ComPose: A Unified Completion-Pose Framework for Robust Category-Level Object Pose Estimation

Huan Ren[1,2] Yihan Chen[1,2] Chuxin Wang[1,2] Nailong Liu[3] Wenfei Yang[1,2*] Tianzhu Zhang[1,2]
[1]University of Science and Technology of China
[2]National Key Laboratory of Deep Space Exploration, Deep Space Exploration Laboratory
[3]Beijing Institute of Control Engineering
{rh_hr_666,yihanchen,wcx0602}@mail.ustc.edu.cn {yangwf,tzzhang}@ustc.edu.cn

Abstract

Category-level object pose estimation aims to predict the pose and size of arbitrary objects in specific categories. Existing methods struggle with the inherent incompleteness of observed point clouds, which limits their ability to capture complete object shapes for robust pose reasoning. While point cloud completion offers a promising solution, naively treating it as a separate preprocessing step for partial observations introduces compounding errors and additional computational overhead, ultimately hindering both accuracy and efficiency. To address these challenges, we propose ComPose, a novel unified framework that tightly integrates shape completion to provide complete geometric cues for enhanced pose estimation. At the core of ComPose is a keypoint-based progressive completion module, which recovers full shape representations by progressively predicting a sparse set of keypoints and their surrounding dense point sets, empowering the keypoints to capture holistic object geometries. A geometric relation encoding module further enriches keypoint features with both local and global geometric context. In addition, we introduce a novel geometric relation consistency loss to enforce structural alignment between observed keypoints and their predicted NOCS coordinates, ensuring globally coherent coordinate transformations. Extensive experiments on standard benchmarks demonstrate that our method outperforms state-of-the-art approaches without relying on category-level shape priors.

1. Introduction

Category-level object pose estimation aims to predict the 6D pose and 3D size of arbitrary objects within predefined categories. As a fundamental task in 3D computer vision, it has attracted significant attention from the research community due to its extensive potential applications in fields such as robotic manipulation [18], augmented reality [30], and autonomous driving [20]. In contrast to traditional instance-level approaches [25, 34, 35], category-level methods do not require instance-specific CAD models for inference, exhibiting stronger generalization in real-world scenarios.

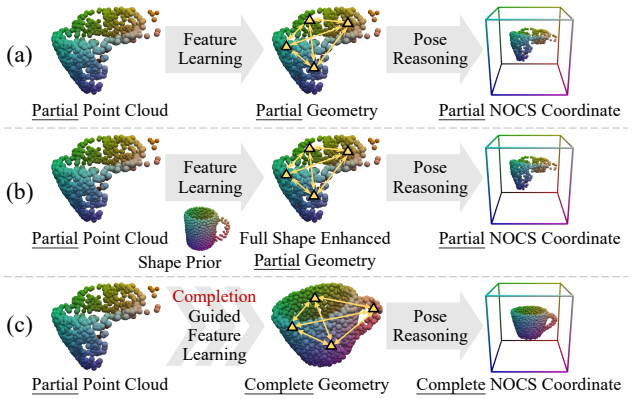

Figure 1. Comparison of geometric representation strategies in category-level object pose estimation. (a) Classic methods directly encode geometric features from partial point clouds, which limits their ability to capture complete object structures. (b) Prior-based approaches resort to category-level shape priors [31] to enhance feature understanding of full object shapes, yet they still operate on incomplete geometries. (c) Our method explicitly integrates shape completion to recover complete geometries, facilitating more comprehensive and robust pose reasoning.

Existing category-level methods typically commence by extracting features from partial observations to comprehend object shapes, which subsequently guides either direct pose regression or the prediction of Normalized Object Coordinate Space (NOCS) [36] as an intermediate representation for pose fitting. Despite the considerable progress achieved by these approaches, their performance remains fundamentally constrained by the inherent incompleteness of partially observed point clouds. Specifically, depth cameras fail to capture the occluded backside of objects due to self-occlusion, yielding incomplete point clouds after back-

*Corresponding author. Website: renhuan1999.github.io/ComPose.

projection. As illustrated in Figure 1(a), most previous methods [5, 15, 16, 29] encode geometric structures directly from such partial point clouds, which restricts their ability to capture complete object shapes for robust pose reasoning. To alleviate this limitation, several works [10, 13] incorporate category-level shape priors [31] to enhance the comprehension of full shape context at the feature level, as shown in Figure 1(b). Nevertheless, these methods only provide indirect shape cues *in the canonical space* while still operate on intrinsically incomplete shape representations *in the observation space*, leaving the fundamental issue of geometric incompleteness unresolved. Moreover, acquiring shape priors requires collecting extensive CAD models and training an extra autoencoder, which is labor-intensive and costly.

To address these challenges, we draw inspiration from recent advancements in point cloud completion [40, 41] and explore reconstructing complete object shapes directly in the observation space. As illustrated in Figure 1(c), the completed point clouds provide a more comprehensive geometric representation of objects, which is crucial for robust pose reasoning. To quantitatively assess the merit of complete geometries, we conduct an oracle experiment by replacing the *partial point cloud* inputs in the depth-only version of the leading AG-Pose [15] network with *ground-truth complete point clouds*, while keeping the network architecture unchanged. As shown in Figure 2, the 10°2cm accuracy improves dramatically from 68.5% to 91.7%, highlighting the significant upper-bound performance gain enabled by full shape information. This finding naturally suggests a straightforward solution that employs a point cloud completion network to first recover the full shape and then feeds it into a standard pose estimator, such as AG-Pose. However, such a naive two-stage pipeline suffers from compounding errors and introduces additional computational overhead, which compromise both accuracy and efficiency. As shown in Figure 2, even an end-to-end jointly optimized variant yields only a marginal improvement of the 10°2cm accuracy to 71.0% while reducing inference speed from 33.5 FPS to 21.5 FPS. This indicates that simply cascading completion and pose estimation networks falls short of fully exploiting the potential of shape completion and incurs a notable efficiency trade-off. These observations thus raise a pivotal question: *how can we effectively and efficiently integrate the complete geometric cues recovered from point cloud completion to enhance object pose estimation?*

Motivated by the above discussions, we propose **ComPose**, a novel category-level framework that seamlessly unifies point cloud **com**pletion and object **pose** estimation within a single network. Unlike naive cascaded pipelines that treat completion as a separate preprocessing step, ComPose tightly integrates completion as a task-driven internal component, thereby enhancing the comprehension of complete object shapes in a more effective and efficient manner.

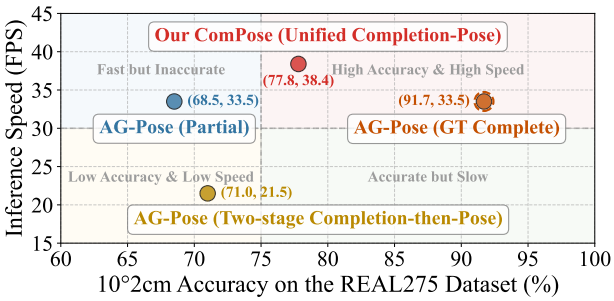

Figure 2. Accuracy and inference speed comparison for the depth-only versions of different methods. The dashed circle indicates the performance upper bound achieved using ground-truth complete point clouds as input. Our ComPose achieves the best balance between accuracy and efficiency with 38.4 FPS on an RTX3090Ti GPU. More implementation details are provided in Section 4.3.

At the core of our framework is a keypoint-based progressive completion module that reconstructs full object shapes from partial observations by predicting a sparse set of complete keypoints along with their surrounding dense point sets. This design not only yields comprehensive shape representations in the observation space, but also empowers keypoints to capture complete geometries from a holistic perspective. Building on this, we incorporate a geometric relation encoding module to enrich keypoint features with both local and global geometric context, which facilitates keypoint-wise prediction of NOCS coordinates. To further ensure globally coherent coordinate transformations, we introduce a geometric relation consistency loss grounded in relational modeling [27, 38], which enforces structural alignment between observed keypoints and their predicted NOCS counterparts. In contrast to conventional point-to-point coordinate mapping losses, the proposed relation-based constraint captures higher-order structural cues, resulting in more robust and precise object pose estimation.

In summary, the contributions of this work are fourfold: (1) We pioneer a novel paradigm that leverages the complete geometric cues recovered from point cloud completion to enhance the understanding of holistic object shapes, unlocking substantial potential for advanced pose reasoning. (2) We propose a unified framework that seamlessly integrates shape completion and pose estimation into a single network, delivering an effective and efficient solution for category-level object pose estimation. (3) We introduce a carefully designed approach that acquires complete shape representations through keypoint-based progressive completion, and further incorporates geometric relation encoding and consistency constraints for robust coordinate transformations. (4) Extensive experimental results demonstrate the superior performance of our method over state-of-the-art approaches. Notably, our depth-only model achieves a significant 9.1% improvement on the 10°2cm metric of the REAL275 dataset without relying on shape priors.

2. Related Works

2.1. Category-level Object Pose Estimation

Category-level object pose estimation aims to determine the 6D pose and 3D size of novel objects in specific categories, without requiring instance-specific CAD models at test time. A core challenge in this field lies in understanding object shapes with significant intra-class variations for robust pose reasoning, especially from partial and incomplete observations. One class of works [2, 12] directly regresses object poses from point cloud data by extracting discriminative geometric features, often incorporating geometry-guided constraints [5] or hybrid-scope geometric perception [44]. Another line of methods [17, 29] leverages Normalized Object Coordinate Space (NOCS) [36] as a shared canonical representation to align various object instances, which enables the establishment of dense point-wise [13] or sparse keypoint-wise [15] correspondences between camera and object coordinate spaces for subsequent pose fitting. However, both types of methods are inherently constrained by the incomplete nature of partial point clouds, which limits their ability to capture complete object shapes. To alleviate this, prior-based methods [1, 10] incorporate category-level shape priors [31] to enhance the feature-level perception of full shape context. For instance, SPD [31] predicts deformation fields from shape priors to reconstruct 3D object models in the canonical space, which in turn guides the prediction of NOCS coordinates from partial point clouds. Nevertheless, these approaches still operate on incomplete shape representations in the observation space, leaving the fundamental issue of geometric incompleteness unresolved. In contrast, our ComPose explicitly integrates shape completion to recover complete shape representations directly within the observation space, enabling the network to capture holistic object geometries and facilitating more robust pose reasoning without relying on external shape priors.

2.2. Point Cloud Completion

Point cloud completion aims to reconstruct the complete 3D shape from partial point clouds, a naturally arising problem in real-world scenarios where sensor data is often incomplete due to inevitable object self-occlusion. Early methods typically adopt an encoder-decoder architecture, where a global shape representation is extracted from the visible part and decoded into the complete point cloud. For instance, FoldingNet [39] proposes a folding-based decoder that deforms a 2D grid onto the 3D surface, while PCN [42] introduces a coarse-to-fine decoding strategy to progressively recover fine-grained geometry. However, these approaches primarily focus on global shape embeddings, which limits their capacity to explicitly model the interactions between visible and missing regions. To deal with this limitation, recent approaches such as PoinTr [40] reformulate point cloud completion as a set-to-set translation problem, leveraging Transformer-based attention mechanisms [33] to model long-range dependencies. AdaPoinTr [41] further improves upon this by introducing an adaptive query generation strategy and an auxiliary denoising task to boost completion quality. Despite these advances, most methods treat point cloud completion as a standalone task, with limited exploration of its integration into downstream applications such as category-level object pose estimation. A notable attempt is DR-Pose [45], which leverages an off-the-shelf completion network [40] to recover the missing object parts. However, the completed shapes are merely used to guide the deformation of shape priors as in SPD [31] and remain decoupled from the actual pose reasoning process. In contrast, we tightly integrate completion as an internal component of our unified ComPose framework to enhance the understanding of complete object shapes for robust pose reasoning.

3. Method

Task Definition. Given an RGB-D image, an off-the-shelf Mask R-CNN [7] network is first used to obtain instance masks, yielding the cropped RGB image $\boldsymbol{I}^{\text{rgb}} \in \mathbb{R}^{H\times W\times 3}$ and segmented depth image for each instance. The partial point cloud $\boldsymbol{P}^{\text{part}} \in \mathbb{R}^{N^{\text{part}}\times 3}$ is then derived by back-projecting and downsampling the segmented depth image, where N^{part} denotes the number of points. Taking $\boldsymbol{I}^{\text{rgb}}$ and $\boldsymbol{P}^{\text{part}}$ as inputs, our ComPose predicts the 3D rotation $\boldsymbol{R} \in \text{SO}(3)$, 3D translation $\boldsymbol{t} \in \mathbb{R}^3$, and 3D size $\boldsymbol{s} \in \mathbb{R}^3$ of the observed object instance within predefined categories.

Overview. As illustrated in Figure 3, the proposed framework consists of four components: partial feature extraction (Section 3.1), keypoint-based progressive completion (Section 3.2), geometric relation encoding (Section 3.3), and correspondence-based pose estimation (Section 3.4).

3.1. Partial Feature Extraction

For the partial point cloud $\boldsymbol{P}^{\text{part}}$, we adopt PointNet++ [26] to extract point-wise geometric features $\boldsymbol{F}^{\text{pn}}$, which serve as the initial representation $\boldsymbol{F}^{\text{init}} \in \mathbb{R}^{N^{\text{part}}\times D}$. Under the RGB-D setting, we further follow SecondPose [3] by employing DINOv2 [21] to extract pose-consistent semantic features $\boldsymbol{F}^{\text{dino}}$ from the RGB image $\boldsymbol{I}^{\text{rgb}}$. These semantic features are associated [29] with the 3D points in $\boldsymbol{P}^{\text{part}}$, concatenated with the geometric features $\boldsymbol{F}^{\text{pn}}$, and then projected into a D-dimensional space to form the initial feature representation $\boldsymbol{F}^{\text{init}} \in \mathbb{R}^{N^{\text{part}}\times D}$. To better capture global context, a stack of Self-Attention (SA) [33] layers are applied to $\boldsymbol{F}^{\text{init}}$, enabling dynamic interactions among all points and yielding the refined partial representation $\boldsymbol{F}^{\text{part}} \in \mathbb{R}^{N^{\text{part}}\times D}$, which is formulated as below:

$$\boldsymbol{F}^{\text{part}} = \text{SA}(\boldsymbol{F}^{\text{init}} + \text{PE}(\boldsymbol{P}^{\text{part}})), \tag{1}$$

$$\text{SA}(\boldsymbol{Q}) = \phi((\boldsymbol{Q}\boldsymbol{W}^{Q})(\boldsymbol{Q}\boldsymbol{W}^{K})^{\top}/\sqrt{D})(\boldsymbol{Q}\boldsymbol{W}^{V}), \tag{2}$$

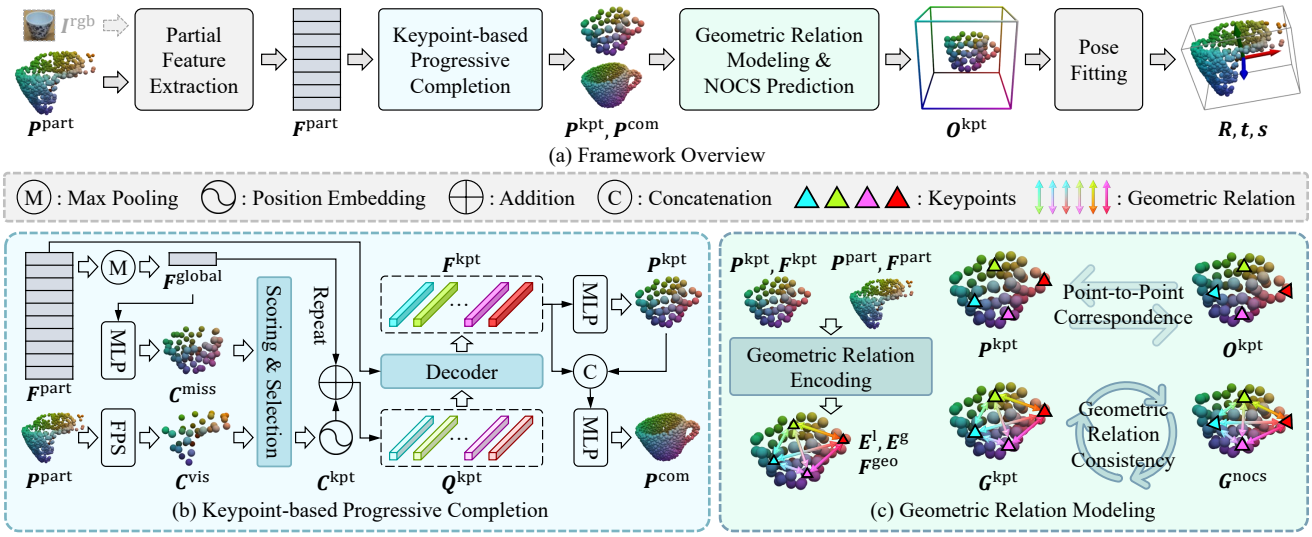

Figure 3. (a) Overview of the proposed ComPose framework, which supports both RGB-D and depth-only settings, where the latter omits the RGB images $\boldsymbol{I}^{\text{rgb}}$. (b) The initial coarse keypoints $\boldsymbol{C}^{\text{kpt}}$ are adaptively selected from missing and visible candidates $\{\boldsymbol{C}^{\text{miss}}, \boldsymbol{C}^{\text{vis}}\}$. These coarse keypoints are then progressively refined through feature interactions with the partial features $\boldsymbol{F}^{\text{part}}$ to recover complete object geometries, including refined keypoints $\boldsymbol{P}^{\text{kpt}}$ and dense shapes $\boldsymbol{P}^{\text{com}}$. (c) The keypoint features $\boldsymbol{F}^{\text{kpt}}$ are enhanced into $\boldsymbol{F}^{\text{geo}}$ via geometric relation encoding, incorporating both local and global geometric context $\{\boldsymbol{E}^{\text{l}}, \boldsymbol{E}^{\text{g}}\}$. To ensure robust coordinate transformations, the pairwise geometric relations $\boldsymbol{G}$ among keypoints are constrained to maintain alignment between the observation and canonical spaces.

where $\boldsymbol{W}^{Q/K/V} \in \mathbb{R}^{D\times D}$ are projection matrices, ϕ and $\top$ denote the Softmax and transpose operations, respectively. PE represents the learnable position embedding [33], which provides explicit spatial guidance for feature interactions.

3.2. Keypoint-based Progressive Completion

After extracting features from partial observations, we draw inspiration from [41] and design a keypoint-based progressive completion module tailored for object pose estimation. Unlike typical point cloud completion approaches [39, 42] that assume canonical object alignments, our method handles the more challenging scenario where partial shapes are observed under arbitrary poses. The progressive process begins by generating an initial set of coarse keypoints $\boldsymbol{C}^{\text{kpt}} \in \mathbb{R}^{N^{\text{kpt}}\times 3}$, where N^{kpt} represents the number of keypoints. These keypoints are then progressively refined through feature interactions to recover complete object geometries, including both sparse keypoints $\boldsymbol{P}^{\text{kpt}} \in \mathbb{R}^{N^{\text{kpt}}\times 3}$ and dense shapes $\boldsymbol{P}^{\text{com}} \in \mathbb{R}^{N^{\text{com}}\times 3}$ with N^{com} points. This module enables a more comprehensive understanding of object shapes, which is crucial for robust pose reasoning.

Coarse Keypoint Generation. Given the partial features $\boldsymbol{F}^{\text{part}}$, we first apply global max pooling to obtain a global representation $\boldsymbol{f}^{\text{global}} \in \mathbb{R}^{D}$, followed by an MLP to predict a set of coarse keypoints $\boldsymbol{C}^{\text{miss}} \in \mathbb{R}^{N^{\text{miss}}\times 3}$ that indicate potentially missing regions. Meanwhile, Farthest Point Sampling (FPS) is applied to the partial point cloud $\boldsymbol{P}^{\text{part}}$ to acquire a set of visible keypoints $\boldsymbol{C}^{\text{vis}} \in \mathbb{R}^{N^{\text{vis}}\times 3}$, providing reliable geometric cues from the visible regions. These two sets of keypoints contribute to the construction of the initial coarse keypoints $\boldsymbol{C}^{\text{kpt}}$. Nevertheless, due to varying levels of incompleteness across different observations, relying on a fixed ratio of missing and visible keypoints lacks flexibility. To adaptively select representative keypoints, we concatenate $\boldsymbol{C}^{\text{miss}}$ and $\boldsymbol{C}^{\text{vis}}$ to form the candidate set $\boldsymbol{C}^{\text{cand}} \in \mathbb{R}^{N^{\text{cand}}\times 3}$, which is then fed into a scoring MLP to predict scores $\boldsymbol{r} \in \mathbb{R}^{N^{\text{cand}}}$ for each candidate. The top N^{kpt} keypoints are retained as $\boldsymbol{C}^{\text{kpt}}$ after ranking, enabling a flexible balance between missing and visible regions.

Progressive Shape Completion. To further refine the initially generated coarse keypoints $\boldsymbol{C}^{\text{kpt}}$ and achieve higher fidelity in shape recovery, we employ a Transformer [33] decoder that facilitates fine-grained feature interactions with observations. Specifically, we first construct the keypoint queries $\boldsymbol{Q}^{\text{kpt}} \in \mathbb{R}^{N^{\text{kpt}}\times D}$ by fusing the position embedding of $\boldsymbol{C}^{\text{kpt}}$ with the global feature $\boldsymbol{f}^{\text{global}}$, written as:

$$\boldsymbol{Q}^{\text{kpt}} = \text{Repeat}(\boldsymbol{f}^{\text{global}}) + \text{PE}(\boldsymbol{C}^{\text{kpt}}). \tag{3}$$

The position embedding of $\boldsymbol{C}^{\text{kpt}}$ provides explicit spatial guidance for the subsequent feature interactions with the partial feature $\boldsymbol{F}^{\text{part}}$ through a decoder composed of Cross-Attention (CA) and Self-Attention (SA) layers, yielding the refined keypoint features $\boldsymbol{F}^{\text{kpt}} \in \mathbb{R}^{N^{\text{kpt}}\times D}$, formulated as:

$$\hat{\boldsymbol{F}}^{\text{kpt}} = \text{CA}(\boldsymbol{Q}^{\text{kpt}}, \boldsymbol{F}^{\text{part}}), \quad \boldsymbol{F}^{\text{kpt}} = \text{SA}(\hat{\boldsymbol{F}}^{\text{kpt}}), \tag{4}$$

$$\text{CA}(\boldsymbol{Q}, \boldsymbol{K}) = \phi((\boldsymbol{Q}\boldsymbol{W}^{Q})(\boldsymbol{K}\boldsymbol{W}^{K})^{\top}/\sqrt{D})(\boldsymbol{K}\boldsymbol{W}^{V}). \tag{5}$$

These features are then passed through an MLP to directly predict the 3D keypoint coordinates $\boldsymbol{P}^{\text{kpt}}$. To reconstruct a

dense and complete point cloud $\boldsymbol{P}^{\text{com}}$, each keypoint feature $\boldsymbol{F}_n^{\text{kpt}}$ is concatenated with its associated coordinate $\boldsymbol{P}_n^{\text{kpt}}$ and passed through an MLP, whose output is reshaped to form the fine-grained local geometry $\boldsymbol{P}_n^{\text{fold}} \in \mathbb{R}^{N^{\text{fold}} \times 3}$ around each keypoint. The final complete dense point cloud $\boldsymbol{P}^{\text{com}} \in \mathbb{R}^{N^{\text{com}} \times 3}$, consisting of $N^{\text{com}} = N^{\text{kpt}} N^{\text{fold}}$ points, is formed by aggregating all $\boldsymbol{P}_n^{\text{fold}}$ outputs.

Supervision and Loss Functions. During training, to supervise the learning of progressive completion, the CAD model $\boldsymbol{M}^{\text{cad}}$ of the object is first transformed into the observation space using the ground-truth pose parameters $\{\boldsymbol{R}^{\text{gt}}, \boldsymbol{t}^{\text{gt}}, \boldsymbol{s}^{\text{gt}}\}$, resulting in the transformed model $\boldsymbol{M}^{\text{obs}}$:

$$\boldsymbol{M}^{\text{obs}} = \|\boldsymbol{s}^{\text{gt}}\|_2 \boldsymbol{R}^{\text{gt}} \boldsymbol{M}^{\text{cad}} + \boldsymbol{t}^{\text{gt}}. \tag{6}$$

The progressive completion loss is then computed using the Chamfer Distance [6] between the transformed model $\boldsymbol{M}^{\text{obs}}$ and the set $\mathbb{S} = \{\boldsymbol{C}^{\text{miss}}, \boldsymbol{P}^{\text{kpt}}, \boldsymbol{P}^{\text{com}}\}$, defined as:

$$\mathcal{L}^{\text{com}} = \sum\nolimits_{\boldsymbol{P}^{\text{sup}} \in \mathbb{S}} \text{CD}(\boldsymbol{P}^{\text{sup}}, \boldsymbol{M}^{\text{obs}}), \tag{7}$$

$$\begin{aligned} \text{CD}(\boldsymbol{X}, \boldsymbol{Y}) = &\frac{1}{|\boldsymbol{X}|} \sum\nolimits_{\boldsymbol{x} \in \boldsymbol{X}} \min_{\boldsymbol{y} \in \boldsymbol{Y}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \\ &+ \frac{1}{|\boldsymbol{Y}|} \sum\nolimits_{\boldsymbol{y} \in \boldsymbol{Y}} \min_{\boldsymbol{x} \in \boldsymbol{X}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2. \end{aligned} \tag{8}$$

During the selection of representative keypoints, we aim to retain those that are close to the transformed model $\boldsymbol{M}^{\text{obs}}$ while filtering out outlier predictions in $\boldsymbol{C}^{\text{miss}}$ and outlier points in $\boldsymbol{C}^{\text{vis}}$ [29]. To achieve this, we employ an MSE loss function on the predicted keypoint scores $\boldsymbol{r}$, defined as:

$$\mathcal{L}^{\text{score}} = \frac{1}{N^{\text{cand}}} \sum\nolimits_n (\boldsymbol{r}_n - \boldsymbol{r}_n^{\text{gt}})^2, \tag{9}$$

$$\boldsymbol{r}_n^{\text{gt}} = \exp(-\boldsymbol{d}_n / \tau), \quad \boldsymbol{d}_n = \min_{\boldsymbol{y} \in \boldsymbol{M}^{\text{obs}}} \|\boldsymbol{C}_n^{\text{cand}} - \boldsymbol{y}\|_2, \tag{10}$$

where $\tau = 0.05$ denotes the temperature hyper-parameter.

3.3. Geometric Relation Encoding

To further enhance the geometric contextual modeling of keypoint features, we follow AG-Pose [15] to explicitly encode their surrounding geometric relations. Specifically, for the n-th keypoint $\boldsymbol{P}_n^{\text{kpt}}$, its N^{knn} nearest neighbors $\boldsymbol{P}_n^{\text{knn}} \in \mathbb{R}^{N^{\text{knn}} \times 3}$ are first identified from $\boldsymbol{P}^{\text{part}}$, and their corresponding features $\boldsymbol{F}_n^{\text{knn}} \in \mathbb{R}^{N^{\text{knn}} \times D}$ are retrieved from $\boldsymbol{F}^{\text{part}}$. We then compute both local and global geometric relation embeddings, denoted as $\boldsymbol{E}_n^{\text{l}} \in \mathbb{R}^{N^{\text{knn}} \times D}$ and $\boldsymbol{E}_n^{\text{g}} \in \mathbb{R}^{N^{\text{kpt}} \times D}$, respectively, as formulated below:

$$\boldsymbol{E}_n^{\text{l}} = \text{MLP}(\text{Repeat}(\boldsymbol{P}_n^{\text{kpt}}) - \boldsymbol{P}_n^{\text{knn}}), \tag{11}$$

$$\boldsymbol{E}_n^{\text{g}} = \text{MLP}(\text{Repeat}(\boldsymbol{P}_n^{\text{kpt}}) - \boldsymbol{P}^{\text{kpt}}). \tag{12}$$

Subsequently, the keypoint features $\boldsymbol{F}^{\text{kpt}}$ are progressively enriched with local and global geometric context through the following alternating enhancement process, yielding the geometric-aware keypoint features $\boldsymbol{F}^{\text{geo}} \in \mathbb{R}^{N^{\text{kpt}} \times D}$:

$$\hat{\boldsymbol{F}}_n^{\text{kpt}} = \text{CA}(\boldsymbol{F}_n^{\text{kpt}}, \text{MLP}(\boldsymbol{F}_n^{\text{knn}} + \boldsymbol{E}_n^{\text{l}})), \tag{13}$$

$$\begin{aligned} \boldsymbol{F}_n^{\text{geo}} = \text{MLP}(&\hat{\boldsymbol{F}}_n^{\text{kpt}} + \text{AvgPool}(\hat{\boldsymbol{F}}^{\text{kpt}}) \\ &+ \text{PE}(\boldsymbol{P}_n^{\text{kpt}}) + \text{AvgPool}(\boldsymbol{E}_n^{\text{g}})). \end{aligned} \tag{14}$$

3.4. Correspondence-based Pose Estimation

We adopt a correspondence-based paradigm [29] for object pose estimation. Specifically, given the geometry-enhanced keypoint features $\boldsymbol{F}^{\text{geo}}$, an MLP is employed to predict the corresponding NOCS coordinates $\boldsymbol{O}^{\text{kpt}} \in \mathbb{R}^{N^{\text{kpt}} \times 3}$. The object pose $\{\boldsymbol{R}, \boldsymbol{t}, \boldsymbol{s}\}$ is then solved from the correspondences between the keypoint coordinates $\boldsymbol{P}^{\text{kpt}}$ and their NOCS counterparts $\boldsymbol{O}^{\text{kpt}}$ with a pose fitting algorithm such as the Umeyama algorithm [32] or a deep estimator [13].

Point-to-Point Correspondence Supervision. To supervise the learning of NOCS coordinates, previous works [4, 15, 29] commonly employ a straightforward point-wise coordinate regression loss, which directly penalizes the deviation between the predicted and ground-truth coordinates in a point-to-point manner. Specifically, the ground-truth NOCS coordinates $\boldsymbol{O}^{\text{gt}}$ are derived by applying the ground-truth pose parameters $\{\boldsymbol{R}^{\text{gt}}, \boldsymbol{t}^{\text{gt}}, \boldsymbol{s}^{\text{gt}}\}$ to the keypoint coordinates $\boldsymbol{P}^{\text{kpt}}$, which is formulated as below:

$$\boldsymbol{O}^{\text{gt}} = \frac{1}{\|\boldsymbol{s}^{\text{gt}}\|_2} (\boldsymbol{R}^{\text{gt}})^\top (\boldsymbol{P}^{\text{kpt}} - \boldsymbol{t}^{\text{gt}}). \tag{15}$$

The correspondence loss is then defined as the point-wise L^2 or Smooth-L^1 [13] distance between the predicted NOCS coordinates $\boldsymbol{O}^{\text{kpt}}$ and the ground-truth $\boldsymbol{O}^{\text{gt}}$. For instance, the L^2-based loss is defined as:

$$\mathcal{L}^{\text{corr}} = \frac{1}{N^{\text{kpt}}} \sum\nolimits_n \|\boldsymbol{O}_n^{\text{kpt}} - \boldsymbol{O}_n^{\text{gt}}\|_2. \tag{16}$$

Geometric Relation Consistency Constraint. Nevertheless, the above point-to-point constraint fails to capture the holistic geometric structure of object. For example, two sets of NOCS coordinates may exhibit similar mean point-wise errors, yet represent substantially different overall shapes. This ambiguity can lead to a structural mismatch between the predicted NOCS coordinates and the underlying object geometry observed in the input space. As a result, it becomes challenging to reliably estimate a globally coherent rigid transformation from these correspondences, which may compromise the accuracy of the final pose estimation. To address this limitation, we propose a novel geometric relation consistency loss that explicitly enforces alignment in the overall geometric relations among keypoints to capture higher-order structural cues. Specifically, we compute the pairwise L^2 distances among the scaled keypoint coordinates $\boldsymbol{P}^{\text{kpt}} / \|\boldsymbol{s}^{\text{gt}}\|_2$ to construct the reference geometric

relation matrix $\boldsymbol{G}^{\text{kpt}} \in \mathbb{R}^{N^{\text{kpt}} \times N^{\text{kpt}}}$, and similarly derive the predicted counterpart $\boldsymbol{G}^{\text{nocs}}$ from the predicted NOCS coordinates $\boldsymbol{O}^{\text{kpt}}$. The geometric relation consistency loss is then computed between these two matrices, defined as:

$$\mathcal{L}^{\text{geo}} = \frac{1}{N^{\text{kpt}} \times N^{\text{kpt}}} \sum\nolimits_{n,m} (\boldsymbol{G}^{\text{kpt}}_{n,m} - \boldsymbol{G}^{\text{nocs}}_{n,m})^2. \quad (17)$$

In summary, the overall loss function is as follows:

$$\mathcal{L}^{\text{all}} = \lambda^{\text{com}}\mathcal{L}^{\text{com}} + \lambda^{\text{score}}\mathcal{L}^{\text{score}} + \lambda^{\text{corr}}\mathcal{L}^{\text{corr}} + \lambda^{\text{geo}}\mathcal{L}^{\text{geo}}, \quad (18)$$

with $\lambda^{\text{com}}, \lambda^{\text{score}}, \lambda^{\text{corr}}, \lambda^{\text{geo}}$ as balance hyper-parameters.

4. Experiments

4.1. Experimental Setup

Datasets. We conduct experiments on three benchmarks including CAMERA25, REAL275 [36] and HouseCat6D [8]. CAMERA25 is a synthetic dataset comprising 275K training and 25K testing images across 6 object categories. These images are generated via mixed-reality techniques by rendering foreground objects onto real backgrounds. REAL275 is a real-world dataset with 4.3K training images from 7 scenes and 2.75K testing images from 6 scenes, covering the same 6 object categories as CAMERA25. HouseCat6D is an emerging real-world benchmark containing 20K training images from 34 scenes and 3K testing images from 5 scenes, spanning 10 household object categories. This dataset includes photometrically challenging objects with diverse viewpoints and *occlusions*, posing significant challenges for accurate object pose estimation.

Evaluation Metrics. Following previous works [15, 29], we report the mean Average Precision (mAP) of $n^\circ m$ cm for 6D pose evaluation, which measures the percentage of predictions with rotation error below n° and translation error below m cm. We also report the mAP of 3D Intersection over Union (IoU_x) at a threshold of $x\%$ for joint 6D pose and 3D size evaluation.

Implementation Details. For a fair comparison, we use the same instance segmentation masks as AG-Pose [15], obtained from Mask R-CNN [7]. We sample $N^{\text{part}} = 1024$ partial points and extract $N^{\text{kpt}} = 64$ keypoints. The dense complete shape is reconstructed with $N^{\text{com}} = 1024$ points, where each keypoint is expanded into $N^{\text{fold}} = 16$ fine-grained local geometry points. The numbers of missing and visible keypoints are set to $N^{\text{miss}} = 64$ and $N^{\text{vis}} = 32$, respectively. The network architecture includes 2 attention layers in each of the following modules: partial feature extraction, progressive shape completion, and geometric relation encoding. In the geometric relation encoding module, the number of nearest neighbors N^{knn} is set to 16 and 32 across the two layers, respectively. The feature dimension D is set to 256 for RGB-D input and 128 for the depth-only setting. For loss balancing, we set the weights $\lambda^{\text{com}} = 15$, $\lambda^{\text{score}} = 1$, $\lambda^{\text{corr}} = 2$, and $\lambda^{\text{geo}} = 1$. The network is trained using the Adam [9] optimizer with an initial learning rate of 0.001 and a cosine annealing schedule. All experiments are conducted on a single RTX3090Ti GPU with a batch size of 24 over 200K iterations.

Table 1. Performance comparison with state-of-the-art methods on the REAL275 dataset. The method marked with '*' is reproduced by us. "Prior" refers to shape priors [31]. For each data setting, the best results are in **bold**, and the second best results are underlined.

Method	Prior	IoU_{50}	IoU_{75}	5°2cm	5°5cm	10°2cm	10°5cm
RGB-D Setting							
SPD [31]	✓	77.3	53.2	19.3	21.4	43.2	54.1
SGPA [1]	✓	80.1	61.9	35.9	39.6	61.3	70.7
DPDN [13]	✓	83.4	76.0	46.0	50.7	70.4	78.4
GCE-Pose [10]	✓	**84.1**	79.8	57.0	65.1	75.6	86.3
VI-Net [14]	×	-	-	50.0	57.6	70.8	82.1
SecondPose [3]	×	-	-	56.2	63.6	74.7	86.0
AG-Pose [15]	×	**84.1**	80.1	57.0	64.6	75.1	84.7
SpherePose [28]	×	84.0	79.0	58.2	67.5	76.2	88.2
SpotPose [29]	×	**84.1**	81.2	59.7	64.8	81.5	88.2
CleanPose [16]	×	-	-	61.5	67.4	78.3	86.2
ComPose	×	84.0	**81.4**	**62.1**	**68.0**	**81.8**	**89.2**
Depth-only Setting							
SAR-Net [11]	✓	79.3	62.4	31.6	42.3	50.3	68.3
RBP-Pose [43]	✓	-	67.8	38.2	48.1	63.1	79.2
DR-Pose [45]	✓	78.9	68.2	41.7	46.0	67.7	76.3
GPV-Pose [5]	×	-	64.4	32.0	42.9	-	73.3
HS-Pose [44]	×	82.1	74.7	46.5	55.2	68.6	82.7
Query6DoF [37]	×	82.5	76.1	49.0	58.9	68.7	83.0
AG-Pose* [15]	×	**83.2**	75.6	48.8	58.8	68.5	80.8
ComPose	×	82.1	**77.0**	**55.6**	**61.3**	**77.8**	**85.0**

4.2. Comparison with State-of-the-art Methods

Results on the REAL275 dataset. Table 1 presents a comprehensive comparison between our method and existing RGB-D and depth-only approaches on the REAL275 dataset. Under the depth-only setting, our ComPose outperforms previous methods by a large margin across all 6D pose evaluation metrics. Notably, when compared to the keypoint-based AG-Pose [15], our ComPose achieves significant improvements of 6.8% on 5°2cm and 9.3% on 10°2cm, demonstrating the importance of shape completion for precise pose estimation. For 3D IoU metrics, ComPose achieves comparable performance on IoU_{50} and sets a new state-of-the-art on the stricter IoU_{75} criterion. When incorporating semantic information from RGB images, ComPose exhibits further performance gains under the RGB-D setting, achieving the best results across all 6D pose metrics without relying on shape priors. These results validate the effectiveness of our unified completion-pose framework.

Results on the HouseCat6D dataset. Table 2 reports the performance comparison on the more challenging HouseCat6D dataset. Our ComPose consistently achieves state-of-the-art results across all evaluation metrics under both

Table 2. Performance comparison with state-of-the-art methods on the HouseCat6D dataset. The method marked with '*' is reproduced by us. For each data setting, the best results are shown in **bold**, and the second best results are underlined.

Method	IoU_{25}	IoU_{50}	5°2cm	5°5cm	10°2cm	10°5cm
RGB-D Setting						
VI-Net [14]	80.7	56.4	8.4	10.3	20.5	29.1
SecondPose [3]	83.7	66.1	11.0	13.4	25.3	35.7
AG-Pose [15]	88.1	76.9	21.3	22.1	51.3	54.3
SpherePose [28]	88.8	72.2	19.3	25.9	40.9	55.3
SpotPose [29]	89.1	77.0	23.8	24.5	52.3	54.8
GCE-Pose [10]	-	79.2	24.8	25.7	55.4	58.4
CleanPose [16]	89.2	79.8	22.4	24.1	51.6	56.5
ComPose	**90.3**	**80.6**	**25.8**	**27.6**	**57.8**	**61.5**
Depth-only Setting						
FS-Net [2]	74.9	48.0	3.3	4.2	17.1	21.6
GPV-Pose [5]	74.9	50.7	3.5	4.6	17.8	22.7
AG-Pose* [15]	81.4	59.9	9.7	10.6	25.9	29.7
ComPose	**81.6**	**65.1**	**11.8**	**12.7**	**34.8**	**38.9**

depth-only and RGB-D settings. Specifically, under the depth-only setting, ComPose outperforms AG-Pose [15] by 5.2% on the IoU_{50} metric and 2.1% on the 5°2cm metric. Under the RGB-D setting, ComPose surpasses GCE-Pose [10] by 1.4% on IoU_{50} and 1.0% on 5°2cm. The superior performance on this more comprehensive and challenging real-world dataset with occlusions further validate the effectiveness and robustness of our ComPose framework.

Results of Completion. Previous methods typically rely on category-level shape priors [31] to reconstruct object CAD models at unit scale *in the canonical space*. In contrast, our approach is the first to perform shape completion directly *in the observation space* under arbitrary poses, which is much more challenging. Table 3 compares the reconstruction performance of different methods on the hard camera category of the REAL275 dataset. To ensure a fair comparison, in addition to the metric-scale CD as defined in Equation (8), we also compute the unit-scale CD^{unit} metric. This involves normalizing the completed shape $\boldsymbol{P}^{com}$ and the ground-truth shape $\boldsymbol{M}^{obs}$ to unit scale using the ground-truth scale $\|\boldsymbol{s}^{gt}\|_2$ before calculating the Chamfer Distance. As shown in the table, without relying on shape priors, our RGB-D model performs best in reconstructing complete 3D shapes.

Robustness to Occlusion. Although the integrated shape completion primarily focuses on addressing point cloud incompleteness caused by *self-occlusion*, our method also performs effectively under *external occlusion*. In Table 4, we simulate severe occlusion by applying a 25% occlusion mask to the object segmentation masks on the REAL275 dataset, with the masked region randomly selected from the top, bottom, left, and right edges of the object. As shown, AG-Pose suffers a performance drop of 16.5% on 5°5cm, while our ComPose exhibits a smaller decrease by 12.6%, demonstrating its superior robustness against occlusion.

Table 3. Reconstruction performance comparisons for the camera category on the REAL275 dataset, measured using Chamfer Distance ($\times 10^{-3}$). CD^{unit} is computed after *unit-scale* normalization of the shape, while CD is computed directly in the observation space under the real-world *metric scale*. A lower value (↓) indicates better performance, with the best results highlighted in **bold**.

Method	Data Setting	Shape Prior	CD^{unit} ↓	CD ↓
Reconstruct 3D Object Models in the Canonical Space				
SPD [31]	RGB-D	✓	8.89	-
SGPA [1]	RGB-D	✓	5.51	-
DR-Pose [45]	D	✓	5.26	-
Reconstruct Complete 3D Shapes in the Observation Space				
ComPose	RGB-D	×	**4.20**	**0.17**
ComPose	D	×	6.09	0.23

Table 4. Performance comparison of different depth-only methods under occlusion-augmented testing on the REAL275 dataset.

Method (D)	Test with OccAug	5°2cm	5°5cm	10°2cm	10°5cm
AG-Pose* [15]	×	48.8	58.8	68.5	80.8
	✓	37.1	49.1	54.3	72.6
	Drop ↓	24.0%	16.5%	20.7%	10.1%
ComPose	×	55.6	61.3	77.8	85.0
	✓	42.7	53.6	62.9	77.7
	Drop ↓	**23.2%**	**12.6%**	**19.2%**	**8.6%**

Table 5. Ablation studies on the shape completion strategy. "Partial Instance" indicates reconstructing only visible object regions.

Reconstruction	$\boldsymbol{P}^{kpt}$	$\boldsymbol{P}^{com}$	5°2cm	5°5cm	10°2cm	10°5cm
Partial Instance [15]	✓	✓	49.6	56.1	72.4	82.0
Complete $\boldsymbol{M}^{obs}$	✓	✓	**55.6**	**61.3**	**77.8**	**85.0**
Complete $\boldsymbol{M}^{obs}$	✓	×	54.9	60.5	76.1	83.3

4.3. Ablation Studies

In this section, we conduct comprehensive ablation studies on the REAL275 dataset under the depth-only setting to shed more light on the superiority of our method.

Efficacy of Shape Completion. Table 5 presents ablation studies on different shape completion strategies. Replacing the *complete shape recovery* with the *partial instance reconstruction* as in AG-Pose [15] results in a 6% decrease on 5°2cm, demonstrating the importance of complete geometric cues provided by shape completion for precise pose estimation. Moreover, the completion of dense point clouds $\boldsymbol{P}^{com}$ enhances the fine-grained geometric awareness of keypoints, leading to a 1.7% performance gain on 10°5cm.

Efficacy of the Unified Framework. Figure 2 illustrates the superiority of the proposed unified Completion-Pose framework over two-stage Completion-then-Pose scheme in terms of both accuracy and efficiency. In the two-stage pipeline, the completed dense shape $\boldsymbol{P}^{com}$ replaces the original partial input $\boldsymbol{P}^{part}$ of AG-Pose [15], introducing additional inference time during completion. In contrast, ComPose directly replaces AG-Pose's *keypoint detection module* with the *keypoint-based completion module*, elimi-

Table 6. Ablation studies on the progressive completion process, where N^{kpt} is kept constant at 64 across all experiments.

Completion	Selection	N^{miss}	N^{vis}	5°2cm	5°5cm	10°2cm	10°5cm
Static Query	-	-	-	51.6	58.6	72.9	82.0
PoinTr [40]	×	64	0	52.7	59.6	74.3	82.7
AdaPoinTr [41]	✓	64	32	53.4	59.8	75.6	83.8
Progressive	✓	64	32	**55.6**	**61.3**	**77.8**	**85.0**
Progressive	×	64	0	53.8	61.0	75.4	84.0
Progressive	×	32	32	54.6	60.0	76.6	83.7

Table 7. Ablation studies on the geometric relation modeling.

Encoding	Consistency	5°2cm	5°5cm	10°2cm	10°5cm
×	×	49.5	55.6	71.9	79.7
✓	×	53.8	60.5	74.8	83.5
✓	✓	**55.6**	**61.3**	**77.8**	**85.0**

nating extra latency. Both frameworks utilize the same completion module, with the depth-only version operating at a feature dimension of 128. The efficiency gains of ComPose over the original AG-Pose mainly stem from reducing the keypoint count from 96 to 64, removing the Self-Attention operation in NOCS prediction, and using the Umeyama algorithm rather than a deep estimator for pose fitting.

Efficacy of Keypoint-based Progressive Completion. In Table 6, we conduct ablation studies on the progressive completion process. "Static Query" indicates that keypoint queries $\boldsymbol{Q}^{\text{kpt}}$ are initialized as learnable embeddings [19], without the construction of coarse keypoints $\boldsymbol{C}^{\text{kpt}}$. Unlike our approach, classic methods such as PoinTr [40] and AdaPoinTr [41] do not enforce constraints on the reconstruction of $\boldsymbol{C}^{\text{miss}}$, resulting in a significant deviation of the predicted coarse coordinates $\boldsymbol{C}^{\text{kpt}}$ from the *actual object shapes*. In addition, AdaPoinTr lacks explicit supervision during the keypoint selection process. As shown in the table, our progressive completion module outperforms all three alternative strategies, with a 2.2% improvement over AdaPoinTr on the 5°2cm metric. The last two rows further demonstrate the necessity of adaptive keypoint selection, which effectively harnesses reliable geometric cues from visible keypoints and flexibly balances missing and visible regions.

Efficacy of Geometric Relation Modeling. We conduct ablation studies on the components of geometric relation modeling in Table 7. The geometric relation encoding module explicitly enhances the keypoint features by capturing geometric context, leading to a 4.3% improvement on the 5°2cm metric. Additionally, the geometric relation consistency constraint enforces structural alignment between the predicted NOCS coordinates and the observed object geometry, further improving performance by 1.8% on 5°2cm.

4.4. Visualization

Figure 4 visualizes the keypoint-based progressive completion process. The results demonstrate that our completion module not only progressively recovers the complete object shape from partial observations, but also effectively filtering

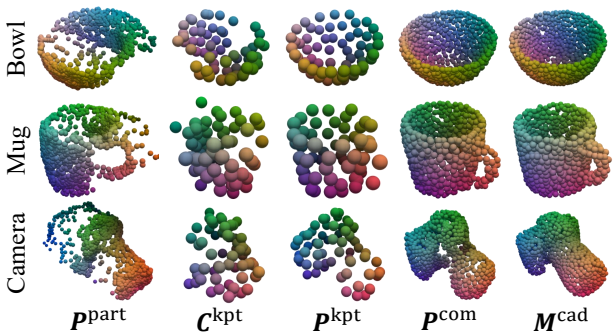

Figure 4. Visualization of the keypoint-based progressive completion. Complete object geometries are progressively recovered.

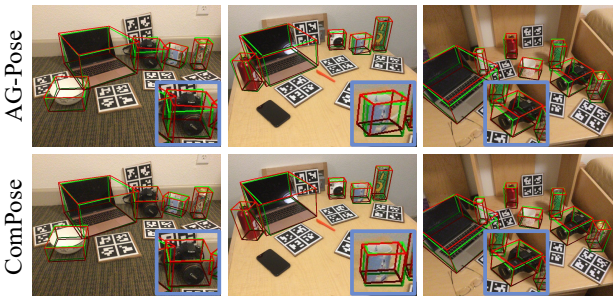

Figure 5. Qualitative comparison between our ComPose and AG-Pose [15]. Red/Green indicates the predicted/GT results.

out outlier points caused by inaccurate segmentation [22–24, 29], yielding a cleaner and more comprehensive object representation. This refined geometric representation enhances the robustness of object pose estimation. As presented in Figure 5, our ComPose achieves more accurate and reliable pose predictions than AG-Pose [15], benefiting from a better comprehension of holistic object geometry.

5. Conclusion

In this work, we propose ComPose, a novel framework that integrates keypoint-based progressive completion to enhance the understanding of complete object shapes, which exhibits significant potential for robust category-level object pose estimation. To further improve structural awareness, we introduce geometric relation encoding and consistency constraints, which explicitly capture object geometric structure and facilitate globally coherent coordinate transformations for robust pose fitting. Extensive experiments on existing benchmarks consistently demonstrate the superior performance of our method under both RGB-D and depth-only settings without relying on category-level shape priors.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. 62306294) and the Open Fund of National Key Laboratory of Deep Space Exploration (Grant NKDSEL2025008).

References

[1] 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. 3, 6, 7

[2] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, 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*, pages 1581–1590, 2021. 3, 7

[3] 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. 3, 6, 7

[4] Yihan Chen, Wenfei Yang, Huan Ren, Shifeng Zhang, Tianzhu Zhang, and Feng Wu. Structure-aware correspondence learning for relative pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11611–11621, 2025. 5

[5] 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, 3, 6, 7

[6] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 5

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3, 6

[8] 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. 6

[9] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015. 6

[10] 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 Computer Vision and Pattern Recognition Conference*, pages 27154–27165, 2025. 2, 3, 6, 7

[11] 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*, pages 6707–6717, 2022. 6

[12] 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*, pages 3560–3569, 2021. 3

[13] 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, 3, 5, 6

[14] 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. 6, 7

[15] 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, 5, 6, 7, 8

[16] Xiao Lin, Yun Peng, Liuyi Wang, Xianyou Zhong, Minghao Zhu, Yi Feng, Jingwei Yang, Chengju Liu, and Qijun Chen. Cleanpose: Category-level object pose estimation via causal learning and knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5990–6000, 2025. 2, 6, 7

[17] Jianhui Liu, Yukang Chen, Xiaoqing Ye, and Xiaojuan Qi. Ist-net: Prior-free category-level pose estimation with implicit space transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13978–13988, 2023. 3

[18] Jian Liu, Wei Sun, Chongpei Liu, Xing Zhang, and Qiang Fu. Robotic continuous grasping system by shape transformer-guided multiobject category-level 6-d pose estimation. *IEEE Transactions on Industrial Informatics*, 19 (11):11171–11181, 2023. 1

[19] Wang Luo, Huan Ren, Tianzhu Zhang, Wenfei Yang, and Yongdong Zhang. Adaptive prototype learning for weakly-supervised temporal action localization. *IEEE Transactions on Image Processing*, 34:3154–3168, 2024. 8

[20] Fabian Manhardt, Wadim Kehl, and Adrien Gaidon. Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2069–2078, 2019. 1

[21] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 3

[22] Yuwen Pan, Rui Sun, Yuan Wang, Wenfei Yang, Tianzhu Zhang, and Yongdong Zhang. Purify then guide: A bi-directional bridge network for open-vocabulary semantic

segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 8

[23] Yuwen Pan, Rui Sun, Yuan Wang, Tianzhu Zhang, and Yongdong Zhang. Rethinking the implicit optimization paradigm with dual alignments for referring remote sensing image segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2031–2040, 2024.

[24] Yuwen Pan, Rui Sun, Wangkai Li, and Tianzhu Zhang. Exploring weather-aware aggregation and adaptation for semantic segmentation under adverse conditions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13952–13962, 2025. 8

[25] 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. 1

[26] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 3

[27] Huan Ren, Wenfei Yang, Tianzhu Zhang, and Yongdong Zhang. Proposal-based multiple instance learning for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2394–2404, 2023. 2

[28] 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. 6, 7

[29] Huan Ren, Wenfei Yang, Shifeng Zhang, and Tianzhu Zhang. Rethinking correspondence-based category-level object pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1170–1179, 2025. 2, 3, 5, 6, 7, 8

[30] Yongzhi Su, Jason Rambach, Nareg Minaskan, Paul Lesur, Alain Pagani, and Didier Stricker. Deep multi-state object pose estimation for augmented reality assembly. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 222–227. IEEE, 2019. 1

[31] 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, 2, 3, 6, 7

[32] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04): 376–380, 1991. 5

[33] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3, 4

[34] 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. 1

[35] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16611–16621, 2021. 1

[36] 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*, pages 2642–2651, 2019. 1, 3, 6

[37] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14055–14064, 2023. 6

[38] Wenfei Yang, Huan Ren, Tianzhu Zhang, Zhe Zhang, Yongdong Zhang, and Feng Wu. Cross-task relation-aware consistency for weakly supervised temporal action detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):10513–10527, 2025. 2

[39] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 206–215, 2018. 3, 4

[40] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12498–12507, 2021. 2, 3, 8

[41] Xumin Yu, Yongming Rao, Ziyi Wang, Jiwen Lu, and Jie Zhou. Adapointr: Diverse point cloud completion with adaptive geometry-aware transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):14114–14130, 2023. 2, 3, 4, 8

[42] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *2018 international conference on 3D vision (3DV)*, pages 728–737. IEEE, 2018. 3, 4

[43] 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. 6

[44] 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*, pages 17163–17173, 2023. 3, 6

[45] Lei Zhou, Zhiyang Liu, Runze Gan, Haozhe Wang, and Marcelo H Ang. Dr-pose: A two-stage deformation-and-registration pipeline for category-level 6d object pose estimation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1192–1199. IEEE, 2023. 3, 6, 7