGB features, while SecondPose [7] employs MLPs to fuse DINOv2 [46] features with Point Pair Features (PPF) [13].

More recent transformer-based approaches, including SAM6D [38] and MatchU [24], integrate both local RGB-D and global CAD model features [50, 74, 75], demonstrating the value of cross-modality fusion. In this work, we not only incorporate categorical semantic priors into object reconstruction but also effectively integrate global context into local embeddings through our fusion module.

Occlusions and object symmetries introduce visual ambiguities [42], necessitating the consideration of multiple correct poses. To address this, various methods frame pose prediction as distribution estimation, learning ambiguity-aware representations [19, 53, 61]. We tackle the occlusion

(A) Robust Partial Feature Extractor

(C) Global Context Enhanced Feature Fusion

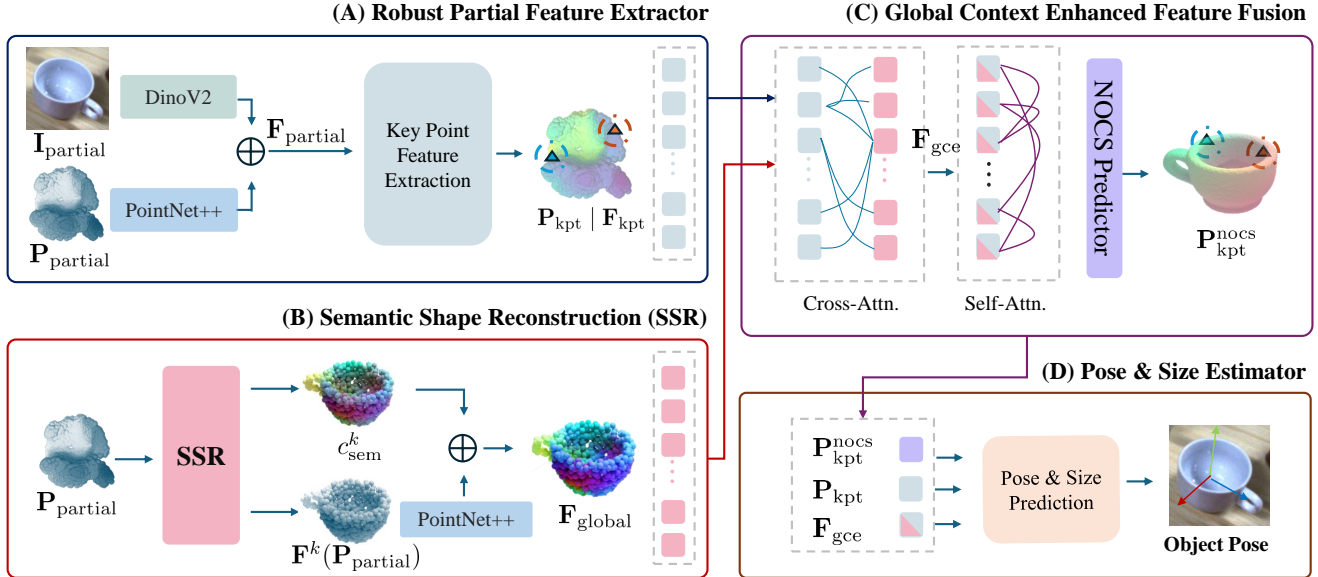**(B) Semantic Shape Reconstruction (SSR)**

(D) Pose & Size Estimator

Figure 2. Illustration of GCE-Pose: (A) Semantic and geometric features are extracted from an RGB-D input. A keypoint feature detector identifies robust keypoints and extracts their corresponding features. (B) An instance-specific and category-level semantic global feature is reconstructed using our SSR module. (C) The global features are fused with the keypoint features to predict the keypoint NOCS coordinates. (D) The predicted keypoints, NOCS coordinates, and fused keypoint features are utilized for pose and size estimation.

challenge by employing a completion task that enables the network to reason about full geometric representations despite missing points.

2.3. Generalizing Object Pose Estimators

The constraints of 3D model-based pose estimation have been relaxed to category-level estimation, where the task is to predict the pose of an unknown instance within a known category (e.g., a fork in "cutlery") on benchmarks like NOCS [64], PhoCal [67], and HouseCat6D [27].

Category-level pose estimation aims to predict 9DoF poses for novel instances within specified categories. Wang *et al.* [64] introduced the Normalized Object Coordinate Space (NOCS) framework, mapping observed point clouds to a canonical space with pose recovery via the Umeyama algorithm [60]. Subsequent methods improve accuracy [4, 6, 10, 35, 39, 77, 79]. Some works adopt prior-free methods, such as VI-Net [37], which separates rotation components, and IST-Net [40], which transforms camera-space features implicitly. AG-Pose [39] achieves state-of-the-art results by learning keypoints from RGB-D without priors. In contrast, we incorporate learned priors for geometry and semantics to complete partial observations and map mean shape semantics onto observed instances.

Self-supervised approaches are also popular in category-level estimation, refining models without annotated real data. CPS++ [43] uses a differentiable renderer to adapt synthetic models with real, unlabeled RGB-D inputs, while other works [66, 68] tackle photometric challenges using RGBP polarization [17, 52] and contextual language cues [68]. Self-DPDN [36] employs a shape deformation network for self-supervision, while our approach leverages a categorical shape prior without network refinement.

In open-vocabulary settings, POPE [15] introduces promptable object pose, (H)Oryon [8, 9] use vision-language models and stereo matching, while NOPE [44] and SpaRP [73] predict pose distributions or relative NOCS-maps. These methods often treat pose estimation as correspondence matching [9, 15] or reconstruction [44]. We instead deform a mean shape within an absolute category space to capture instance-specific correspondences.

3. Method

The objective of GCE-Pose is to estimate the 6D object pose and size from RGB-D data. Given a single RGB-D frame and the category instance mask, we obtain the partial RGB observation $\mathbf{I}_{partial}$ and its corresponding partial point cloud $\mathbf{P}_{partial}$ derived from the depth map. Utilizing $\mathbf{I}_{partial}$ and $\mathbf{P}_{partial}$, the objective is to recover the 3D rotation $\mathbf{R} \in SO(3)$, the 3D translation $\mathbf{t} \in \mathbb{R}^3$, and the size $\mathbf{s} \in \mathbb{R}^3$ of the target object.

GCE-Pose consists of four main modules (Fig. 2): Robust Partial Feature Extraction (Sec. 3.1), Semantic Shape Reconstruction (Sec. 3.2), Global Context Enhanced Feature Fusion (Sec. 3.3), and Pose & Size Estimator (Sec. 3.5).

3.1. Robust Partial Feature Extraction

Partial observations from RGB-D sensors often contain significant noise and incomplete geometry, making dense cor-

respondence prediction unreliable. We address this challenge with a keypoint-based approach [39] that focuses on the most discriminative and reliable object regions.

The N input points are put in order within $\mathbf{P}_{\text{partial}} \in \mathbb{R}^{N \times 3}$ and we extract point features $\mathbf{F}_{\text{P}} \in \mathbb{R}^{N \times C_1}$ using PointNet++ [49]. For the RGB image $\mathbf{I}_{\text{partial}}$, we extract the image feature $\mathbf{F}_{\text{I}} \in \mathbb{R}^{N \times C_2}$ using DINOv2 [46] and concatenate $\mathbf{F}_{\text{I}}$ to $\mathbf{F}_{\text{P}}$ to obtain $\mathbf{F}_{\text{partial}} \in \mathbb{R}^{N \times C}$. We follow AG-Pose [39] for keypoint detection. First, M keypoint features are extracted using a learnable embedding $\mathbf{F}_{\text{emb}} \in \mathbb{R}^{M \times C}$, which undergoes cross-attention with $\mathbf{F}_{\text{partial}}$ to attend to critical regions in $\mathbf{P}_{\text{partial}}$. This process yields a feature query matrix $\mathbf{F}_{\text{q}} = \text{CrossAttention}(\mathbf{F}_{\text{emb}}, \mathbf{F}_{\text{partial}})$. We then compute correspondences via cosine similarity, forming a matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$, and select M keypoints from $\mathbf{P}_{\text{partial}}$ as $\mathbf{P}_{\text{kpt}} = \text{softmax}(\mathbf{A})\mathbf{P}_{\text{partial}}$. To ensure keypoints lie on the object surface and minimize outliers, an object-aware Chamfer distance loss $\mathcal{L}_{\text{ocd}}$ is applied. With ground truth pose $\mathbf{T}_{\text{gt}}$, we filter outliers by comparing each point $x \in \mathbf{P}_{\text{partial}}$ to the instance model $\mathbf{M}_{\text{obj}}$:

$$\min_{y \in \mathbf{M}_{\text{obj}}} \|\mathbf{T}_{\text{gt}}(x) - y\|_2 < \tau_1, \tag{1}$$

where τ_1 is an outlier threshold. The object-aware Chamfer distance loss is then:

$$\mathcal{L}_{\text{ocd}} = \frac{1}{|\mathbf{P}_{\text{kpt}}|} \sum_{x \in \mathbf{P}_{\text{kpt}}} \min_{y \in \mathbf{P}^*_{\text{partial}}} \|x - y\|_2. \tag{2}$$

To prevent keypoints from clustering, a diversity regularization loss is added:

$$\mathcal{L}_{\text{div}} = \sum_{x \neq y \in \mathbf{P}_{\text{kpt}}} \max\{0, \tau_2 - \|x - y\|_2\}, \tag{3}$$

where τ_2 controls keypoint distribution. To enhance features with geometric context, the Geometric-Aware Feature Aggregation (GAFA) module [39] is applied. GAFA augments each keypoint with (1) local geometric details from K-nearest neighbors and (2) global information from all keypoints, improving feature discriminability for correspondence estimation.

3.2. Semantic Shape Reconstruction

Intra-class variation is a key challenge in category-level pose estimation. To tackle this issue, category-level shape priors have been extensively used in object pose estimation. By representing the shape with mean shapes and deformations [59] or learning implicit neural representations for geometry recovery [25, 26], pose estimators can better learn correspondences in NOCS space, benefiting from accurate shape priors. While geometric shape reconstruction provides valuable priors, it cannot fully capture the rich semantic information of object parts. Recent advances in 2D foundation models, particularly DINO [46], have demonstrated

remarkable capabilities in extracting zero-shot semantic information from single RGB images. Building upon this insight, we propose **Semantic Shape Reconstruction (SSR)** to learn a per-category linear shape model similar to [41] that describes an object using instance-specific geometry and category-level semantic features.

Deep Linear Semantic Shape Model. To overcome the challenges posed by partial observations from depth sensors, such as occlusions and incomplete geometry, we employ a variation of the deep linear shape model [41]. This approach is motivated by the need to robustly and efficiently parameterize object shapes with shape parameters and produce a completed 3D object representation, even when faced with limited input data. We represent each point in our model as a tuple (x, f) where $x \in \mathbb{R}^3$ represents a spatial coordinate and $f \in \mathbb{R}^C$ represents its semantic feature vector. For I points of an object instance within category k, we learn a linear shape model. The model for category k consists of (i) a geometric prototype $c^k \in \mathbb{R}^{I \times 3}$ with associated semantic features $c^k_{\text{sem}} \in \mathbb{R}^{I \times C}$, (ii) a set of geometric deformation basis vectors $v^k = \{v^k_1, \ldots, v^k_D\}$ where $v^k_i \in \mathbb{R}^{I \times 3}$, and (iii) a scale parameter vector $s^k \in \mathbb{R}^3$. The key insight of our approach is that semantic features remain coupled to their corresponding points during geometric deformation. Any semantic shape $\mathbf{U}_{\text{k}}$ in the model family is defined by:

$$\mathbf{U}_{\text{k}} = (\mathbf{X}_{\text{k}}, \mathbf{F}_{\text{k}}) = \left(s^k \odot (c^k + \sum_{i=1}^{D} a^k_i v^k_i), c^k_{\text{sem}} \right) \tag{4}$$

where $\mathbf{X}_{\text{k}} \in \mathbb{R}^{I \times 3}$ are the I points in shape prior k and $\mathbf{F}_{\text{k}} \in \mathbb{R}^{I \times C}$ their associated features. The shape parameter vector is given by $a^k = (a^k_1, \ldots, a^k_D) \in \mathbb{R}^D$, $s^k \in \mathbb{R}^3$ controls scaling, and $\odot$ defines the element-wise Hadamard product. We train two neural networks for each category k to predict shape parameter a^k with network $\mathcal{D}^k$ and scale s^k with $\mathcal{S}^k$.

To optimize the model, we minimize the Chamfer distance loss, $\mathcal{L}_{\text{CD}}$, which ensures accurate shape reconstruction through:

$$\mathcal{L}_{\text{CD}} = \sum_{x \in \mathbf{P}} \min_k d(x, \mathbf{U}_{\text{k}}), \tag{5}$$

with the Chamfer distance d, ground truth point clouds $\mathbf{P}$ from category k and shape reconstruction $\mathbf{U}_{\text{k}}$ defined in Eq. (4). Training with ground truth yields the optimal parameters $\bar{a}^k$, $\bar{s}^k$, c^k, and v^k which allow to formulate an additional loss to refine shape reconstruction under partial observations $\mathbf{P}_{\text{partial}}$ by freezing c^k, and v^k within

$$\mathcal{L}_{\text{para}} = \sum_{x' \in \mathbf{P}_{\text{partial}}} \lambda_1 \left| \mathcal{D}^k(x') - \bar{a}^k \right| + \lambda_2 \left| \mathcal{S}^k(x') - \bar{s}^k \right|. \tag{6}$$

Finally, we combine the reconstruction and parameter loss to formulate the overall loss

$$\mathcal{L}_{\text{rec}} = \lambda_{\text{CD}} \cdot \mathcal{L}_{\text{CD}} + \lambda_{\text{para}} \cdot \mathcal{L}_{\text{para}}, \tag{7}$$
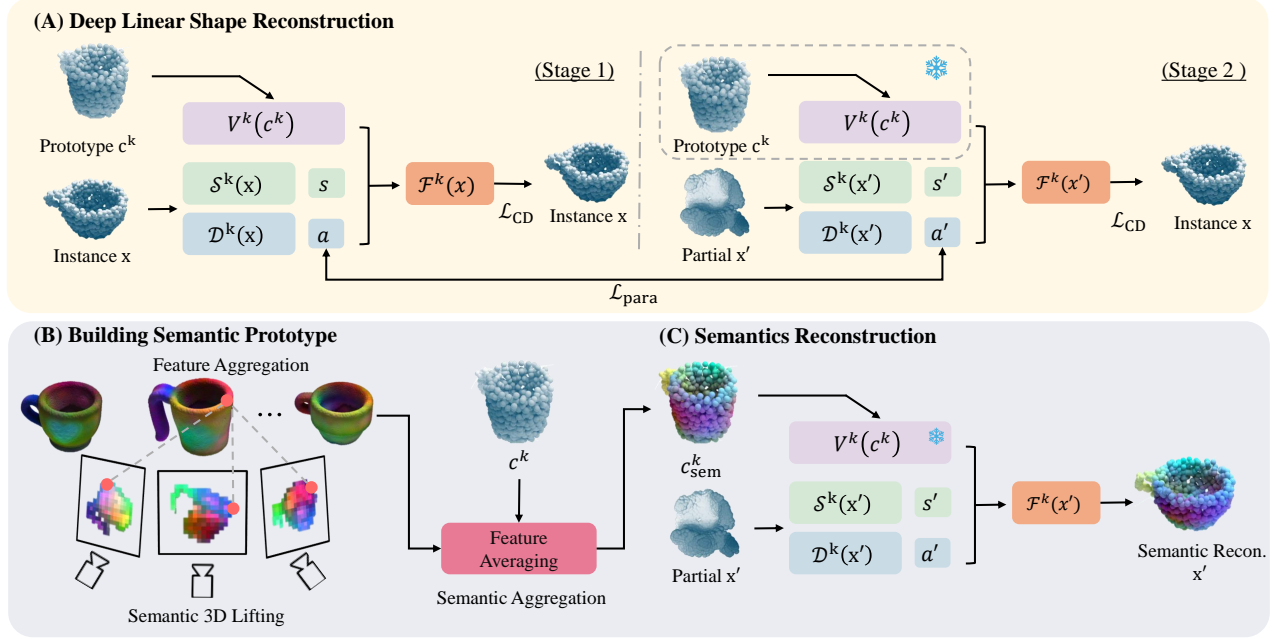
4

Figure 3. Illustration of Deep Linear Semantic Shape Model. A Deep Linear Semantic Shape model is composed of a prototype shape c, a scale network $\mathcal{S}$, a deformation network $\mathcal{D}$, a Deformation field $\mathcal{V}$ and a category-level semantic features c_{sem}^k. At stage 1, we build a Deep Linear Shape (DLS) model using sampled point clouds from all ground truth instances within each category, training a linear parameterization network to represent each instance. At stage 2, we retrain the DLS model to regress the corresponding DLS parameters from partial point cloud inputs using a deformation and scale network. During testing, the network predicts DLS parameters for unseen objects and reconstructs their point clouds based on the learned deformation field to get semantic reconstruction.

where λ_{CD} and λ_{para} are the hyperparameters that weight the contributions of $\mathcal{L}_{\text{CD}}$ and $\mathcal{L}_{\text{para}}$, respectively.

Semantic Prototype Construction. To effectively integrate rich semantic information into our 3D shape reconstruction, we employ a process that begins by extracting dense semantic features from multiple RGB images of each object instance using the DINOv2 [46]. For each object instance without texture, we position multiple virtual cameras around the object to capture RGB images and depth maps from diverse viewpoints. This setup ensures full coverage of the object's surface and mitigates occlusion effects. The RGB images are processed through the DINOv2 model to extract dense 2D semantic feature maps. Using the corresponding depth maps and known camera intrinsics and extrinsic, we project the 2D semantic features into 3D space. For each pixel (u, v) in the image, we compute its 3D position P using the depth value z and project the associated semantic feature $\mathbf{f}_{\text{2D}}(u, v)$ to this point $P = zK^{-1}[u, v, 1]^T$, , where K is the camera intrinsic matrix. As a result, we obtain a dense semantic point cloud $\mathbf{F}_{\text{sem}}$. To ensure computational efficiency and point-wise correspondence, we downsample this dense semantic point cloud to I points aligned with our geometric reconstruction. For each point P_i in the deep linear shape reconstruction, we aggregate semantic

features from its k nearest neighbors in the dense cloud:

$$\mathbf{F}_{\text{instance}}(P_i) = \frac{1}{k} \sum_{P_j \in N_k(P_i)} \mathbf{F}_{\text{sem}}(P_j). \quad (8)$$

The category-level semantic prototype c_{sem}^k is then constructed by averaging N instance features across the category k while maintaining point-wise correspondence with the geometric prototype c^k:

$$c_{\text{sem}}^k = \frac{1}{N} \sum_{i=1}^{N} \mathbf{F}_{\text{instance}}(P_i^k) \quad (9)$$

Semantic Reconstruction. The key advantage of our approach is that semantic reconstruction becomes straightforward once the semantic prototype is established. Given a partial point cloud x', we first reconstruct its geometry and then directly inherit the semantic feature from the prototype.

3.3. Global Context Enhanced Feature Fusion

Traditional pose estimation methods rely primarily on partial observations, but they often struggle with challenges such as occlusion and viewpoint variations. To overcome these limitations, we propose a **Global Context Enhanced (GCE) Feature Fusion** module that effectively integrates complete semantic shape reconstructions with partial observations, establishing robust feature correspondences.

Firstly, we aim to extract global features from our semantic reconstruction. Given partial observation $\mathbf{P}_{\text{partial}}$, we can reconstruct the shape $\mathbf{P}_{\text{global}}$ using Eq. (4). We leverage PointNet++ [49] to obtain a geometry feature from $\mathbf{P}_{\text{global}}$, then concatenate it with category-level semantic feature c_{sem} to obtain the global feature $\mathbf{F}_{\text{global}} \in \mathbb{R}^{I \times C}$.

Given keypoint Features $\mathbf{F}_{\text{kpt}} \in \mathbb{R}^{M \times C}$ and global feature $\mathbf{F}_{\text{global}} \in \mathbb{R}^{I \times C}$, our goal is to enrich keypoint features with global semantic context. The primary challenge lies in bridging the domain gap between partial observations, captured in noisy camera-space coordinates, and the global reconstruction, represented in normalized object-space coordinates with complete shape information.

To fuse the partial feature $\mathbf{F}_{\text{kpt}}$ with $\mathbf{F}_{\text{global}}$, we transform both the partial and global features by concatenating learnable positional embedding network that maps these 3D positions $\mathbf{P}_{\text{kpt}}$ and $\mathbf{P}_{\text{global}}$. into high-dimensional positional tokens, respectively.

$$\mathbf{F}'_{\text{kpt}} = \text{concat}(\mathbf{F}_{\text{kpt}}, \text{PE}_{\text{kpt}}), \quad \mathbf{F}'_{\text{global}} = \text{concat}(\mathbf{F}_{\text{global}}, \text{PE}_{\text{global}}) \tag{10}$$

where PE is positional encoding network. We then apply an attention mechanism to merge both sources of information, where the global features provide the semantic context for refining the keypoint features. We first project keypoint features and global features into a shared embedding space:

$$\mathbf{F}''_{\text{kpt}} = \text{LayerNorm}(\text{MLP}_{\text{proj}}(\mathbf{F}'_{\text{kpt}})) \tag{11}$$

$$\mathbf{F}''_{\text{global}} = \text{LayerNorm}(\text{MLP}_{\text{proj}}(\mathbf{F}'_{\text{global}})) \tag{12}$$

then, the global context enhancement is aggregated through cross-attention and residual connection:

$$\mathbf{F}_{\text{context}} = \text{CrossAttn}(\mathbf{F}''_{\text{kpt}}, \mathbf{F}''_{\text{global}}) \tag{13}$$

$$\mathbf{F}_{\text{gce}} = \mathbf{F}_{\text{kpt}} + \mathbf{F}_{\text{context}} \tag{14}$$

After fusing the keypoint $\mathbf{F}_{\text{kpt}}$ feature with our global feature $\mathbf{F}_{\text{global}}$. The resulting global context enhanced keypoint feature $\mathbf{F}_{\text{gce}}$ are then passed through the self-attention module and MLP following [36] to predict the corresponding NOCS coordinates $\mathbf{P}_{\text{kpt}}^{\text{nocs}}$.

To ensure that our keypoints and associated features effectively represent the partial observation $\mathbf{P}_{\text{partial}}$, we additionally employ a reconstruction module to recover its 3D geometry. This module takes keypoint positions and features as input, applies positional encoding to the keypoints, and refines their features through a MLP. The encoded and refined features are aggregated, and a shape decoder predicts reconstruction deltas to recover the geometry. The reconstruction loss is defined as the object-aware Chamfer distance (CD) between the partial observation $\mathbf{P}_{\text{partial}}$ and the reconstructed point cloud $\mathbf{P}_{\text{recon}}$ following Eq. (1):

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathbf{P}_{\text{recon}}|} \sum_{x \in \mathbf{P}_{\text{recon}}} \min_{y \in \mathbf{P}_{\text{partial}}^{\star}} \|x - y\|_2. \tag{15}$$

3.4. Pose Size Estimator

Given the NOCS coordinates of keypoints, $\mathbf{P}_{\text{kpt}}^{\text{nocs}} \in \mathbb{R}^{M \times 3}$, the enhanced keypoint features $\mathbf{F}_{\text{gce}}$ and the position of keypoint $\mathbf{P}_{\text{kpt}}$, we can establish keypoint-level correspondences, which are then used to regress the final pose and size parameters, $\mathbf{R}$, $\mathbf{t}$, and $\mathbf{s}$. The process is formulated as follows:

$$\mathbf{f}_{\text{pose}} = \text{concat}\left[\mathbf{P}_{\text{kpt}}, \mathbf{F}_{\text{gce}}, \mathbf{P}_{\text{kpt}}^{\text{nocs}}\right] \tag{16}$$

$$(\mathbf{R}, \mathbf{t}, \mathbf{s}) = (\text{MLP}_R(\mathbf{f}_{\text{pose}}), \text{MLP}_t(\mathbf{f}_{\text{pose}}), \text{MLP}_s(\mathbf{f}_{\text{pose}})) \tag{17}$$

For the rotation representation $\mathbf{R}$, we utilize the 6D continuous representation proposed in [80]. For the translation $\mathbf{t}$, we adopt the strategy from [79] by predicting the residual translation between the ground truth and the mean position of the point cloud.

3.5. Overall Loss Function

The overall loss function for pose estimation is as follows:

$$\mathcal{L}_{\text{all}} = \lambda_1 \mathcal{L}_{\text{ocd}} + \lambda_2 \mathcal{L}_{\text{div}} + \lambda_3 \mathcal{L}_{\text{rec}} + \lambda_4 \mathcal{L}_{\text{nocs}} + \lambda_5 \mathcal{L}_{\text{pose}} \tag{18}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ are hyperparameters to balance the contribution of each term. For $\mathcal{L}_{\text{pose}}$, we use:

$$\mathcal{L}_{\text{pose}} = \|\mathbf{R}_{\text{gt}} - \mathbf{R}\|_{\text{F}} + \|\mathbf{t}_{\text{gt}} - \mathbf{t}\|_2 + \|\mathbf{s}_{\text{gt}} - \mathbf{s}\|_2. \tag{19}$$

We generate ground truth NOCS coordinates of keypoints $\mathbf{P}_{\text{kpt}}^{\text{gt}}$ by projecting their coordinates under camera space $\mathbf{P}_{\text{kpt}}$ into NOCS using the ground-truth $\mathbf{T}_{\text{gt}} = (\mathbf{R}_{\text{gt}}, \mathbf{t}_{\text{gt}}, \mathbf{s}_{\text{gt}})$. For $\mathcal{L}_{\text{nocs}}$, we use the Smooth L_1 loss with

$$\mathcal{L}_{\text{nocs}} = \|\mathbf{T}_{\text{gt}}(\mathbf{P}_{\text{kpt}}^{\text{gt}}) - \mathbf{P}_{\text{kpt}}^{\text{nocs}}\|_{\text{SL1}} \tag{20}$$

4. Experiment

4.1. Implementation Details

For the HouseCat6D dataset [27], cropped images are resized to 224×224 for feature extraction, and 1024 points are sampled from inputs. For Partial Feature Extraction, the number of keypoints is $M = 96$, and the feature dimensions for geometric and DINO features are $C_1 = 128$, $C_2 = 128$, and $C = 256$. In deep linear shape reconstruction, we set the basis dimension D to 5, and the number of points in the prototype is 1024. The pose estimation network is trained with batch size 36 with an ADAM [30] optimizer with a triangular2 cyclical learning rate schedule [55] on a single NVIDIA 4090 GPU for 150 epochs. We attach more implementation details in our Appendix.

4.2. Evaluation Benchmarks

Datasets. We evaluate our method on two challenging real-world benchmarks: HouseCat6D [27] and NOCS-REAL275 [64]. HouseCat6D contains 21K images of 194
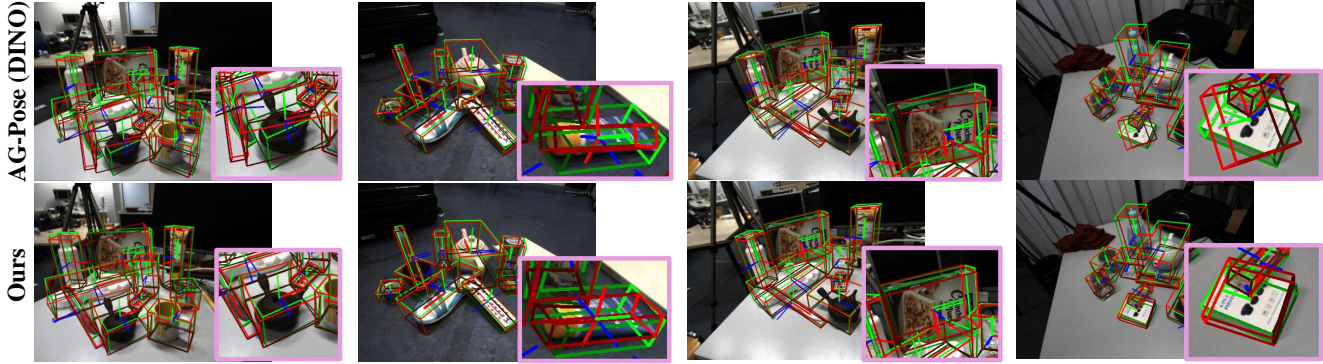
Figure 4. Visualization of category-level object pose estimation results on HouseCat6D dataset [27]. Predicted 3D bounding boxes are shown in red, with ground truth in green. Challenging cases are highlighted in pink side squares. Leveraging our global context-enhanced pose prediction pipeline, GCE-Pose outperforms the SOTA AG-Pose [39] (DINO), demonstrating robustness to occlusions and strong generalization to novel instances.

Dataset	Method	Shape Prior	Semantic Prior	5°2cm	5°5cm	10°2cm	10°5cm	IoU50	IoU75
HouseCat6D [27]	VI-Net [37]			8.4	10.3	20.5	29.1	56.4	-
	SecondPose [7]			11.0	13.4	25.3	35.7	66.1	-
	AG-Pose [39]			11.5	12.0	32.7	35.8	66.0	45.0
	AG-Pose (DINO) [39]			21.3	22.1	51.3	54.3	76.9	53.0
	GCE-Pose (Ours)	✓	✓	**24.8**	**25.7**	**55.4**	**58.4**	**79.2**	**60.6**
NOCS-REAL275 [64]	SecondPose [7]			56.2	63.6	74.7	86.0	-	-
	AG-Pose [39]			56.2	62.3	73.4	81.2	83.8	77.6
	AG-Pose(DINO) [39]			**57.0**	64.6	75.1	84.7	**84.1**	**80.1**
	RBP-Pose [78]	✓		38.2	48.1	63.1	79.2	-	67.8
	DPDN [36]	✓		46.0	50.7	70.4	78.4	83.4	76.0
	Query6DoF [70]	✓		49.0	58.9	68.7	83.0	82.5	76.1
	GS-Pose [69]	✓	✓	-	28.8	-	60.1	-	63.2
	GCE-Pose (Ours)	✓	✓	**57.0**	**65.1**	**75.6**	**86.3**	**84.1**	79.8

Table 1. Quantitative comparison of category-level object pose estimation on HouseCat6D and NOCS-REAL275 datasets.

household instances across 10 categories, with 2,929 images of 50 instances reserved for evaluation, covering diverse shapes, occlusions, and lighting. NOCS-REAL275 consists of 7K images across 6 object categories and 13 scenes, with 4.3K images used for training and 2,750 for testing. We compare our method against baselines on both datasets with the same setup for segmentation mask and conduct the ablation studies on HouseCat6D. **Evaluation Metrics.** Following prior work [7, 10, 37, 39], we evaluate performance with two metrics: $n°m$ cm and *3D IoU*. $n°m$ cm metric computes mean Average Precision (mAP) for rotation and translation accuracy, considering predictions correct if the rotation error is within $n°$ and translation error within m cm. *3D IoU* is a mAP-based metric that assesses 3D bounding box IoU with thresholds at 50% and 75%, capturing both pose and object size.

4.3. Comparison with the State-of-the-Art

Results on HouseCat6D. Table 1 presents GCE-Pose's performance on the HouseCat6D dataset, where it outperforms state-of-the-art methods across all metrics against the existing approaches. Compared to AG-Pose, even strengthened with DINOv2 [39], our method shows 16% improvement on

the most strict 5°2cm metric and notable gains on other metrics, demonstrating the efficacy of integrating global context priors into the local feature-based pose estimation pipeline. Against SecondPose [7], which fuses PPF geometric and DINOv2 semantic features from only the partial observations, our method achieves more than 30% improvement in IoU75 metric and over 100% in $n°m$ cm metrics. Qualitative comparisons of our method against AG-Pose are shown in Figure 4.

Results on NOCS-REAL275. Table 1 also shows GCE-Pose's performance on NOCS-REAL275 dataset, achieving the highest scores on most metrics. GCE-Pose surpasses prior-free baselines by integrating robust global context priors into a strong pose estimator, proving the benefits of our design. We also compare with methods using categorical priors: RBP-Pose [78], Query6DoF [70] and DPND [36], which rely solely on shape priors, while GS-Pose [69] uses both shape and semantic priors with a single-instance reference. In contrast, GCE-Pose learns robust deformation priors across multiple instances and completes partial inputs into a full semantic shape, outperforming all prior-based methods and by more than 23% on 5°2cm metric, demon-
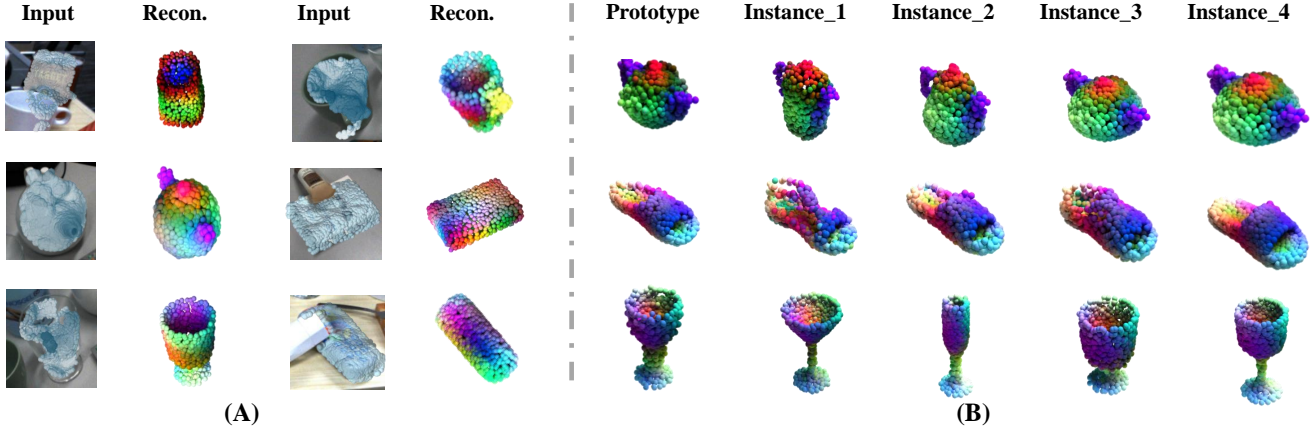
Figure 5. (A) shows the visualization of input partial points and the output semantic shape reconstructions; (B) visualizes the semantic prototypes of different categories and the aggregated instance semantics.

strating the effectiveness of our SSR and GCE fusion modules.

Method	Ins. recon.	Mean shape	Geo.	Sem.	5°2cm	5°5cm
(0) AG-Pose DINO (Baseline)	×	×	×	×	21.3	22.1
(1) Ours (Instance Geo.)	✓	×	✓	×	22.2	23.7
(2) Ours (Categorical Sem.)	×	✓	×	✓	22.7	24.3
(3) Ours (Mean shape Geo. & Categorical Sem.)	×	✓	✓	✓	23.4	24.2
(4) GCE-Pose (Full pipeline)	✓	✓	✓	✓	**24.8**	**25.7**

Table 2. Ablation study on different global priors.

4.4. Ablation Studies

To show the efficacy of our design choices, we conduct exhaustive ablation experiments on the HouseCat6D dataset.
Effects of Global Context Priors. We evaluate the effect of global context priors on the pose estimation backbone by comparing different global prior configurations, including: *(0) Baseline.* AG-Pose with the DINOv2 backbone, excluding SSR and GCE fusion modules (no global shape or semantic priors). *(1) Instance geometric prior.* Geometric features alone are passed into the GCE fusion module, omitting semantic information. *(2) Categorical semantic prior.* The shape reconstruction module is excluded, and the categorical semantic prior is fused with local features. *(3) Mean shape geometric & Categorical semantic prior.* The SSR module applies mean shape reconstruction and categorical semantic features. *(4) GCE-Pose (Full pipeline).* Our full method combines instance-specific geometric and categorical semantic priors. As shown in Table 2, *(1)* yields a 4% improvement on the 5°2cm metric over the baseline, *(2)* yields 7%, and *(3)* achieves 10%. Our method (5), combining instance geometric and semantic priors as global context guidance, reaches the highest performance and surpasses the baseline by 16%.
Robustness of Semantic Shape Reconstruction. Qualitative results of our semantic shape reconstruction are shown in Figure 5. Specifically, (a) shows partial input points and their reconstructed points with semantic features, illustrating the robustness of our SSR module to noisy and occluded scenarios. (b) displays the categorical semantic prototype

Fusion features	5°2cm	5°5cm	10°2cm	10°5cm	IoU50	IoU75
Value feature	21.9	23.2	48.9	53.1	76.0	55.2
Key feature (Ours)	**24.8**	**25.7**	**55.4**	**58.4**	**79.2**	**60.6**

Table 3. Ablation study on different feature fusion strategies.

where the 3D semantics are lifted from DINO features. We also show the instances where the semantics are aggregated from our Deep Linear Semantic Shape Models. We visualize the semantic features by sharing the PCA centers, demonstrating our semantic feature aggregation is robust against shape variance.
Effects of the GCE Feature Fusion Module. In Table 3, we evaluate our GCE feature fusion by experimenting with different DINOv2 tokens. Our results indicate that using the key tokens to aggregate the global semantic prior yields better pose estimation performance than using value tokens. We attribute this improvement to the design of our cross-attention layer where the local features serve as the query and global semantic prior as the key and value. This setup allows the global semantic prior to initialize the attentional weights between local and global cues effectively, enhancing the feature fusion for pose estimation.

5. Conclusions and Limitations

We propose GCE-Pose, a category-level pose estimation method that leverages category-specific shape and semantic priors to improve pose learning. Our approach achieves state-of-the-art results on two challenging datasets through a Semantic Shape Reconstruction (SSR) module and a Global Context Enhanced (GCE) feature fusion module. GCE-Pose reconstructs partial inputs into complete semantic shapes, providing robust prior guidance for pose estimation. Although our semantic prior is limited to specific categories without scaling to instance-level alignment, our semantic shape reconstruction strategy still performs accurate pose estimation on category-level real data, even under occlusions and sensor artifacts, moving towards application-ready accuracy in category-level pose estimation.

References

[1] Philipp Ausserlechner, David Haberger, Stefan Thalhammer, Jean-Baptiste Weibel, and Markus Vincze. Zs6d: Zero-shot 6d object pose estimation using vision transformers. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 463–469. IEEE, 2024. 1, 2

[2] Dingding Cai, Janne Heikkilä, and Esa Rahtu. Gs-pose: Cascaded framework for generalizable segmentation-based 6d object pose estimation. *arXiv preprint arXiv:2403.10683*, 2024. 2

[3] Andrea Caraffa, Davide Boscaini, Amir Hamza, and Fabio Poiesi. Freeze: Training-free zero-shot 6d pose estimation with geometric and vision foundation models. In *European Conference on Computer Vision*, pages 414–431. Springer, 2025. 1, 2

[4] 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. 3

[5] Kai Chen and Qi Dou. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2773–2782, 2021. 2

[6] 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. 3

[7] 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, 7

[8] 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. 3

[9] 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. 3

[10] 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. 3, 7

[11] Francesco Di Felice, Alberto Remus, Stefano Gasperini, Benjamin Busam, Lionel Ott, Federico Tombari, Roland Siegwart, and Carlo Alberto Avizzano. Zero123-6d: Zero-shot novel view synthesis for rgb category-level 6d pose estimation. *arXiv preprint arXiv:2403.14279*, 2024. 2

[12] Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 998–1005. Ieee, 2010. 1

[13] Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In *CVPR*, 2010. 2

[14] Zhaoxin Fan, Zhengbo Song, Jian Xu, Zhicheng Wang, Kejian Wu, Hongyan Liu, and Jun He. Acr-pose: Adversarial canonical representation reconstruction network for category level 6d object pose estimation, 2021. 2

[15] Zhiwen Fan, Panwang Pan, Peihao Wang, Yifan Jiang, Dejia Xu, and Zhangyang Wang. Pope: 6-dof promptable pose estimation of any object in any scene with one reference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7771–7781, 2024. 3

[16] 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. 2

[17] Daoyi Gao, Yitong Li, Patrick Ruhkamp, Iuliia Skobleva, Magdalena Wysocki, HyunJun Jung, Pengyuan Wang, Arturo Guridi, and Benjamin Busam. Polarimetric pose prediction. In *European Conference on Computer Vision*, pages 735–752. Springer, 2022. 3

[18] Walter Goodwin, Sagar Vaze, Ioannis Havoutis, and Ingmar Posner. Zero-shot category-level object pose estimation. In *European Conference on Computer Vision*, pages 516–532. Springer, 2022. 2

[19] Rasmus Laurvig Haugaard, Frederik Hagelskjær, and Thorbjørn Mosekjær Iversen. Spyropose: Se (3) pyramids for object pose distribution estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2082–2091, 2023. 2

[20] 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. 2

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

[22] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3003–3013, 2021. 2

[23] Tomas Hodan, Martin Sundermeyer, Yann Labbe, Van Nguyen Nguyen, Gu Wang, Eric Brachmann, Bertram Drost, Vincent Lepetit, Carsten Rother, and Jiri Matas. Bop challenge 2023 on detection segmentation and pose estimation of seen and unseen rigid objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5610–5619, 2024. 1

[24] Junwen Huang, Hao Yu, Kuan-Ting Yu, Nassir Navab, Slobodan Ilic, and Benjamin Busam. Matchu: Matching unseen

objects for 6d pose estimation from rgb-d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10095–10105, 2024. 1, 2

[25] Muhammad Zubair Irshad, Thomas Kollar, Michael Laskey, Kevin Stone, and Zsolt Kira. Centersnap: Single-shot multi-object 3d shape reconstruction and categorical 6d pose and size estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 10632–10640. IEEE, 2022. 4

[26] Muhammad Zubair Irshad, Sergey Zakharov, Rares Ambrus, Thomas Kollar, Zsolt Kira, and Adrien Gaidon. Shapo: Implicit representations for multi-object shape, appearance, and pose optimization. In *European Conference on Computer Vision*, pages 275–292. Springer, 2022. 4

[27] 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. 3, 6, 7

[28] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the IEEE international conference on computer vision*, pages 1521–1529, 2017. 2

[29] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 6

[31] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *ECCV*, 2020. 2

[32] 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. 2

[33] Fu Li, Shishir Reddy Vutukur, Hao Yu, Ivan Shugurov, Benjamin Busam, Shaowu Yang, and Slobodan Ilic. Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2123–2133, 2023. 2

[34] Haitao Lin, Zichang Liu, Chilam Cheang, Yanwei Fu, Guodong Guo, and Xiangyang Xue. Sar-net: Shape alignment and recovery network for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6707–6717, 2022. 2

[35] 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. 3

[36] 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. 1, 2, 3, 6, 7

[37] 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. 3, 7

[38] 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. 1, 2

[39] 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. 1, 3, 4, 7

[40] Jianhui Liu, Yukang Chen, Xiaoqing Ye, and Xiaojuan Qi. Ist-net: Prior-free category-level pose estimation with implicit space transformation, 2023. 3

[41] Romain Loiseau, Tom Monnier, Mathieu Aubry, and Loïc Landrieu. Representing shape collections with alignment-aware linear models, 2021. 4

[42] Fabian Manhardt, Diego Martin Arroyo, Christian Rupprecht, Benjamin Busam, Tolga Birdal, Nassir Navab, and Federico Tombari. Explaining the ambiguity of object detection and 6d pose from visual data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6841–6850, 2019. 2

[43] Fabian Manhardt, Gu Wang, Benjamin Busam, Manuel Nickel, Sven Meier, Luca Minciullo, Xiangyang Ji, and Nassir Navab. Cps++: Improving class-level 6d pose and shape estimation from monocular images with self-supervised learning. *arXiv preprint arXiv:2003.05848*, 2020. 3

[44] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Yinlin Hu, Renaud Marlet, Mathieu Salzmann, and Vincent Lepetit. Nope: Novel object pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17923–17932, 2024. 3

[45] 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

[46] 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. 1, 2, 4, 5

[47] 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, 2

[48] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4561–4570, 2019. 2

[49] 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. 4, 6

[50] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, Slobodan Ilic, Dewen Hu, and Kai Xu. Geotransformer: Fast and robust point cloud registration with geometric transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9806–9821, 2023. 2

[51] 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. 2

[52] Patrick Ruhkamp, Daoyi Gao, Nassir Navab, and Benjamin Busam. S2 p3: Self-supervised polarimetric pose prediction. *International Journal of Computer Vision*, 132(6): 2177–2194, 2024. 3

[53] Yifei Shi, Junwen Huang, Xin Xu, Yifan Zhang, and Kai Xu. Stablepose: Learning 6d object poses from geometrically stable patches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15222–15231, 2021. 2

[54] 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. 1

[55] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017. 6

[56] Yongzhi Su, Mahdi Saleh, Torben Fetzer, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari. Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6738–6748, 2022. 1

[57] 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. 2

[58] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *Proceedings of the european conference on computer vision (ECCV)*, pages 699–715, 2018. 2

[59] 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. 2, 4

[60] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04): 376–380, 1991. 1, 3

[61] Shishir Reddy Vutukur, Rasmus Laurvig Haugaard, Junwen Huang, Benjamin Busam, and Tolga Birdal. Alignist: Cad-informed orientation distribution estimation by fusing shape and correspondences. In *European Conference on Computer Vision*, pages 351–369. Springer, 2025. 2

[62] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3343–3352, 2019. 2

[63] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. GDR-Net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16611–16621, 2021. 1

[64] 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, 3, 6, 7

[65] Jiaze Wang, Kai Chen, and Qi Dou. Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4807–4814. IEEE, 2021. 2

[66] Pengyuan Wang, Lorenzo Garattoni, Sven Meier, Nassir Navab, and Benjamin Busam. Crocps: Addressing photometric challenges in self-supervised category-level 6d object poses with cross-modal learning. In *BMVC*, page 390, 2022. 3

[67] Pengyuan Wang, HyunJun Jung, Yitong Li, Siyuan Shen, Rahul Parthasarathy Srikanth, Lorenzo Garattoni, Sven Meier, Nassir Navab, and Benjamin Busam. Phocal: A multi-modal dataset for category-level object pose estimation with photometrically challenging objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21222–21231, 2022. 3

[68] Pengyuan Wang, Lorenzo Garattoni, Sven Meier, Nassir Navab, and Benjamin Busam. Improving self-supervised learning of transparent category poses with language guidance and implicit physical constraints. *IEEE Robotics and Automation Letters*, 2024. 3

[69] 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, 7

11

[70] Ruiqi Wang, Xinggang Wang, Te Li, Rong Yang, Minhong Wan, and Wenyu Liu. Query6dof: Learning sparse queries as implicit shape prior for category-level 6dof pose estimation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14009–14018, 2023. 2, 7

[71] 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

[72] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 2

[73] Chao Xu, Ang Li, Linghao Chen, Yulin Liu, Ruoxi Shi, Hao Su, and Minghua Liu. Sparp: Fast 3d object reconstruction and pose estimation from sparse views. In *European Conference on Computer Vision*, pages 143–163. Springer, 2025. 3

[74] Hao Yu, Zheng Qin, Ji Hou, Mahdi Saleh, Dongsheng Li, Benjamin Busam, and Slobodan Ilic. Rotation-invariant transformer for point cloud matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5384–5393, 2023. 2

[75] Hao Yu, Ji Hou, Zheng Qin, Mahdi Saleh, Ivan Shugurov, Kai Wang, Benjamin Busam, and Slobodan Ilic. Riga: Rotation-invariant and globally-aware descriptors for point cloud registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[76] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1941–1950, 2019. 2

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

[78] Ruida Zhang, Yan Di, Zhiqiang Lou, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Rbp-pose: Residual bounding box projection for category-level pose estimation. In *European Conference on Computer Vision*, pages 655–672. Springer, 2022. 2, 7

[79] 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. 3, 6

[80] 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. 6