MemPose: Category-level Object Pose Estimation with Memory

Xiao Lin[1], Minghao Zhu[1], Yun Peng[1], Liuyi Wang[1], Qiyi Wang[1], Chengju Liu[1,2✉], and Qijun Chen[1,2✉]

[1] Tongji University, Shanghai, China
[2] State Key Laboratory of Autonomous Intelligent Unmanned Systems
{linx_xx, zmh_hh, pengyun, wly, wqy126179, liuchengju, qjchen}@tongji.edu,cn

Abstract. In the pursuit of robust and generalizable category-level object pose estimation, most existing methods adopt parametric formulations that learn effective representations from data, yet they primarily encode category-level patterns into fixed shape priors or static parameter weights, which limits their scalability to highly diverse instances. In this paper, we rethink category-level pose estimation from a memory-centric perspective and present **MemPose**, a memory-augmented framework that explicitly incorporates category-level geometric memory into the pose estimation pipeline. We introduce an external memory buffer that stores and dynamically updates structural representations from previously observed instances, enabling the model to leverage accumulated experience to support current perception. Extensive experiments on four challenging benchmarks (REAL275, CAMERA25, Housecat6D and Wild6D) demonstrate the superiority of our proposed method over previous state-of-the-art approaches.

Keywords: Category-level Pose Estimation · 3D Vision · Memory System

1 Introduction

As a critical application in human-robot interaction [45], the **C**ategory-level **O**bject **P**ose **E**stimation (COPE) [38] has attracted increasing attention. Unlike instance-level methods [22], COPE is model-free and aims to estimate the 9-DoF pose for arbitrary objects within predefined categories, without relying on instance-specific CAD models. However, this setting is inherently challenging due to the significant differences among objects within the category. To overcome these challenges, humans typically leverage memory from previous observations to perform analogical reasoning across instances. Such memory not only supports immediate perception but is also continuously updated as new objects are encountered.

Drawing on an intuitive understanding of human perception, a line of existing works [4, 18, 36] introduces explicit shape priors, often by extracting average

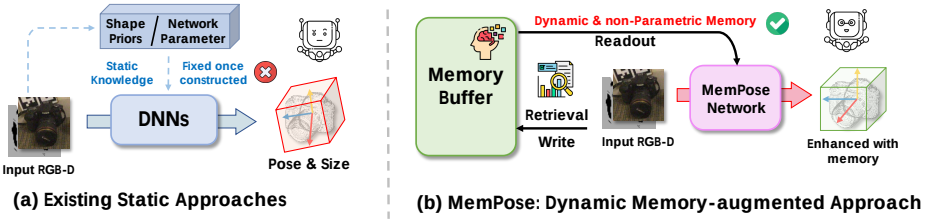

Fig. 1: Overview of the category-level pose estimation pipeline: (a) Existing methods rely on static patterns, such as shape priors or fixed network parameters, to regress object pose and size. (b) In contrast, our approach introduces a dynamic, memory-augmented pipeline that explicitly incorporates category-level geometric memory to enhance pose and size estimation.

shape for each category. Specifically, these approaches first reconstruct instance models by deforming a categorical shape prior and then match observations with the reconstructed models to regress pose. While priors provide explicit category-level cues, they are fixed once constructed and more like static prototypes. Essentially, such priors are hard to function as memory, as they would not update with new observations, thus capturing the diversity of instances. Moreover, acquiring the high-quality priors requires costly pre-processing pipelines.

More recently, benefiting from the success of deep neural networks (DNNs), another line of existing methods adopt a parametric paradigm, which learns effective feature representations from input modalities via finely designed networks. For instance, HS-Pose [44] proposes a 3D graph convolution network to enhance pose-sensitive feature extraction, while AG-Pose [23] introduces a local feature aggregation module to establish robust keypoint-level correspondences. KeyPose [42] propose a graph-based detection method to strengthen the understanding of geometric structures. Furthermore, foundational models like DINOv2 [28] have been widely adopted [6, 21] to enhance robustness and contextual understanding. Although effective, these approaches implicitly encode category-level knowledge into network parameters, which remain static during inference, lacking a mechanism to provide dynamic memory support like human perception.

Despite their differences, two lines of existing methods share a common limitation: they both lack a dynamic mechanism to accumulate and reuse category-level geometric memory, as shown in Fig. 1 (a). This observation prompts us to rethink category-level pose estimation from a memory-centric perspective and explore how geometric memory can be effectively modeled to support robust pose prediction.

To this end, we present **MemPose**, an innovative architecture by proposing the memory module for effective and robust category-level pose estimation, as shown in Fig. 1 (b). The key insight of our method is to integrate parametric perception with non-parametric memory in a unified framework. Towards this goal, we maintain a *memory buffer* that stores structural feature representations. The

buffer serves as an external repository of category-level knowledge, enabling the model to retrieve relevant structural patterns to support the current observation. To ensure dynamic nature of the memory buffer, we introduce a similarity-based token merge update mechanism. As new objects are inputted, the memory buffer is continually updated with the latest keypoint context features. Finally, by fusing memory-derived information with current features, MemPose can predict 9-DoF pose of objects. As demonstrated by extensive experiments, our findings reveal the impact of memory-centric designs for category-level pose estimation and may inspire future work in this direction.

To summarize, our main contributions are as follows:

- We introduce MemPose, a novel architecture with memory module that unifies parametric perception and non-parametric memory for category-level pose estimation.
- We design a similarity-based memory update strategy to ensure the adaptability of the memory. Such s mechanism allows category-level representations to be continuously refined rather than being fixed.
- Extensive experiments on three mainstream challenge benchmarks, REAL275, Housecat6D and Wild6D, demonstrate that the proposed MemPose outperforms other existing methods.

2 Related Works

Category-level Object Pose Estimation. To improve the generalization ability on unseen instances, traditional methods suggest mapping input shape to a normalized canonical space (NOCS) [38] and recovering the pose via the Umeyama algorithm [37]. Furthermore, SPD [36] proposes a deformation and matching strategy that match observations to the reconstructed models to solve poses. Inspired by SPD, many subsequent prior-based works [16, 17] further improve the use of shape priors, continuously enhancing the pose estimation performance. More recently, prior-free methods [7, 20, 23, 24] have achieved impressive performance. VI-Net [20] separates rotation into viewpoint and in-plane rotations, while AG-Pose [23] explicitly extracts local and global geometric keypoint information of different instances. CleanPose [21] introduces causal learning into the formulation of COPE to mitigate the negative effects caused by confounders. However, these methods either encode category-level structural patterns as fixed shape priors or store them within static network parameters, lacking a flexible memory mechanism to accumulate and reuse category-level experience.

Memory-augmented Methods. Memory mechanism is initially introduced in Large Language Models (LLMs) to enhance long-context reasoning performance [3]. Memory mechanism has been widely used in video prediction [40, 46, 47], point tracking [9] and robotic tasks [27, 33, 34], demonstrating their effectiveness. For instance, Memflow [8] leverages memory for optical flow estimation, while MemoryVLA [34] explicitly incorporates memory to model temporal dependencies in robotic manipulation. In thr 3D domain, MAD [1] constructs a memory bank to store past prediction and trajectory for 3D detection. Some ap-

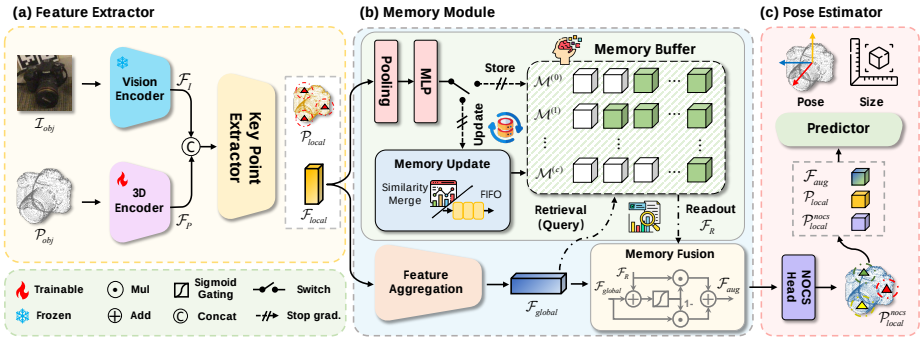

Fig. 2: Illustration of MemPose. (a) The model extracts semantic and geometric features from RGB-D inputs and detects robust keypoints to obtain keypoint-level representations. (b) A dynamic memory module is introduced to store and retrieve historical feature representations. Through attention-based querying and memory fusion, the model leverages short-term geometric information to enhance robust feature learning. (c) The fused keypoint features, together with the predicted keypoints and NOCS coordinates, are utilized for joint object pose and size estimation.

proaches also attempt to utilize memory mechanism to solve point cloud tracking [41] and 3D reconstruction [39]. However, above methods primarily focus on temporal modeling, it is non-trivial to adapt these approaches to pose estimation due to the inherent differences in human modeling among these tasks. In this work, our MemPose exploits category-level structural patterns to form memory mechanism. Importantly, our motivation originates from a deep analysis of intra-category generalization in pose estimation, which is naturally suitable for modeling with memory.

3 Methodology

3.1 Definition and Overview

Problem Setup. Given an RGB-D image containing objects from a predefined set of categories, we first employ segmentation masks to obtain the cropped RGB image $\mathcal{I}_{obj} \in \mathbb{R}^{H\times W\times 3}$ and the point cloud $\mathcal{P}_{obj} \in \mathbb{R}^{N\times 3}$, where N is the number of points and $\mathcal{P}_{obj}$ is acquired by back-projecting the cropped depth image with camera intrinsics followed by a downsampling process. With the input $\mathcal{I}_{obj}$ and $\mathcal{P}_{obj}$, the objective of COPE [38] is to recover the 9-DoF pose of the target object, including the 3D rotation $R \in SO(3)$, the 3D translation $t \in \mathbb{R}^3$, and size $s \in \mathbb{R}^3$.

Overview. We present an overview of MemPose as in Fig. 2. Specifically, our method first encodes the current observation to construct an RGB-D representation. To explicitly capture geometric information, we further perform keypoint detection to extract local features (Sec. 3.2). Then, these local structural patterns are stored in a memory buffer $\mathcal{M}$, forming a short-term memory. By feature

enqueue operation, memory buffer is continually updated with the latest geometric context (Sec. 3.3). Subsequently, the aggregated global feature reads from the memory and fuses the retrieved information. Finally, MemPose regresses the object pose based on the fused feature representation (Sec. 3.4).

3.2 Partial Feature Extraction

Following [21,23], we utilize the PointNet++ [31] to extract point feature $\mathcal{F}_P \in \mathbb{R}^{N\times C_1}$ of input point cloud $\mathcal{P}_{obj} \in \mathbb{R}^{N\times 3}$. As for the RGB image $\mathcal{I}_{obj}$, we adopt DINOv2 (ViT-S/14) [28] as our image feature extractor, which has been proven to extract abundant semantic-aware information from RGB images [29,30]. We select those pixel features corresponding to $\mathcal{P}_{obj}$ and utilize linear interpolation to propagate the original DINOv2 features into the final RGB features $\mathcal{F}_I \in \mathbb{R}^{N\times C_2}$. Moreover, we concatenate $\mathcal{F}_P$ and $\mathcal{F}_I$ to form $\mathcal{F}_{obj} \in \mathbb{R}^{N\times C}$. The local geometric information is indispensable to establish robust correspondences. Typically, methods like Farthest Point Sampling (FPS) can be used to extract local keypoints. In this work, we follow previous method [23] and utilize a instance-adaptive approach that focuses on the most discriminative and reliable object regions. Similar to DETR [2], we initial M learnable category embeddings $\mathcal{F}_{cat} \in \mathbb{R}^{M\times C}$, which undergoes cross-attention with $\mathcal{F}_{obj}$ to attend to critical regions in $\mathcal{P}_{obj}$. This process obtains an instance feature $\mathcal{F}_{ins} = \text{CrossAttn}(\mathcal{F}_{cat}, \mathcal{F}_{obj})$, We then compute correspondences between $\mathcal{F}_{ins}$ and $\mathcal{F}_{obj}$ via cosine similarity, forming a matrix $H \in \mathbb{R}^{M\times N}$, and M keypoints are selected as $\mathcal{P}_{local} = \text{softmax}(H)\mathcal{P}_{obj}$, along with their corresponding feature $\mathcal{F}_{local} = \text{softmax}(H)\mathcal{F}_{obj}$.

3.3 Memory Module

To better exploit short-term working memory, we propose a novel memory module and processing pipeline for the COPE task. Specifically, we maintain a category-aware memory buffer defined as follows:

$$\mathcal{M} = \{\mathcal{M}^{(c)} \mid c = 0, \ldots, k\}, \mathcal{M}^{(c)} \in \mathbb{R}^{L\times C} \tag{1}$$

where k denotes the number of object categories, and $\mathcal{M}^{(c)}$ stores the memory features associated with category c. Here, L represents the length of memory buffer.

Memory Construction. Each memory buffer $\mathcal{M}^{(c)}$ is initialized as an empty set. During training, newly batch features are first pushed into the corresponding category buffer. The memory retrieval and update mechanisms are activated only after the buffer reaches its capacity L. This warm-up strategy effectively prevents noisy and unstable representations of the early training stage, improving the robustness of the stored patterns. Note that, most existing approaches [8,21] directly store the final output features as memory entries. However, such a design typically affects the attention-based retrieval process by biasing it towards recently updated entries, as these features are more closely aligned with the current network state rather than being geometrically relevant. As a result, memory

retrieval may be dominated by temporal recency instead of meaningful structural similarity. To this end, our design stores pooled local features as memory entries. Specifically, we first employ a simple average pooling on local features and get $\mathcal{F}_{local}^{avg} \in \mathbb{R}^{1\times C}$. Then, we linear project pooled features to form memory entries $\mathcal{F}_{mem}$,

$$\mathcal{F}_{local}^{avg} = \text{AvgPool}(\mathcal{F}_{local}) \tag{2}$$

$$\mathcal{F}_{mem} = \text{Norm}(\text{MLP}(\mathcal{F}_{local}^{avg})), \tag{3}$$

where the pooled local features capture explicit local geometric patterns, providing a more stable and geometry-aware reference for subsequent memory retrieval.

Update Mechanism. The memory update process is performed before memory retrieval to ensure that the stored representations reflect the most informative geometric context. When the number of stored entries exceeds L, the update mechanism is activated. Specifically, given a newly incoming pooled local feature $\mathcal{F}_{mem}$, we compute its cosine similarity with all existing memory entries in the corresponding buffer $\mathcal{M}^{(c)}$. The memory entry with the highest similarity is selected and updated by averaging it with the incoming feature.

$$i^* = \arg\max_i \cos\Big(\mathcal{F}_{mem}, \mathcal{M}_i^{(c)}\Big), \quad i = 1, \ldots, l, \tag{4}$$

$$\mathcal{M}_{i^*}^{(c)} \leftarrow \frac{1}{2}\left(\mathcal{M}_{i^*}^{(c)} + \mathcal{F}_{mem}\right). \tag{5}$$

This update mechanism mitigates memory bloat by reducing redundancy. Furthermore, to prevent imbalance in similarity-based feature selection, we introduce several lightweight balancing strategies. For instance, the update target is randomly sampled from the top-k most similar memory entries, or mild stochastic perturbations are applied during the update process. Here, we adopt Gumbel noise g [12] to add random perturbations

$$g = -\log(-\log(u)) \tag{6}$$

$$\cos\Big(\mathcal{F}_{mem}, \mathcal{M}_i^{(c)}\Big) \leftarrow \cos\Big(\mathcal{F}_{mem}, \mathcal{M}_i^{(c)}\Big) + \lambda g, \tag{7}$$

where $u \sim \mathcal{U}(0, 1)$ is drawn from a uniform distribution, and λ (set to 0.1) is a scaling factor controlling the magnitude of the perturbation. Meanwhile, the classic FIFO strategy can also be used to update the buffer. The effectiveness of these strategies is validated in ablation studies in Sec. 4.2.

Memory Retrieval and Fusion. Recall that we expect the model to focus on geometric semantics rather than temporal similarity during retrieval. To this end, we first enhance feature representations with explicit geometric context before interacting with the memory. Specifically, we apply a geometric aggregation approach [23] on the local features to form a global feature representation $\mathcal{F}_{global} \in \mathbb{R}^{M\times C}$, which is jointly enriched by the local geometric details from k-nearest neighbors and the global structural information across all keypoints. Subsequently, we treat global features as the query for memory retrieval and

stack all memory entries stored in the category-specific memory buffers $\mathcal{M}^{(c)}$ to form a unified tensor, which serves as the key–value set

$$\mathcal{F}'_{mem} = \left[\mathcal{F}^0_{mem}; \cdots; \mathcal{F}^L_{mem}\right] \in \mathbb{R}^{L\times C} \tag{8}$$

$$q = \mathcal{F}_{global}W_q,\ k = \mathcal{F}'_{mem}W_k,\ v = \mathcal{F}'_{mem}W_v, \tag{9}$$

where $[;]$ is the concatenation operation along the first dimension, and W_q, W_k, W_v are the learnable projection parameters, Hence, the retrieved memory features can be read-out by

$$\mathcal{F}_R = \text{Softmax}(1/\sqrt{D_k} \times q \times k^T) \times v. \tag{10}$$

Furthermore, to enhance the stability of memory learning, we introduce an adaptive gate fusion method to integrate memory retrieved features and current global features.

$$w_a = \sigma(\mathcal{F}_R W_R + \mathcal{F}_{global}W_f) \tag{11}$$

$$\mathcal{F}_{aug} = w_a \odot \mathcal{F}_R + (1 - w_a) \odot \mathcal{F}_{global}, \tag{12}$$

where σ and $\odot$ mean the Sigmoid gate function and element-wise multiplication. $W_{R/f} \in \mathbb{R}^{C\times 1}$ is a learnable weight parameter. The resulting memory-augmented features $\mathcal{F}_{aug}$ is then forwarded to the pose and size predictor.

3.4 Pose Estimation and Loss Function

With the obtained memory-enhanced features, we follow [23] to perform self-attention and MLP modules to predict the corresponding NOCS coordinates $\mathcal{P}^{nocs}_{local} \in \mathbb{R}^{M\times 3}$. Then, given the NOCS coordinates of keypoints $\mathcal{P}^{nocs}_{local}$, the memory-enhanced features $\mathcal{F}_{aug}$ and the position of keypoint $\mathcal{P}_{local}$, we recover the final pose and size R, t, s via a set of keypoint-level correspondences containing global features and points. We use L2-norm Loss to supervise the predicted pose, in formula:

$$\mathcal{L}_{pose} = \|R_{gt} - R\|_2 + \|t_{gt} - t\|_2 + \|s_{gt} - s\|_2, \tag{13}$$

where R_{gt}, t_{gt}, s_{gt} means the ground truth rotation, translation and size.

In addition, there are some additional loss functions to balance keypoints selection and pose prediction. First, to encourage the keypoints to focus on different parts, the diversity loss $\mathcal{L}_{div}$ is used to force the detected keypoints to be away from each other, in detail:

$$\mathcal{L}_{div} = \sum_{i=1}^{N_{local}} \sum_{j=1, j\neq i}^{N_{local}} \mathbf{d}\left(\mathcal{P}^{(i)}_{local}, \mathcal{P}^{(j)}_{local}\right) \tag{14}$$

$$\mathbf{d}\left(\mathcal{P}^{(i)}_{local}, \mathcal{P}^{(j)}_{local}\right) = \max\left\{th_1 - \left\|\mathcal{P}^{(i)}_{local} - \mathcal{P}^{(j)}_{local}\right\|_2, 0\right\}, \tag{15}$$

where th_1 is a hyper-parameter and is set as 0.01, $\mathcal{P}^{(i)}_{local}$ means the i-th keypoint. Then, to encourage the keypoints to locate on the surface of the object and

exclude outliers simultaneously, an object-aware chamfer distance loss $\mathcal{L}_{ocd}$ is employed to constrain the distribution of $\mathcal{P}_{local}$. In formula:

$$\mathcal{L}_{ocd} = \frac{1}{|\mathcal{P}_{local}|} \sum_{x_i \in \mathcal{P}_{local}} \min_{y_j \in \mathcal{P}'_{obj}} \|x_i - y_j\|_2 , \tag{16}$$

where $\mathcal{P}'_{obj}$ denotes the point cloud of objects without outlier points. Moreover, we also use MLP to predict the NOCS coordinates of keypoints $\mathcal{P}^{nocs}_{local} \in \mathbb{R}^{N_{local} \times 3}$. Then, we generate ground truth NOCS of keypoints $\mathcal{P}^{gt}_{local}$ by projecting their coordinates under camera space $\mathcal{P}_{local}$ into NOCS using the ground truth R_{gt}, t_{gt}, s_{gt}. And we use the $SmoothL_1$ loss to supervise the NOCS projection:

$$\mathcal{P}^{gt}_{local} = \frac{1}{\|s_{gt}\|_2} R_{gt} \left(\mathcal{P}_{local} - t_{gt}\right) \tag{17}$$

$$\mathcal{L}_{nocs} = SmoothL_1(\mathcal{P}^{gt}_{local}, \mathcal{P}^{nocs}_{local}). \tag{18}$$

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

$$\mathcal{L}_{rec} = \frac{1}{|\mathcal{P}_{recon}|} \sum_{x \in \mathcal{P}_{recon}} \min_{y \in \mathcal{P}^*_{local}} \|x - y\|_2. \tag{19}$$

Finally, the NOCS loss $\mathcal{L}_{nocs}$ is used to supervise the prediction of $\mathcal{P}^{nocs}_{local}$. ence, the overall loss function is as follow

$$\mathcal{L}_{all} = \alpha_1 \mathcal{L}_{pose} + \alpha_2 \mathcal{L}_{div} + \alpha_3 \mathcal{L}_{ocd} + \alpha_4 \mathcal{L}_{rec} + \alpha_5 \mathcal{L}_{nocs}, \tag{20}$$

where $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$ are hyper-parameters to balance the contribution of each term.

4 Experiments

Datasets and Metrics. Following previous works, we conduct experiments on four mainstream benchmarks, REAL275, CAMERA25 [38], HouseCat6D [13] and Wild6D [10] datasets. REAL275 is a challenging real-world dataset that contains objects from six categories. The training data consists of 4.3k images from 7 scenes, while the testing data includes 2.75k images from 6 scenes and 3 objects from each category. HouseCat6D is a comprehensive multi-modal real-world dataset and encompasses ten household categories, including photometrically challenging objects like glass and cutlery, with occlusions. Wild6D is a large dataset designed for self-supervised learning of COPE. It is only annotated on the test set with images from 486 different background videos, containing 162

Dataset		REAL275				Housecat6D				
Methods	Prior	IoU_{75}	5°2cm	5°5cm	10°2cm	IoU_{50}	IoU_{75}	5°2cm	10°2cm	10°5cm
SGPA ICCV'21 [4]	✓	61.9	35.9	39.6	61.3	-	-	-	-	-
RBP-Pose ECCV'22 [43]	✓	67.8	38.2	48.1	63.1	-	-	-	-	-
DPDN ECCV'22 [18]	✓	76.0	46.0	50.7	70.4	-	-	-	-	-
GCE-Pose CVPR'25 [16]	✓	79.8	57.0	65.1	75.6	76.1	<u>55.6</u>	<u>22.2</u>	49.3	53.5
FS-Net CVPR'21 [5]	✗	-	-	28.2	-	48.0	14.8	3.3	17.1	21.6
GPV-Pose CVPR'22 [7]	✗	64.4	32.0	42.9	-	50.7	15.2	3.5	17.8	22.7
VI-Net ICCV'23 [20]	✗	48.3	50.0	57.6	70.8	56.4	20.4	8.4	20.5	29.1
PRD-Pose ICCV'25 [15]	✗	-	52.6	-	73.4	-	-	8.9	26.2	-
SecondPose CVPR'24 [6]	✗	77.7	56.2	63.6	74.7	66.1	24.9	11.0	25.3	35.7
AG-Pose CVPR'24 [23]	✗	80.1	57.0	64.6	75.1	<u>76.9</u>	53.0	21.3	<u>51.3</u>	54.3
KeyPose AAAI'25 [42]	✗	<u>80.8</u>	57.7	66.0	<u>78.8</u>	75.4	-	-	-	-
SpherePose ICLR'25 [32]	✗	79.0	<u>58.2</u>	<u>67.4</u>	76.2	72.2	-	19.3	40.9	<u>55.3</u>
MemPose (ours)	✗	**81.0**	**59.9**	**67.7**	**79.0**	**81.5**	**56.4**	**23.1**	**52.6**	**57.3**

Table 1: Comparison with state-of-the-art methods on REAL275 and Housecat6D dataset. A higher value indicates better performance. '-' means unavailable statistics. Overall best results are in **bold** and the second best results are <u>underlined</u>.

objects from five categories. In this work, we only use the test set of Wild6D for evaluation to enrich the real-world experiments. When validating REAL275 and Wild6D, we also used CAMERA25 [38] for training, which is a synthetic dataset that contains the same categories as REAL275. It provides 300k synthetic RGB-D images of objects rendered on virtual backgrounds, with 25k images are withheld for testing.

We evaluate the model performance with two metrics. (**i**) **3D IoU**. As for the NOCS dataset, we report the mean average precision (mAP) of Intersection over Union (IoU) with the thresholds of 75%. For the HouseCat6D dataset, we report the mAP of 3D IoU under thresholds of 25%, 50% and 75%. (**ii**) **n°m *cm***. We also utilize the combination of rotation and translation metrics of 5°2 *cm*, 5°5 *cm*, 10°2 *cm* and 10°5 *cm*, which means the estimation is considered correct when the error is below a threshold.

Implementation Details. For a fair comparison, we utilize the same segmentation masks as AG-Pose [23] and DPDN [18] from MaskRCNN [11] and resize them to 224×224. For model parameters, the number of points N in point cloud is 1024 and the number of keypoints M is set as 96. the feature dimensions are set as $C_1 = C_2 = 128$ and $C = 256$, respectively. For memory module, we set the size of category-specific buffer to 96. The number of object categories k is 6 in datasets REAL275 and Wild6D, and 10 in dataset Housecat6D. Technically, the memory buffer is registered as a buffer tensor during training via `register_buffer`, which would be saved and restored together with the model. It does not participate in gradient backpropagation and is not updated by the optimizer. For the hyper-parameters in the loss functions, $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$ are $0.3, 10.0, 2.0, 15.0, 2.0$, respectively. For model optimizing, we employ the same data augmentation approach as previous works [18, 23], which leverage random rotation degree sampled from $U(0,\ 20)$ and rotation $\Delta t \sim U(\text{-}0.02,\ 0.02)$ and

Methods	Prior	IoU_{50}	IoU_{75}	5°2cm	5°5cm	10°2cm	10°5cm
SPD ECCV'20 [36]	✓	93.2	83.1	54.3	59.0	73.3	81.5
SGPA ICCV'21 [4]	✓	93.2	88.1	70.7	74.5	82.7	88.4
RBP-Pose ECCV'22 [43]	✓	93.1	89.0	73.5	79.6	82.1	89.5
NOCS CVPR'19 [38]	✗	83.9	69.5	32.3	40.9	48.2	64.4
DualPoseNet ICCV'21 [19]	✗	92.4	86.4	64.7	70.7	77.2	84.7
GPV-Pose CVPR'22 [7]	✗	93.4	88.3	72.1	79.1	-	89.0
HS-Pose CVPR'23 [44]	✗	93.3	89.4	73.3	80.5	80.4	89.4
VI-Net ICCV'23 [20]	✗	-	-	74.1	81.4	79.3	87.3
CLIPose TCSVT'24 [24]	✗	-	91.0	74.8	82.2	82.0	90.6
AG-Pose CVPR'24 [23]	✗	93.8	91.3	77.8	82.8	<u>85.5</u>	91.6
SpherePose ICLR'25 [32]	✗	**94.8**	<u>92.4</u>	<u>78.3</u>	<u>84.3</u>	84.8	<u>92.3</u>
MemPose (ours)	✗	<u>94.2</u>	**92.5**	**78.4**	**84.6**	**87.0**	**92.6**

Table 2: Comparisons with state-of-the-art methods on CAMERA25 dataset. A higher value indicating better performance. '-' means unavailable statistics. Overall best results are in **bold** and the second best results are <u>underlined</u>.

scaling $\Delta s \sim U$(-0.18, 1.2). We train the model on a single NVIDIA L40 GPU for a total of 120k iterations by the Adam [14] optimizer, with a mini training batch is 36 and a learning rate range from 2e-5 to 5e-4 based on triangular2 cyclical schedule [35].

4.1 Comparison with State-of-the-Art Methods

Results on REAL275 and Housecat6D datasets. Tab. 1 shows the comparison of our method with previous methods on REAL275 and Housecat6D datasets. Overall, our MemPose consistently outperforms previous methods and achieves state-of-the-art performance on both datasets. On REAL275, our method obtains the best results on all reported metrics. In particular, when compared with prior-based methods, our approach outperforms GCE-Pose [16] by 2.9% on 5°2 *cm* and 2.6% on 5°5 *cm*. And when compared with prior-free methods, MemPose surpasses the previous best method SpherePose [32] by 1.7% on 5°2 *cm* and 2.1% on 10°2 *cm*, which also employs DINOv2 [28] as the image backbone. On the more challenging Housecat6D dataset, MemPose still outperforms the prior state-of-the-art methods on the precision of COPE. Specifically, when compared with previous best prior-based methods, MemPose surpasses GCE-Pose by 5.4% on IoU_{50} and 3.3% on 10°2 *cm*. And when compared with recent prior-free strong baselines, our method outperforms SpherePose by 3.8% on 5°2 *cm* and 11.7% on 10°2 *cm*, and surpasses AG-Pose [23] by 4.6% on IoU_{50} and 3.4% on IoU_{75}. The significant improvements on these two benchmarks demonstrate the effectiveness of the proposed method.

Results on CAMERA25 dataset. In Tab. 2, we compare our method with the existing approaches for category-level object pose estimation on CAMERA25 [38] dataset. From the results, we can see that MemPose achieves the best performance. In detail, MemPose outperforms the current state-of-the-art method SpherePose by 2.2% on 10°2cm, and surpasses AG-Pose [23] by 1.2% on IoU_{75}, 1.8% on 5°5cm and 1.5% on 10°2cm, respectively.

Methods	IoU_{75}	5°2cm	5°5cm	10°2cm	10°5cm
SPD [36]	20.3	6.9	9.3	20.1	27.8
GPV-Pose [7]	-	14.1	21.5	23.8	41.1
SGPA [4]	56.4	10.3	20.5	29.1	39.5
AG-Pose [23]	36.2	21.4	27.3	29.1	40.1
Diff9D [26]	38.2	25.5	30.5	32.5	<u>40.9</u>
MH6D [25]	<u>41.9</u>	<u>27.0</u>	<u>31.2</u>	<u>34.4</u>	40.5
MemPose	**46.2**	**29.7**	**33.5**	**35.8**	**43.2**

Table 3: Comparisons with state-of-the-art methods on Wild6D dataset. A higher value indicating better performance. Overall best results are in **bold** and the second best results are <u>underlined</u>.

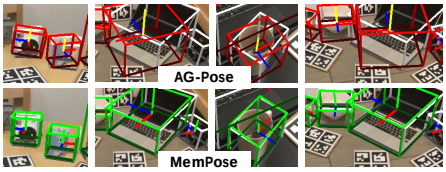

Fig. 3: Qualitative comparison on REAL275. We compare the predictions of MemPose and the baseline AG-Pose. The ground truth is marked by white borders.

Type	IoU_{75}	5°2cm	5°5cm
Local	**81.0**	**59.9**	**67.7**
Global	80.5	58.0	66.5
Fused	80.4	58.1	66.8

(a) Effect of distinct stored memory.

Strategy	IoU_{75}	5°2cm	5°5cm
w/ warm-up	**81.0**	**59.9**	**67.7**
w/o warm-up	80.8	59.0	66.0
Random	80.4	58.0	65.0

(b) Effect of memory construction strategies.

Table 4: Ablation studies on memory modules. Settings used in our final model are colored in gray.

Results on Wild6D dataset. We further evaluate the proposed method on the Wild6D dataset, using CAMERA25 and REAL275 datasets for training. Tab. 3 reports the quantitative results of existing methods on Wild6D dataset. Once again, our MemPose achieves the best performance over state-of-the-art approaches by a large margin. Concretely, our MemPose exceeds previous best method MH6D [25] by 4.3% on IoU_{75}. Moreover, when compared under the n°m cm metric, our method achieved an average improvement of 2.5% across all thresholds. The superior performance on this more comprehensive and challenging real-world dataset further demonstrates the effectiveness of our approach.
Qualitative Comparison. The qualitative results of AG-Pose and proposed MemPose are shown in Fig. 3. It can be observed that our method yields a more accurate pose estimation across diverse shapes and poses. These exceptional outcomes further support the efficacy of our approach.

4.2 Ablation Studies

To further demonstrate the superiority of our method, we conduct comprehensive ablation studies on REAL275 dataset.

Effect of Distinct Stored Memory. We further present the different stored memory as mentioned in Sec. 3.3, the results are shown in Tab. 4a. The table indicates that storing local features as memory entries achieves the best overall

Update Period	IoU_{75}	5°2cm	5°5cm
Pre-update	**81.0**	**59.9**	**67.7**
Post-update	80.7	59.0	65.2

(a) Effect of different update periods.

Method	Balance	IoU_{75}	5°2cm	5°5cm
FIFO	-	80.7	59.1	66.0
Merge	Noise	**81.0**	**59.9**	**67.7**
	Top-k	**81.0**	59.2	65.3

(b) Effect of update mechanism.

Size	IoU_{75}	5°2cm	5°5cm
All	**81.0**	**59.9**	**67.7**
Random-48	80.4	58.0	65.5
Random-16	80.1	57.7	65.0
Top-16	80.2	57.9	65.3

(c) Effect of retrieved size

Fusion	IoU_{75}	5°2cm	5°5cm
Add	80.4	58.5	65.3
Gate	**81.0**	**59.9**	**67.7**
Concat	80.5	58.0	65.0

(d) Effect of memory fusion mechanism.

Table 5: Ablation studies on memory mechanisms. Settings used in our final model are colored in gray.

performance. This suggests that local features encode more explicit geometric information, which guides the model to focus on structural similarity rather than temporal proximity.

Effect of Memory Construction Strategies. As discussed in the previous section, the warm-up strategy refers to newly features are first fed into buffer until the buffer reaches full capacity, after which the update mechanism is activated. Tab. 4b reports the comparison of different memory construction strategies. The results show that the warm-up strategy can bring the best performance compared to the direct update strategy and random initialization. Consequently, the warm-up strategy provides a more reliable foundation for subsequent memory updates and retrieval.

Effect of Memory Update Mechanism. We then analyze the impact of the memory update stage. First, as shown in 5a, performing memory updates before retrieval yields better performance than post-retrieval updates. This confirms that updating the memory in advance allows the stored representations to reflect the most informative and up-to-date geometric context, leading to more effective retrieval. Second, we evaluate different memory update mechanisms, as reported in Tab. 5b. Compared to the FIFO strategy, the similarity-based token merge mechanism achieves superior performance on REAL275. To further prevent imbalanced updates like only a small subset of memory entries is repeatedly modified, we introduce lightweight balancing strategies. Specifically, the noise variant injects mild perturbations into cosine similarity computation, while the top-k variant randomly selects an update target from the top-k most similar entries. The noise-based token merge achieves the best overall results. These findings demonstrate that similarity-guided and balanced memory updates are crucial for maintaining a compact yet expressive memory buffer.

Effect of Retrieved Memory Size. We are also interested in how many memory entries of the buffer we need for retrieval. As shown in Tab. 5c, retrieving all memory entries yields the best performance, indicating that more memory

Buffer Length	IoU_{75}	5°2cm	5°5cm	10°2cm
0 (w/o mem)	79.5	57.0	64.6	75.1
16	79.6	57.2	64.9	77.1
48	80.0	58.5	66.5	78.3
96	**81.0**	**59.9**	67.7	**79.0**
256	80.8	59.1	**67.9**	78.6

Table 6: Effect of different buffer length. Setting 96 as the buffer length provides a balance between accuracy and efficiency.

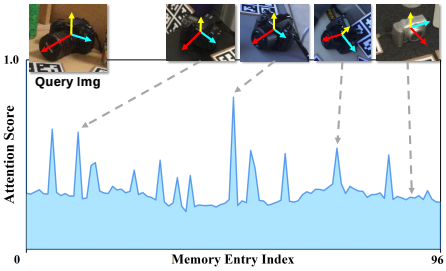

Fig. 4: Visualization of Memory Mechanism. The figure visualizes the retrieved memory elements and their attention weights during inference.

entries allow the model to access richer geometric context. Although attention-guided selection (Top-16) outperforms random sampling with the same retrieval size, the performance gap with full retrieval remains.

Effect of Memory Fusion Mechanism. We present the ablation of different fusion mechanisms in Tab. 5d. From the results, we can observe that the proposed gating fusion achieves the best performance, outperforming the simple addition and direct concatenation.

Effect of Different Buffer Length. We investigate the impact of different buffer length and report the performance in Tab. 6. We adopt AG-Pose [23] as baseline, which does not incorporate memory module, meaning the length is 0. Generally, we find that length 96 is good enough to train MemPose. Using longer length only results in comparable results but with much more computational overhead, *e.g.* GPU memory and training time. Therefore, we finally choose to use 96 as the buffer length, which provides a balance between accuracy and efficiency.

Discussion: Whether the memory mechanism retrieves relevant elements? To provide a direct view of how the memory mechanism functions, we present a qualitative case study that visualizes the retrieved memory elements and their attention scores during inference, as shown in Fig. 4. The horizontal axis denotes the index of memory entries in the buffer, while the vertical axis represents the attention score assigned to each entry. During inference, the target object queries the memory buffer, and the attention scores reflect the relative importance of different memory entries for the current prediction. From the visualization, we observe that memory entries with high attention scores correspond to instances whose poses and sizes are highly similar to those of the query object. In contrast, memory entries with low attention scores often exhibit significantly different or even opposite poses. These observations suggest that the proposed memory mechanism effectively retrieves and leverages pose-relevant category-level geometric information, enabling the model to focus on the most informative instances.

ID	Method	Visual Enc.	Param.↓	IoU_{75} ↑	5°2cm↑	TT↓	IT↑
1	AG-Pose	DINOv2 (VIT-S/14)	223M	80.1	57.0	47.5	**35**
2	ours		225M	**81.0**	**59.9**	47.8	<u>33</u>
3	ours†		225M	<u>80.5</u>	<u>57.8</u>	47.8	<u>33</u>
4	AG-Pose	ResNet18	**220M**	77.6	56.2	**46.9**	**35**
5	ours		<u>222M</u>	80.3	57.4	<u>47.1</u>	<u>33</u>

Table 7: Detailed comparison results on REAL275. '†' represents replacement of memory module with MLPs of the same number of parameters. TT: Traning Time (min/epoch), IT: Inference Speed (frame/sec). Overall best results are in **bold** and the second best results are <u>underlined</u>. Default settings are colored in gray.

Detailed Comparison with AG-Pose. Following common practice in this domain, we report the standard accuracy metrics in Tab. 1. To provide a more transparent and comprehensive comparison beyond accuracy, Tab. 7 further summarizes key practical factors, including the visual encoder type, the total number of parameters, training time per epoch (TT), and inference throughput (IT, FPS). As confirmed in (#1) and (#2) of Tab. 7, the memory module introduce only negligible parameter overhead when using the same DINOv2 ViT-S/14 backbone (225M *vs*. 223M), while the running time remains nearly unchanged (33 *vs*. 35 in FPS). More importantly, this gain does not come at the cost of efficiency: our method achieves comparable training time (47.8 vs. 47.5 min/epoch) while maintaining essentially the same inference budget and even slightly higher throughput (33 vs. 35 FPS). Notably, the memory buffer stores non-parametric entries and therefore does not introduce additional learnable weights; the parameter change mainly comes from lightweight integration components, making the overhead minimal in practice.

To ensure a fair comparison, we also replace the memory module with MLPs that have the same number of parameters (#3). These results indicate that the gains stem from the proposed memory mechanism rather than from additional parameters. What's more, we include additional results with ResNet18. Specifically, our method still outperforms AG-Pose with ResNet18 setting (#4 *vs*. #5), further supporting the efficacy of our approach.

Memory in Inference. During training, the category-specific memory buffer is registered as a non-trainable buffer and updated online with incoming samples. Once the corresponding category buffer is full, memory update and retrieval are activated to provide additional geometric context for pose estimation. In principle, the same memory mechanism can also be applied during inference. As new objects are perceived, the memory buffer can be continuously updated, allowing the model to accumulate additional category-level geometric knowledge from the test stream. However, to ensure a fair comparison with previous methods, all main results reported in this paper are obtained under a frozen-memory inference protocol, where the memory buffer is fixed after training and is not updated on the test set. Therefore, our standard evaluation does not introduce test-set leak-

Method	IoU_{75}	$5°2cm$	$5°5cm$	$10°2cm$
AG-Pose [23]	80.1	57.0	64.6	75.1
MemPose (frozen memory)	81.0	59.9	67.7	79.0
MemPose (updating memory)	**81.2**	**63.6**	**68.3**	**80.9**

Table 8: Performance of updating memory buffer during inference on REAL275. Settings used in our final model are colored in gray.

age or test-time adaptation effects. Additional experiments and detailed analysis of inference-time memory updating are provided in Tab. 8.

5 Conclusion

In this work, we perform a comprehensive analysis of existing parametric COPE paradigms and introduce a core essential: the dynamic mechanism to accumulate and reuse category-level geometric memory, which is essential for robust COPE. This analysis reveals that relying solely on static shape priors or fixed model parameters limits the ability of current approaches to generalize across diverse object instances. Based on these insights, we propose MemPose, a simple yet effective memory-augmented framework that explicitly incorporates category-level geometric memory into the COPE pipeline. MemPose provides meaningful contextual support for pose estimation, enabling the model to leverage accumulated geometric knowledge from previous observations. Extensive experimental results on three challenging benchmarks demonstrate the effectiveness of the proposed method.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (No.62233013, 62333017, 62403358).

References

1. Agro, B., Casas, S., Wang, P., Gilles, T., Urtasun, R.: Mad: Memory-augmented detection of 3d objects. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 1449–1460 (2025)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision (ECCV). pp. 213–229 (2020)
3. Chen, C., Guan, M., Lin, X., Li, J., Lin, L., Wang, Q., Chen, X., Luo, J., Sun, C., Zhang, D., et al.: Telemem: Building long-term and multimodal memory for agentic ai. arXiv preprint arXiv:2601.06037 (2025)

4. Chen, K., Dou, Q.: Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR). pp. 2773–2782 (2021)
5. Chen, W., Jia, X., Chang, H.J., Duan, J., Shen, L., Leonardis, A.: Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1581–1590 (2021)
6. Chen, Y., Di, Y., Zhai, G., Manhardt, F., Zhang, C., Zhang, R., Tombari, F., Navab, N., Busam, B.: 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 (CVPR). pp. 9959–9969 (2024)
7. Di, Y., Zhang, R., Lou, Z., Manhardt, F., Ji, X., Navab, N., Tombari, F.: 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 (CVPR). pp. 6781–6791 (2022)
8. Dong, Q., Fu, Y.: Memflow: Optical flow estimation and prediction with memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19068–19078 (2024)
9. Dong, Q., Fu, Y.: Online dense point tracking with streaming memory. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
10. Fu, Y., Wang, X.: Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. Advances in Neural Information Processing Systems (NeurIPS) **35**, 27469–27483 (2022)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (CVPR). pp. 2961–2969 (2017)
12. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: International Conference on Learning Representations (ICLR) (2017)
13. Jung, H., Wu, S.C., Ruhkamp, P., Zhai, G., Schieber, H., Rizzoli, G., Wang, P., Zhao, H., Garattoni, L., Meier, S., 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 (CVPR). pp. 22498–22508 (2024)
14. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Lee, S., Kim, T.K.: Joint learning of pose regression and denoising diffusion with score scaling sampling for category-level 6d pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5757–5768 (2025)
16. Li, W., Xu, H., Huang, J., Jung, H., Yu, P.K., Navab, N., Busam, B.: Gce-pose: Global context enhancement for category-level object pose estimation. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 27154–27165 (2025)
17. Lin, H., Liu, Z., Cheang, C., Fu, Y., Guo, G., Xue, X.: Sar-net: Shape alignment and recovery network for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6707–6717 (2022)
18. Lin, J., Wei, Z., Ding, C., Jia, K.: Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks. In: European Conference on Computer Vision (ECCV). pp. 19–34. Springer (2022)

19. Lin, J., Wei, Z., Li, Z., Xu, S., Jia, K., Li, Y.: 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 (CVPR). pp. 3560–3569 (2021)
20. Lin, J., Wei, Z., Zhang, Y., Jia, K.: 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 (CVPR). pp. 14001–14011 (2023)
21. Lin, X., Peng, Y., Wang, L., Zhong, X., Zhu, M., Feng, Y., Yang, J., Liu, C., Chen, Q.: Cleanpose: Category-level object pose estimation via causal learning and knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5990–6000 (2025)
22. Lin, X., Wang, D., Zhou, G., Liu, C., Chen, Q.: Transpose: 6d object pose estimation with geometry-aware transformer. Neurocomputing **589**, 127652 (2024)
23. Lin, X., Yang, W., Gao, Y., Zhang, T.: 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 (CVPR). pp. 21040–21049 (2024)
24. Lin, X., Zhu, M., Dang, R., Zhou, G., Shu, S., Lin, F., Liu, C., Chen, Q.: Clipose: Category-level object pose estimation with pre-trained vision-language knowledge. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) (2024)
25. Liu, J., Sun, W., Liu, C., Yang, H., Zhang, X., Mian, A.: Mh6d: Multi-hypothesis consistency learning for category-level 6-d object pose estimation. IEEE Transactions on Neural Networks and Learning Systems (TNNLS) (2024)
26. Liu, J., Sun, W., Yang, H., Deng, P., Liu, C., Sebe, N., Rahmani, H., Mian, A.: Diff9d: Diffusion-based domain-generalized category-level 9-dof object pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2025)
27. Liu, T., Qi, X., Wang, L., Li, J., Lin, X., Zhu, M., Cui, Y., Liu, C., Chen, Q.: Stamp: Spatial-temporal anchored motion planning for zero-shot continuous vision-and-language navigation. Sensors **26**(12), 3698 (2026)
28. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research Journal (TMLR) pp. 1–31 (2024)
29. Peng, Y., Lin, X., Ma, N., Du, J., Liu, C., Liu, C., Chen, Q.: Sam-lad: Segment anything model meets zero-shot logic anomaly detection. Knowledge-Based Systems **314**, 113176 (2025)
30. Peng, Y., Lin, X., Ma, N., Liu, C., Chen, Q.: Vllm-lad: Visual large language model for zero-shot logical anomaly detection. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) (2026)
31. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In Advances in Neural Information Processing Systems (NeurIPS) **30** (2017)
32. Ren, H., Yang, W., Liu, X., Zhang, S., Zhang, T.: Learning shape-independent transformation via spherical representations for category-level object pose estimation. In: The Thirteenth International Conference on Learning Representations (ICLR) (2025)
33. Sheng, K., Wang, L., He, Z., Lin, X., Liu, C., Chen, Q.: Dream: Dynamic routing of experts via attention-based mixture for vision-language-action modeling. Knowledge-Based Systems p. 115585 (2026)

34. Shi, H., Xie, B., Liu, Y., Sun, L., Liu, F., Wang, T., Zhou, E., Fan, H., Zhang, X., Huang, G.: Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. In: International Conference on Learning Representations (ICLR) (2026)
35. Smith, L.N.: Cyclical learning rates for training neural networks. In: 2017 IEEE winter conference on applications of computer vision (WACV). pp. 464–472. IEEE (2017)
36. Tian, M., Ang, M.H., Lee, G.H.: Shape prior deformation for categorical 6d object pose and size estimation. In: European Conference on Computer Vision (ECCV). pp. 530–546. Springer (2020)
37. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **13**(04), 376–380 (1991)
38. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2642–2651 (2019)
39. Wang, H., Agapito, L.: 3d reconstruction with spatial memory. In: 2025 International Conference on 3D Vision (3DV). pp. 78–89. IEEE (2025)
40. Wang, Q., Chen, S., Shen, Y.: Causalvtg: Towards robust video temporal grounding via causal inference. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)
41. Xu, T.X., Guo, Y.C., Lai, Y.K., Zhang, S.H.: Mbptrack: Improving 3d point cloud tracking with memory networks and box priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9911–9920 (2023)
42. Yu, S., Zhai, D.H., Xia, Y.: Keypose: Category-level 6d object pose estimation with self-adaptive keypoints. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 39, pp. 9653–9661 (2025)
43. Zhang, R., Di, Y., Lou, Z., Manhardt, F., Tombari, F., Ji, X.: Rbp-pose: Residual bounding box projection for category-level pose estimation. In: European Conference on Computer Vision (ECCV). pp. 655–672. Springer (2022)
44. Zheng, L., Wang, C., Sun, Y., Dasgupta, E., Chen, H., Leonardis, A., Zhang, W., Chang, H.J.: Hs-pose: Hybrid scope feature extraction for category-level object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17163–17173 (2023)
45. Zhong, X., Shen, M., Lin, X., Liu, C., Chen, Q.: An adaptive path planning algorithm with singularity consistency for robotic arms. Robotica pp. 1–18 (2025)
46. Zhu, M., Lin, X., Dang, R., Liu, C., Chen, Q.: Fine-grained spatiotemporal motion alignment for contrastive video representation learning. In: Proceedings of the 31st ACM International Conference on Multimedia (ACMMM). pp. 4725–4736 (2023)
47. Zhu, M., Wang, Z., Hu, M., Dang, R., Lin, X., Zhou, X., Liu, C., Chen, Q.: Mote: Reconciling generalization with specialization for visual-language to video knowledge transfer. Advances in Neural Information Processing Systems (NeurIPS) **37**, 55403–55424 (2024)