TGVPOSE: LIGHTWEIGHT CATEGORY-LEVEL OBJECT POSE ESTIMATION VIA TOPOLOGY GRAPH-GUIDED ADAPTIVE GRAPH VECTOR NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Object pose estimation serves as a crucial approach to enabling reliable robotic grasping. To achieve high generalization in real-world scenarios, existing pose estimation methods commonly adopt category-level object pose estimation strategies. However, current high-precision methods often suffer from large parameter volumes, rendering them difficult to deploy on computation-constrained edge devices such as mobile robots and unmanned aerial vehicles (UAVs). To address this issue, this paper proposes TGVPose, a high-precision and lightweight category-level object pose estimation method via topology graph-guided adaptive graph vector networks. First, to overcome the fixed-graph limitation inherent in conventional graph vector networks, we introduce an Adaptive Multi-Relation Gated Vector Neuron Layer (AMG-VN), which enables the network to adaptively extract both local and global features. Second, to enhance the network's understanding of object shapes, we present a category-prototype persistent topology conditioning method, thereby achieving accurate object pose estimation. Finally, we evaluate TGVPose across multiple datasets and real-world object pose estimation tasks. The experimental results demonstrate that TGVPose effectively balances accuracy and lightweight efficiency, delivering superior pose estimation performance compared to existing lightweight methods.

1 INTRODUCTION

Robotic grasping is a fundamental capability for autonomous systems, and accurate 6D object pose estimation Tremblay et al. (2018); Wang et al. (2019a) is a prerequisite for reliable grasp execution. Although instance-level pose estimation methods Tremblay et al. (2018); Wang et al. (2019a); He et al. (2020); Lin et al. (2022c) have achieved promising results, their reliance on object-specific CAD models limits generalization to unseen objects and restricts their applicability in open-world environments. Category-level pose estimation Tian et al. (2020); Chen & Dou (2021); Wang et al. (2021); Lin et al. (2022b) addresses this issue by predicting the poses of unseen instances from known categories without requiring their CAD models.

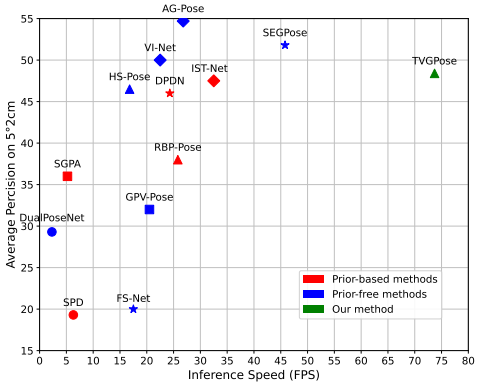


Figure 1: Comparison between previous methods and our method.

Recent category-level methods Lin et al. (2021; 2023; 2024; 2025) can produce accurate pose estimates, but their large model sizes and computational costs hinder deployment on resource-constrained edge devices, such as mobile robots and unmanned aerial vehicles. Lightweight methods Lin et al. (2022a); Du et al. (2025) reduce the reliance on RGB feature extraction and directly estimate poses from point clouds. However, their accuracy is often insufficient for practical grasping tasks because conventional point-cloud convolutions cannot effectively capture the intrinsic geometric structure of objects. Vector Neuron Networks (VNNs) Deng et al. (2021) represent point features as vectors and provide rotation-equivariant mappings from input point clouds to learned features.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

When combined with SE(3)-equivariant learning, VNN-based methods Du et al. (2025); Yu et al. (2026) improve pose estimation performance; nevertheless, they still face several limitations.

First, existing SE(3)-equivariant pose estimation networks commonly contain multiple VNN layers, repeated KNN graph construction, high-dimensional vector transformations, and auxiliary boundary distance field reconstruction. These components introduce substantial computation and memory consumption, making low-latency and low-power deployment on mobile robots difficult. Second, existing methods usually aggregate KNN features through mean or max pooling, treating neighboring points as equally important. Such aggregation cannot effectively distinguish structurally informative points from redundant observations, nor can it adaptively coordinate local geometry, feature similarity, and global context. Third, existing methods lack explicit category-level topological priors. Under severe occlusion or partial observations, local geometric evidence and implicit reconstruction alone may not recover stable global structures, leading to inaccurate rotation, translation, and size estimation.

To address these issues, we propose TGVPose, a topology-guided adaptive graph vector network for category-level object pose estimation. TGVPose introduces an adaptive multi-relation graph vector-neuron layer that jointly models local geometric neighborhoods, feature-space neighborhoods, and global context while preserving the rotation equivariance of vector features. Rotation-invariant gating weights adaptively determine the contributions of different relation branches, enabling the network to aggregate informative local structures and global semantic cues more effectively.

We further propose a category-prototype persistent topology conditioning module. Persistent-homology images are computed offline from canonical point clouds for each category, providing topological priors that characterize connectivity, loop structures, and cavities. During training, the predicted topology descriptors are supervised by category prototypes and residually injected into invariant point features, thereby constraining global shape representations under occlusion and missing observations. Since the topology priors are precomputed offline, the inference stage requires no persistent-homology computation and introduces only lightweight topology prediction and feature mapping operations.

In summary, the main contributions of this work are as follows:

1) We design an adaptive multi-relation graph vector-neuron layer that jointly models local geometric neighborhoods, feature-space neighborhoods, and global context. A rotation-invariant gating mechanism adaptively fuses these relations while preserving SE(3) equivariance.

2) We propose a category-prototype persistent topology conditioning module that injects category-level global shape priors into invariant features through topology prediction and prototype supervision, improving robustness to occlusion and partial point clouds without online persistent-homology computation.

3) We evaluate TGVPose on the REAL275 dataset and conduct real-world grasping experiments on a UR3 robot. The results demonstrate that TGVPose achieves competitive pose estimation performance with efficient inference.

2 Related Works

2.1 RGB-D-based Category-Level Object Pose Estimation

Recent RGB-D category-level pose estimation methods often utilize categorical shape priors to eliminate reliance on exact 3D CAD models Tian et al. (2020); Chen & Dou (2021); Wang et al. (2021); Lin et al. (2022b); Yu et al. (2024). While prior-based paradigms refine generic templates via iterative deformation Wang et al. (2021), dynamic cross-attention Chen & Dou (2021), or coarse-to-fine alignment Yu et al. (2024), they incur significant computational overhead.

To improve efficiency, prior-free frameworks have been explored Lin et al. (2021; 2023; 2024; 2025). Early approaches like DualPoseNet Lin et al. (2021) enforce consistency between explicit and implicit pose encodings, though lacking shape priors often compromises accuracy. To bridge this gap, AG-Pose Lin et al. (2024) introduced adaptive keypoint detection, and CleanPose Lin et al. (2025) leveraged knowledge distillation to enhance category-level guidance.

Nonetheless, these methods rely on heavily parameterized transformer backbones, creating substantial latency that hinders real-time deployment on edge devices.

2.2 Depth-based Category-Level Object Pose Estimation

To enhance prediction accuracy and improve real-time performance, researchers attempt to perform object pose estimation using exclusively depth point clouds, thereby reducing the processing time spent on RGB image feature extraction. FS-Net Chen et al. (2021) decodes orientation by predicting orthogonal basis vectors from fused multimodal features. Similarly, GPV-Pose Di et al. (2022) applies 3D Graph Convolutions to extract local geometric patterns for direct pose regression. HS-Pose Zheng et al. (2023) introduces an HS-layer to enhance point cloud features and encode rotation-translation information. SAR-Net Lin et al. (2022a) introduces a lightweight model with shape alignment for better object center and 3D size estimation. SEGPose Yu et al. (2026) propose a point cloud reconstruction method guided by geometric and topological information, achieving precise reconstruction of object clouds.

Although these methods effectively mitigate the parameter overhead associated with RGB-D approaches, they struggle to achieve an optimal trade-off on edge devices such as mobile robots and unmanned aerial vehicles. Specifically, while certain methods Zheng et al. (2023); Yu et al. (2026) deliver high precision, they suffer from poor real-time responsiveness; conversely, lightweight alternatives Lin et al. (2022a); Du et al. (2025) achieve faster inference at the expense of estimation accuracy. Consequently, achieving accurate pose estimation while maintaining real-time performance on computation-constrained devices remains a critical research challenge.

3 Method

Given the object point cloud $\mathcal{P}$ of a scene TGVPose aims to detect object instances and estimate their category-level 9DoF poses. For each object instance, the pose is represented by rotation $R \in SO(3)$, translation $t \in \mathbb{R}^3$, and object size $s \in \mathbb{R}^3$.

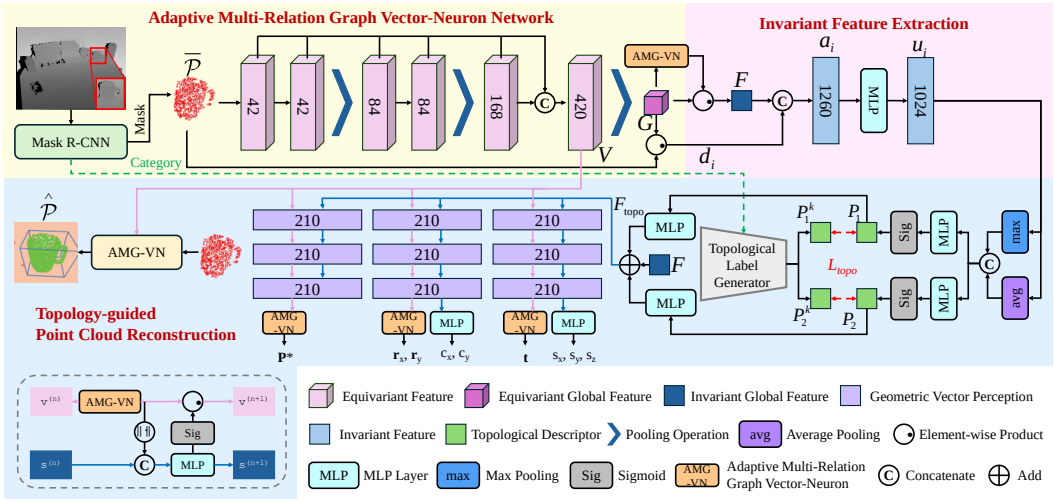


Figure 2: The pipeline of TGVPose.

3.1 Overview

The TGVPose training pipeline is shown in Figure 2. We first center the input point cloud $\bar{\mathcal{P}}$ where $\bar{p}_i \in \bar{\mathcal{P}}$, $\bar{p}_i = p_i - o$, $o = \frac{1}{N}\sum_i p_i$, and feed it into a hierarchical vector-neuron encoder to extract a 420-channel equivariant representation V. Next, V is mapped to a rotation-invariant feature F to predict category-level topological descriptors (persistent images). These descriptors are residually fused with F to generate a topology-conditioned invariant feature F_{topo}. By constructing scalar features exclusively from vector norms and inner products, the network strictly maintains SE(3)-equivariance, ensuring $V^l(Q\bar{\mathcal{P}}) = QV^l(\bar{\mathcal{P}})$ for any rotation $Q \in \text{SO}(3)$. Finally, the equivariant V and invariant F_{topo} jointly guide the 3D shape reconstruction and SE(3)-consistent pose estimation (predicting object axes, translation, and size), with a boundary distance field providing auxiliary supervision.

3.2 Adaptive Multi-Relation Graph Vector-Neuron Layer

A single-relation KNN aggregation describes only one notion of neighborhood similarity and may mix structurally important points with redundant ones. We therefore design an adaptive multi-relation graph vector-neuron layer to jointly exploit complementary local relations and object-level context. Let $X = \{x_i\}_{i=1}^{N}$ denote the input vector feature, where N is the number of points, C is the number of vector channels, and $x_i \in \mathbb{R}^{C\times 3}$ contains the C vectors associated with point i. We use $x_{ci} \in \mathbb{R}^3$ to denote the c-th vector channel of point i, and $[\cdot,\cdot]$ to denote concatenation. We first flatten each x_i into a vector descriptor and perform Euclidean KNN search to obtain the neighborhood $\mathcal{N}_i^{\mathrm{vec}}$. The superscript vec indicates that this graph is constructed directly from the vector-neuron representation, thereby preserving the structure of the equivariant feature space. To capture a complementary similarity, we form a rotation-invariant descriptor $s_i = [\|x_{1i}\|_2, \ldots, \|x_{Ci}\|_2] \in \mathbb{R}^C$, where $\|\cdot\|_2$ is the Euclidean norm of a vector channel.

Euclidean KNN search in this descriptor space gives $\mathcal{N}_i^{\mathrm{inv}}$, where inv indicates that the neighborhood is defined by rotation-invariant channel responses. Thus, the two local graphs describe different but complementary relations. In the implementation, $K = 10$ denotes the number of neighbors in each graph.

For either local relation $\rho \in \{\mathrm{vec}, \mathrm{inv}\}$ and a neighbor $j \in \mathcal{N}_i^{\rho}$, we encode the relative variation and center-point context by $e_{ij} = [x_j - x_i, x_i]$.

Let $\psi_\rho(\cdot)$ denote the relation-specific equivariant vector-neuron transformation. The message associated with relation ρ is obtained by mean aggregation: $m_i^{\rho} = \frac{1}{|\mathcal{N}_i^{\rho}|}\sum_{j\in\mathcal{N}_i^{\rho}} \psi_\rho(e_{ij}), \rho \in \{\mathrm{vec}, \mathrm{inv}\}$. Here, $|\mathcal{N}_i^{\rho}|$ is the number of neighbors in the corresponding graph. This relation-specific transformation allows the two graphs to encode different local cues while retaining a compact mean-aggregation operation rather than introducing per-neighbor attention. To provide complementary object-level context, we introduce a global relation, denoted by the superscript g, and compute $m_i^{g} = \psi_g\left(\frac{1}{N}\sum_{j=1}^{N} x_j\right)$, where $\psi_g(\cdot)$ is the equivariant vector-neuron transformation for the global branch. The same message is broadcast to all points, providing global shape context without constructing another point-to-point graph.

The three messages are fused using a point-wise invariant gate. We concatenate their channel-wise norms with the invariant descriptor of the center point: $h_i = [s_i, \|m_i^{\mathrm{vec}}\|_2, \|m_i^{\mathrm{inv}}\|_2, \|m_i^{g}\|_2]$. Each norm is evaluated channel-wise, so h_i is invariant to the input orientation. A lightweight point-wise multilayer perceptron $\gamma(\cdot)$ maps h_i to relation logits.

The softmax operator converts them into scalar weights shared across the three vector coordinates: $(\alpha_i^{\mathrm{vec}}, \alpha_i^{\mathrm{inv}}, \alpha_i^{g}) = \mathrm{softmax}(\gamma(h_i))$, where $\sum_{\rho\in\{\mathrm{vec},\mathrm{inv},g\}} \alpha_i^{\rho} = 1$, α_i^{ρ} represents the contribution of relation ρ at point i. The fused output is then computed as $y_i = \psi_o\left(\sum_{\rho\in\{\mathrm{vec},\mathrm{inv},g\}} \alpha_i^{\rho} m_i^{\rho}\right) + S(x_i)$, where $\psi_o(\cdot)$ is the output equivariant vector-neuron transformation. The residual mapping $S(\cdot)$ is an identity when the input and output channel dimensions are equal and is otherwise implemented as an equivariant linear projection.

The layer preserves rotation equivariance. Let $Q \in \mathrm{SO}(3)$ denote an arbitrary 3D rotation. Under $x_{ci} \mapsto Qx_{ci}$ for every channel c, the distances used to construct $\mathcal{N}_i^{\mathrm{vec}}$ remain unchanged because the rotation is orthogonal, while s_i and $\mathcal{N}_i^{\mathrm{inv}}$ remain unchanged because they are constructed from vector norms. Moreover, $e_{ij} \mapsto Qe_{ij}$ and $m_i^{\rho} \mapsto Qm_i^{\rho}$ for every branch. Consequently, h_i and the scalar weights α_i^{ρ} are invariant, whereas their weighted message sum and the residual projection transform equivariantly, yielding $y_i \mapsto Qy_i$. Since translation is removed by centering the input point cloud, the proposed layer is consistent with the SE(3)-equivariant backbone. The scalar branch gate also avoids the additional parameter and memory cost of vector-valued or per-neighbor attention coefficients.

3.3 Invariant Feature Extraction and Topology Conditioning

After hierarchical encoding, the multi-scale vector features are concatenated into $V \in \mathbb{R}^{420\times 3\times N}$, where N is the number of input points and each of the 420 channels is a three-dimensional vector.

We first apply point-wise average pooling to obtain the global equivariant feature $G \in \mathbb{R}^{420\times 3}$. A vector-neuron projection produces two learned reference vectors $Z = \{z_l\}_{l=1}^{2}$, with $z_l \in \mathbb{R}^3$. We construct an invariant global descriptor by taking channel-wise inner products, $b_{cl} = \langle G_c, z_l\rangle, c = 1, \ldots, 420, l = 1, 2$, where G_c denotes the c-th vector channel of G and $\langle\cdot,\cdot\rangle$ is the Euclidean inner product. The 840 responses are flattened into a global descriptor $b \in \mathbb{R}^{840}$ and broadcast to all points.

For point i, we additionally define $d_i = [d_{1i}, \ldots, d_{420i}] \in \mathbb{R}^{420}$ with $d_{ci} = \langle G_c, \bar{p}_i\rangle$, and compute the radial distance $r_i = \|\bar{p}_i\|_2$, where $\bar{p}_i = p_i - o$ is the centered point. The descriptor $[b, d_i, r_i, e_k]$ combines global shape responses, point-specific geometry, radial information, and the category one-hot code $e_k \in \mathbb{R}^{N_{\text{cat}}}$. A point-wise projection maps this descriptor to the invariant feature $F \in \mathbb{R}^{420\times N}$. All its components are norms, inner products, or category-fixed scalars; therefore, F is invariant to a global rotation of the input point cloud.

Partial observations can preserve similar local geometry while differing substantially in their complete category-level shape. To provide a compact global prior, we introduce category-prototype persistent topology conditioning. For each category k, we construct an Alpha complex offline from its canonical point cloud and compute persistent diagrams in homology dimensions one and two. The dimension-one diagram describes persistent one-dimensional cycles, while the dimension-two diagram captures enclosed voids. We convert them into persistent images $P_1^k, P_2^k \in \mathbb{R}^{50\times 50}$ and flatten each image into a 2500-dimensional target vector. These category-specific prototypes are precomputed once and loaded during training, so persistent-homology computation is not required by the online network.

Given the invariant point feature F, a shared point-wise transformation produces $U = [u_1, \ldots, u_N] \in \mathbb{R}^{1024\times N}$. We summarize the point set using element-wise maximum and average pooling, $q = \left[\text{MAX}_{i=1}^{N}(u_i), \frac{1}{N}\sum_{i=1}^{N} u_i\right] \in \mathbb{R}^{2048}$, where u_i is the feature of point i and $[\cdot,\cdot]$ denotes concatenation. The two pooled statistics provide complementary summaries of salient and global responses. After a shared fully connected projection, two sigmoid heads predict the flattened persistent images, $\widehat{P}_1 = \text{sigmoid}(h_1(q)), \widehat{P}_2 = \text{sigmoid}(h_2(q))$, where h_1 and h_2 are the prediction heads for H_1 and H_2, respectively, and each prediction has 2500 elements.

The predicted topology descriptors are mapped back to the invariant feature space by learned projections $\phi_1, \phi_2 : \mathbb{R}^{2500} \to \mathbb{R}^{420}$ and broadcast to all points: $F_{\text{topo}} = F + \phi_1(\widehat{P}_1)1_N^\top + \phi_2(\widehat{P}_2)1_N^\top$, where $1_N \in \mathbb{R}^N$ is an all-one vector and F_{topo} denotes the topology-conditioned invariant feature. The residual formulation preserves instance-specific observations while injecting category-level cycle and void information into every point. Since this module operates exclusively on invariant scalar features, it does not change the transformation law of the equivariant vector stream. At inference time, the topology branch only performs point-wise projections, pooling, sigmoid prediction, and feature addition; all Alpha-complex and persistent-image computations are completed offline.

3.4 Pose and Shape Prediction

The prediction heads jointly consume V and F_{topo}. The reconstruction head predicts a point set $\widehat{\mathcal{P}}$ to regularize the learned representation with observable geometry. Two rotation heads predict normalized object-axis directions $\widehat{v}_y$ and $\widehat{v}_x$, together with confidence scores $\widehat{c}_y$ and $\widehat{c}_x$; these directions are assembled into the rotation $\widehat{R}$ while respecting category symmetry. The translation head predicts a residual relative to the input centroid, yielding $\widehat{t} = \Delta t + o$, and the size head predicts a residual relative to the category mean size. During training, an auxiliary decoder also estimates the boundary distance field at sampled query points. This auxiliary decoder and the explicit point reconstruction are not required for pose inference.

3.5 Loss Functions

TGVPose is trained with a multi-task objective that combines direct pose supervision, symmetry-aware point matching, geometric consistency, boundary distance prediction, and category-level topology regularization: $\mathcal{L} = \mathcal{L}_{\text{pose}} + \mathcal{L}_{\text{prop}} + \lambda_{\text{bdf}}\mathcal{L}_{\text{bdf}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} + \lambda_{\text{topo}}\mathcal{L}_{\text{topo}}$.

The pose term directly supervises the two object-axis directions, translation, size, axis orthogonality, and rotation confidence. Let v_y and v_x denote the ground-truth object axes and let M_{ns} indicate objects without continuous rotational symmetry. Since the secondary axis is ambiguous for symmetric instances, the corresponding supervision is masked. The pose loss is formulated as

$$\begin{aligned}\mathcal{L}_{\text{pose}} = \lambda_y \ell_{\text{dir}}(\widehat{v}_y, v_y) + \lambda_x M_{\text{ns}} \ell_{\text{dir}}(\widehat{v}_x, v_x) + \lambda_o M_{\text{ns}} |\widehat{v}_y^\top \widehat{v}_x| \\ + \lambda_t \|\widehat{t} - t\|_1 + \lambda_s \|\widehat{s} - s\|_1 + \lambda_c \mathcal{L}_{\text{conf}},\end{aligned} \tag{1}$$

where $\ell_{\text{dir}}(\widehat{v}, v) = \|\widehat{v} - v\|_1 + 2(1 - \widehat{v}^\top v)$ combines component-wise regression and angular consistency. For each axis, the confidence target is generated from its direction error as $c^* = \exp(-13.7\|\widehat{v} - v\|_2^2)$, and $\mathcal{L}_{\text{conf}} = |\widehat{c}_y - c_y^*| + M_{\text{ns}}|\widehat{c}_x - c_x^*|$. This error-aware target encourages the rotation head to expose the reliability of each predicted direction.

The symmetry-aware term couples pose estimation with point reconstruction. Let $\mathcal{P}^* = R^\top(\mathcal{P} - t)$ and $\widehat{\mathcal{P}}^* = \widehat{R}^\top(\mathcal{P} - \widehat{t})$ denote the point clouds transformed to the ground-truth and predicted canonical frames. The point-matching loss penalizes $\|\widehat{\mathcal{P}}^* - \mathcal{P}^*\|_1$. For categories with reflectional symmetry, valid reflected counterparts are additionally constructed to constrain both the reconstructed points and their consistency with the predicted pose. Accordingly,

$$\mathcal{L}_{\text{prop}} = \lambda_{\text{pm}} \mathcal{L}_{\text{pm}} + \lambda_{\text{sym}}(\mathcal{L}_{\text{sym-rec}} + \mathcal{L}_{\text{sym-pose}}). \tag{2}$$

This formulation prevents symmetry-equivalent poses from being incorrectly penalized while maintaining point-level geometric supervision.

We further impose geometric and boundary constraints. For each point p_i, its canonical coordinate is $p_i^* = R^\top(p_i - t)$. The projections $(p_i - \widehat{t})^\top \widehat{v}_y$ and $(p_i - \widehat{t})^\top \widehat{v}_x$ are matched to the corresponding coordinates of p_i^*, with the secondary-axis term again masked for symmetric objects. The resulting loss is $\mathcal{L}_{\text{geo}} = \|\widehat{y} - y^*\|_1 + M_{\text{ns}}\|\widehat{x} - x^*\|_1$, $\mathcal{L}_{\text{bdf}} = \|\widehat{d} - d^*\|_1$, where $\widehat{d}$ and d^* are the predicted and analytically generated boundary distance values at the sampled queries. These objectives constrain pose predictions through the observed geometry and encourage the equivariant representation to preserve object-boundary information.

Finally, the topology predictor is supervised by the offline category prototypes. For category k, the topology loss is $\mathcal{L}_{\text{topo}} = \|\widehat{P}_1 - P_1^k\|_1 + \|\widehat{P}_2 - P_2^k\|_1$. Empty topology targets are excluded using a validity mask. The topology term encourages partial instances to retain stable category-level cycle and void structures, complementing the local geometric losses with a global structural constraint. In the implementation, $\lambda_{\text{bdf}} = 20$, $\lambda_{\text{geo}} = 1$, and $\lambda_{\text{topo}} = 1$; the point-matching and symmetry terms are weighted by 2 and 1, respectively.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate TGVPose on standard benchmark datasets for category-level 6D object pose estimation, specifically REAL275 Wang et al. (2019b) and Wild6D Fu & Wang (2022). The REAL275 dataset comprises real-world RGB-D sequences, including 4,300 training images and 2,750 testing images captured across 6 distinct evaluation scenes. To assess generalization across broader real-world environments, we also test on the Wild6D dataset, which encompasses 486 video sequences covering over 162 unique object instances.

Evaluation Metrics. We evaluate the performance of SEGPose using widely used evaluation metrics Wang et al. (2019b); Tian et al. (2020); Lin et al. (2021); Di et al. (2022); Lin et al. (2022a). For rotation and translation evaluation, we utilize 3D Intersection-Over-Union (IoU) with thresholds of 0.5 and 0.75. Additionally, we employ $5°2cm$, $5°5cm$, $10°2cm$, and $10°5cm$ to directly assess rotation and translation accuracy. If the errors fall within the thresholds, the predictions are deemed correct. Based on these evaluation metrics, we will use overall mAP to assess the performance of TGVPose compared to other SOTA methods.

Implementation details. The network is trained on the REAL275 Wang et al. (2019b) training split using instance-masked point clouds, each uniformly sampled to $N = 1,024$ points. The adaptive graph layer sets the k-nearest neighbors to $K = 10$, enabling vector-feature, invariant-feature, and

global relation branches along with relation gating and residual connections. Category topology prototypes are constructed offline via Alpha complexes and 50×50 persistence images, while the topology predictor employs a dropout rate of 0.5. We optimize the model using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with an initial learning rate of 10^{-4} and a batch size of 32.

4.2 Experiments on the REAL275 Dataset

First, we evaluate the pose estimation performance of TGVPose in real-world scenarios on the REAL275 dataset, and the obtained experimental results are reported in Table 1.

Compared with RGB-D-based methods, TGVPose outperforms most relevant approaches while achieving a faster inference speed. However, a performance gap remains between TGVPose and the SOTA method CleanPose. This discrepancy is attributed to two primary reasons: (1) TGVPose relies solely on depth information, without utilizing RGB images to assist the network in extracting richer object details. (2) Methods like CleanPose employ large-scale Transformer architectures, such as DINOv2, for feature extraction to obtain richer geometric information, whereas TGVPose is specifically designed for lightweight edge scenarios and refrains from using heavily parameterized structures like Transformers for feature extraction.

Table 1: Comparison with state-of-the-art methods on REAL275 dataset with supervised training. RD indicates the RGB-D-based methods and D indicates the depth-based methods.

Method	Data	mAP(%)						FPS
		IoU_{50}	IoU_{75}	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$	
NOCSWang et al. (2019b)	RD	78.0	30.1	7.2	9.5	13.8	25.2	4.8
DualPoseLin et al. (2021)	RD	79.8	62.2	29.3	35.9	50.0	66.8	2.3
SPDTian et al. (2020)	RD	77.3	53.2	19.2	21.4	43.2	54.1	5.9
SGPAChen & Dou (2021)	RD	80.1	61.9	35.9	39.6	61.3	70.7	5.0
CR-NetWang et al. (2021)	RD	79.3	55.9	27.8	34.3	47.2	60.8	-
IST-NetLiu et al. (2023)	RD	82.5	76.6	47.5	53.4	72.1	80.5	32.4
DPDNLin et al. (2022b)	RD	83.4	76.0	46.0	50.7	70.4	78.4	24.3
VI-NetLin et al. (2023)	RD	-	48.3	50.0	57.6	70.8	82.1	-
CatFormerYu et al. (2024)	RD	83.1	73.8	47.7	53.7	69.0	78.5	8.0
AG-PoseLin et al. (2024)	RD	83.7	79.5	54.7	61.7	74.7	83.1	26.8
CleanPose Lin et al. (2025)	RD	-	62.6	**61.5**	**67.4**	**78.3**	**86.2**	-
FS-NetChen et al. (2021)	D	81.1	63.5	19.9	33.9	-	69.1	19.6
SAR-NetLin et al. (2022a)	D	79.3	62.4	31.6	42.3	50.3	68.3	-
GPV-PoseDi et al. (2022)	D	-	64.4	32.0	42.9	-	73.3	22.5
HS-PoseZheng et al. (2023)	D	82.1	74.7	46.5	55.2	68.6	82.7	16.7
RBP-PoseZhang et al. (2022)	D	-	67.8	38.2	48.1	63.1	79.2	24.5
Query6DoFWang et al. (2023)	D	82.5	76.1	49.0	58.9	68.7	83.0	34.9
TG-PoseZhan et al. (2024)	D	82.6	76.2	49.8	59.0	71.7	86.6	50.0
SEGPoseYu et al. (2026)	D	83.1	78.3	51.8	61.1	73.9	87.1	45.8
SEGPose(Jetson)	D	83.1	78.3	51.8	61.1	73.9	87.1	13.4
Ours(RTX 3090)	D	80.8	66.4	48.4	56.1	67.6	76.9	**73.7**
Ours(Jetson)	D	80.8	66.4	48.4	56.1	67.6	76.9	30.3

When benchmarked against depth-only baselines, TGVPose significantly outperforms methods like FS-Net and GPV-Pose with substantially fewer parameters. We attribute these gains to two key components: the adaptive multirelation vector-neuron layer, which jointly models local geometric neighborhoods, invariant feature similarities, and global context; and the topology conditioning branch, which injects structural cycle and void priors to resolve shape ambiguity under partial observations.

To assess real-time efficiency, we evaluate inference throughput on both a desktop GPU (NVIDIA RTX 3090) and an edge computing platform (NVIDIA Jetson). Compared with the state-of-the-art SEGPose, TGVPose sacrifices a minor 3.4% and 5.0% in $5°2\,\text{cm}$ and $5°5\,\text{cm}$ metrics, respectively, but increases throughput from 45.8 to 73.7 FPS on the RTX 3090 ($1.61\times$ speedup) and from 13.4 to 30.3 FPS on the Jetson ($2.26\times$ speedup). While SEGPose achieves top-tier accuracy on power-sufficient platforms, its heavy computational overhead prevents real-time inference on edge hardware.

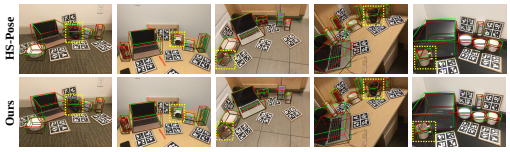


Figure 3: Qualitative comparison between HS-Pose and TGVPose on the REAL275 benchmark. Predictions are marked with red bounding boxes, while ground truth is indicated by green bounding boxes.

Qualitative comparisons in Figure 3 further demonstrate that TGVPose generates 3D bounding boxes that align more tightly with target objects than HS-Pose, particularly under severe occlusion and noisy monocular geometries.

4.3 Experiments on Wild6D Dataset

To test TGVPose's generalization, we apply it to pose estimation on the Wild6D dataset Fu & Wang (2022), which includes five object categories: bottle, bowl, camera, laptop, and mug, placed in

various real-world scenes. We evaluate performance on the Wild6D test set and compare results with depth-based methods, as shown in Table 2.

According to the experimental results presented in the table, TGVPose effectively balances accuracy and real-time capability, outperforming most depth-based pose estimation methods and demonstrating seamless deployment suitability for computation-constrained edge scenarios. The key factor underlying this performance guarantee lies in AMG-VN, which plays a crucial role in capturing adaptive graph vector features from point clouds. It enables the network to effectively comprehend the shape and structure of input point clouds and extract invariant features. Furthermore, the topological features of the point cloud impose additional constraints on point cloud reconstruction and completion, ensuring that the reconstructed point cloud closely aligns with the actual point cloud geometry. These combined feature extraction and constraint mechanisms allow TGVPose to maintain high prediction precision while preserving a lightweight architecture.

Table 2: Quantitative comparison with other methods on the Wild6D dataset. Bold numbers represent the best values in depth-based methods.

Method	Data	mAP (%)				
		IoU_{25}	IoU_{50}	5°5cm	10°5cm	15°5cm
NOCSWang et al. (2019b)	RGB-D	51.0	32.1	13.3	34.9	39.6
RePoNetZe & Wang (2022)	RGB-D	83.3	67.1	30.3	40.2	43.0
DualPoseNetLin et al. (2021)	RGB-D	90.1	66.7	24.5	37.9	41.9
CatFormerYu et al. (2024)	RGB-D	86.6	62.8	29.7	41.9	45.6
SAR-NetLin et al. (2022a)	D	77.6	6.1	8.8	25.6	35.6
CPPFYou et al. (2022)	D	79.6	21.0	7.8	18.1	21.7
GPV-PoseDi et al. (2022)	D	82.0	55.6	27.0	39.4	43.0
HS-PoseZheng et al. (2023)	D	85.4	58.1	28.6	41.6	44.7
SEGPoseYu et al. (2026)	D	**88.1**	**72.2**	**35.6**	**48.1**	**52.2**
Ours	D	86.3	63.4	30.7	45.3	48.3

To highlight TGVPose's superiority in real-world pose prediction, we compare it with HS-Pose through visualization, as shown in Figure 4. The red bounding box represents the ground truth, and the green line shows the prediction. TGVPose outperforms HS-Pose by more accurately capturing object poses and generalizing better to pose estimation for unknown objects in real-world scenarios.



Figure 4: Qualitative results of HS-Pose and TGVPose on the Wild6D Dataset.

4.4 Ablation Studies

To verify the effectiveness of the proposed method and the rationality of the parameter selection, we conduct a series of ablation studies, and the experimental results are summarized in Table 3.

Table 3: Ablation studies on REAL275. Module variants are cumulative, whereas topology, loss, and hyperparameter variants modify only the indicated component of TGVPose.

Configuration	IoU_{50}	IoU_{75}	5°2 cm	5°5 cm	10°2 cm	10°5 cm
(A) Network components						
Original SE(3)	78.4	63.0	42.0	49.5	61.8	71.2
+ relative VN message	78.9	63.8	43.0	50.3	62.7	72.0
+ invariant/global relations	79.5	64.8	44.5	51.9	64.1	73.6
+ adaptive relation gate	80.2	65.6	46.5	53.4	66.0	75.1
+ topology conditioning (TGVPose)	**80.8**	**66.4**	**48.4**	**56.1**	**67.6**	**76.9**
(B) Topology dimensions and training losses						
Topology supervision only	80.3	65.7	46.9	53.8	66.3	75.4
H_1 only	80.5	66.0	47.6	54.6	66.8	76.0
H_2 only	80.4	65.8	47.2	54.2	66.5	75.7
Without $\mathcal{L}_{\mathrm{bdf}}$	80.1	65.9	47.5	54.5	66.6	75.8
Without $\mathcal{L}_{\mathrm{prop}}$	79.9	65.4	46.8	54.0	65.9	75.0
(C) Neighborhood size and persistent-image resolution						
$K = 5$, resolution 50×50	79.9	65.5	47.0	54.0	66.2	75.5
$K = 10$, resolution 50×50	80.8	66.4	48.4	55.2	67.6	76.9
$K = 15$, resolution 50×50	80.5	66.1	48.0	54.8	67.2	76.5
$K = 20$, resolution 50×50	80.1	65.7	47.4	54.2	66.6	75.9
$K = 10$, resolution 20×20	80.2	65.7	47.3	54.1	66.5	75.8
$K = 10$, resolution 30×30	80.5	66.0	47.9	54.7	67.1	76.4

Table 3(A) evaluates the network components cumulatively. The original SE(3) backbone relies on a single neighborhood relation, so pooled features may obscure neighbor–center variations around boundaries and thin structures. Relative VN messages explicitly encode these displacements. Removing the invariant-feature and global relations confines inference to one local graph, losing both rotation-invariant similarity and object-level context, particularly under occlusion. Without the adaptive gate, relation branches are fused without point-dependent selection and cannot emphasize the most informative relation at each location. Removing topology conditioning further makes the pose heads rely entirely on partial observations, increasing ambiguity between instances with similar visible regions but different complete structures.

Table 3(B) analyzes topology and geometric supervision. Topology supervision without residual conditioning regularizes the auxiliary predictor but does not expose the prior directly to the reconstruction and pose heads. Using only H_1 discards persistent voids, whereas using only H_2 removes

cycle structures such as handles, showing that the two dimensions encode complementary shape cues. Removing $\mathcal{L}_{\text{bdf}}$ weakens boundary and object-extent constraints, which affects translation and scale. Removing $\mathcal{L}_{\text{prop}}$ eliminates symmetry-aware matching, causing inconsistent penalties among equivalent correspondences and destabilizing canonical reconstruction and rotation estimation.

Table 3(C) studies neighborhood size and topology resolution. A small K provides insufficient context, whereas an excessively large K may connect unrelated structures and dilute discriminative messages. Low-resolution persistent images coarsely quantize the birth–persistence distribution and may suppress weak topology responses; higher resolution retains finer cycle and void information but enlarges the topology heads. The final configuration is selected according to strict pose accuracy, with its computational cost reported separately in the efficiency evaluation.

4.5 Robotic Grasping Experiments

We further evaluate TGVPose on a UR3 robot using 4 manipulation tasks: (A) single-object grasping, (B) occluded scene grasping, (C) object handover, and (D) object stacking. In Figure 5, we illustrate the grasping process of the UR3 robot across different tasks.

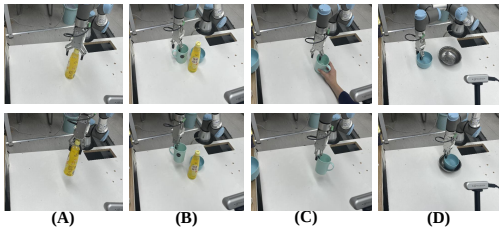


Figure 5: Category-level grasp experiments in real world.

In Table 4, we present a comparison of grasping success rates between TGVPose and previous related methods. As shown in Table 4, TGVPose consistently outperforms the compared methods across all tasks. Its advantage is particularly evident in occluded and disturbed scenes, where the adaptive multi-relation VN layer integrates complementary local and global cues, while topology conditioning provides complete-shape constraints from incomplete observations. Stacking remains more challenging because small rotation, translation, or scale errors can cause misalignment, collision, or unstable contact. Most failed trials arise from severe occlusion, missing depth around reflective surfaces, inaccurate instance masks, or perception-to-control delay after object movement. Overall, the results demonstrate that TGVPose provides accurate and robust pose predictions for practical robotic manipulation.

Table 4: Comparison of grasp success rates on different tasks.

Method	Grasp Success Rate (%)			
	(A)	(B)	(C)	(D)
6-DOF GraspNetMousavian et al. (2019)	68.9	46.7	57.8	26.7
GPV-PoseDi et al. (2022)	71.1	62.2	66.7	26.7
HS-PoseZheng et al. (2023)	75.6	66.7	68.9	55.6
ours	**80.0**	**71.1**	**73.3**	**57.8**

As shown in Table 4, TGVPose achieves higher grasp success rates than previous methods across all tasks. Existing approaches mainly rely on local geometric features and are susceptible to incomplete point clouds and pose ambiguity. In contrast, the adaptive multi-relation VN layer jointly captures local geometry, invariant similarity, and global context, while topology conditioning provides category-level structural cues for recovering occluded object parts. Moreover, the lightweight inference pipeline reduces the delay between perception and execution, improving robustness in handover and disturbance scenes.

Remaining failures mainly arise from severe occlusion, reflective surfaces, missing depth measurements, or inaccurate masks, which may remove critical structures such as handles and boundaries. Rotation ambiguity in symmetric objects can also produce incorrect approach directions, while calibration errors and control latency may shift the gripper away from the predicted contact region.

5 Conclusion

In this paper, we presented TGVPose, a lightweight topology-guided framework for category-level object pose estimation. Its adaptive multi-relation graph vector-neuron layer captures complementary geometric relations through invariant scalar gating, while category-prototype topology conditioning provides global H_1 and H_2 shape priors without online persistent-homology computation. Experiments on REAL275, Wild6D, and the UR3 platform demonstrate a favorable balance between pose accuracy, $\mathrm{SE}(3)$ generalization, and inference efficiency, supporting deployment on resource-constrained robotic systems.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for writing refinement, polishing, and language editing of the manuscript. We have not used generative AI tools for generating research ideas, producing research code, conducting experiments, or analyzing data, and generating synthetic datasets or visual artifacts are not applicable to this work. Additionally, we used generative AI tools for improving the readability, grammar, and style of the text. We have reviewed all AI-assisted work. All LLM-suggested edits to the manuscript were manually reviewed and verified by the authors to ensure that the scientific meaning, technical claims, and original tone remained accurate and unaltered. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Kai Chen and Qi Dou. SGPA: Structure-guided prior adaptation for category-level 6D object pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2773–2782, 2021.

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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1581–1590, 2021.

Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J Guibas. Vector neurons: A general framework for so (3)-equivariant networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12200–12209, 2021.

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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6781–6791, 2022.

Hongzhi Du, Yanyan Li, Yan Di, Teng Zhang, Yanbiao Sun, and Jigui Zhu. Se(3)-equivariance learning for category-level object pose estimation. *IEEE Transactions on Instrumentation and Measurement*, 74:1–13, 2025.

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.

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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11632–11641, 2020.

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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717, 2022a.

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 *IEEE International Conference on Computer Vision*, pp. 3560–3569, 2021.

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*, pp. 19–34, 2022b.

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 *IEEE/CVF International Conference on Computer Vision*, pp. 14001–14011, 2023.

Shifeng Lin, Zunran Wang, Yonggen Ling, Yidan Tao, and Chenguang Yang. E2EK: End-to-end regression network based on keypoint for 6D pose estimation. *IEEE Robotics and Automation Letters*, 7(3):6526–6533, 2022c.

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*, pp. 21040–21049, 2024.

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*, pp. 5990–6000, 2025.

Jianhui Liu, Yukang Chen, Xiaoqing Ye, and Xiaojuan Qi. IST-Net: Prior-free category-level pose estimation with implicit space transformation. In *IEEE/CVF International Conference on Computer Vision*, pp. 13978–13988, October 2023.

Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-DoF GraspNet: Variational grasp generation for object manipulation. In *IEEE/CVF International Conference on Computer Vision*, pp. 2901–2910, 2019.

Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6D object pose and size estimation. In *European Conference on Computer Vision*, pp. 530–546, 2020.

Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. In *Conference on Robot Learning*, 2018.

Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Fei-Fei Li, and Silvio Savarese. Densefusion: 6D object pose estimation by iterative dense fusion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3343–3352, 2019a.

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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2642–2651, 2019b.

Jiaze Wang, Kai Chen, and Qi Dou. Category-level 6D object pose estimation via cascaded relation and recurrent reconstruction networks. In *IEEE International Conference on Intelligent Robots and Systems*, pp. 4807–4814, 2021.

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 *IEEE/CVF International Conference on Computer Vision*, pp. 14055–14064, 2023.

Yang You, Ruoxi Shi, Weiming Wang, and Cewu Lu. CPPF: Towards robust category-level 9D pose estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6866–6875, 2022.

Sheng Yu, Di-Hua Zhai, and Yuanqing Xia. CatFormer: Category-level 6D object pose estimation with transformer. In *AAAI Conference on Artificial Intelligence*, volume 38, pp. 6808–6816, 2024.

Sheng Yu, Di-Hua Zhai, and Yuanqing Xia. Se (3)-equivariance with geometric and topological guidance for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35114–35123, 2026.

Yanjie Ze 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.

Yue Zhan, Xin Wang, Lang Nie, Yang Zhao, Tangwen Yang, and Qiuqi Ruan. Tg-pose: Delving into topology and geometry for category-level object pose estimation. *IEEE Transactions on Multimedia*, 26:9749–9762, 2024.

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*, pp. 655–672, 2022.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17163–17173, 2023.