Pre-defined Keypoints Promote Category-level Articulation Pose Estimation via Multi-Modal Alignment

Wenbo Xu[1] , **Li Zhang**[2,3] , **Liu Liu** [1*] , **Yan Zhong**[4] , **Haonan Jiang**[5] , **Xue Wang** [2] , **Rujing Wang** [2]
[1]Hefei University of Technology, China
[2]Hefei Institute of Physical Science, Chinese Academy of Sciences, China
[3]University of Science and Technology of China, China
[4]School of Mathematical Sciences, Peking University, China
[5]Zhejiang University of Technology, China
2023170714@mail.hfut.edu.cn, zanly20@mail.ustc.edu.cn, liuliu@hfut.edu.cn

Abstract

Articulations are essential in everyday interactions, yet traditional RGB-based pose estimation methods often struggle with issues such as lighting variations and shadows. To overcome these challenges, we propose a novel **P**re-defined keypoint based framework for category-level articulation pose estimation via multi-modal **A**li**G**nm**E**nt, coined **PAGE**. Specifically, we first propose a customized keypoint estimation method, aiming to avoid the divergent distance pattern between heuristically generated keypoints and visible points. In addition, to reduce the mutual information redundancy between point clouds and RGB images, we design the geometry-color alignment, which fuses the features after aligning two modalities. This is followed by decoding the radius for each visible point, and applying our proposal integration scoring strategy to predict keypoints. Ultimately, the framework outputs the per-part 6D pose of the articulation. We conduct extensive experiments to evaluate PAGE across a variety of datasets, from synthetic to real-world scenarios, demonstrating its robustness and superior performance.

1 Introduction

Articulated objects are common in daily life, encompassing household items like scissors and cabinets, as well as office tools such as laptops. These objects are composed of multiple rigid components linked by joints, setting them apart from standard rigid objects through their kinematic constraints. This unique structure complicates articulation pose estimation, making it more challenging than for rigid objects. Accurate and efficient pose estimation of articulated objects remains a crucial hurdle in various downstream applications, particularly in robot manipulation [Xiong *et al.*, 2023], human-object interactions [Liu *et al.*, 2021], embodied AI [Yu *et al.*, 2023; Guo *et al.*, 2024], and augmented reality [Amin and Govilkar, 2015].

*Corresponding author: Liu Liu.

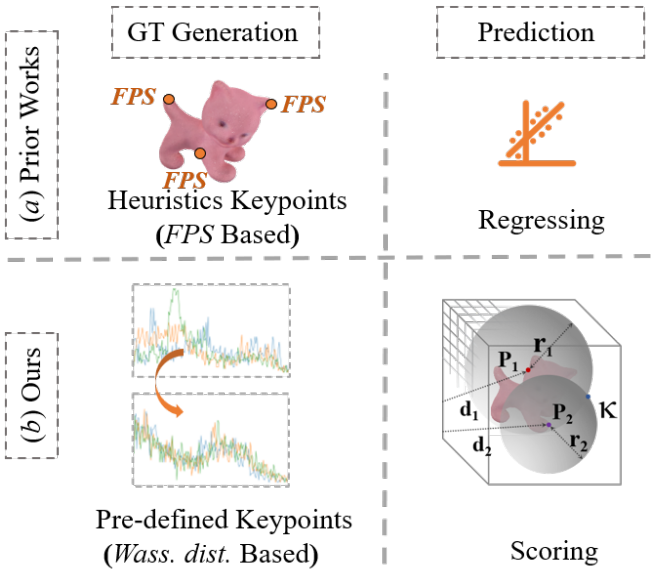

Figure 1: Motivation. 1) Traditional methods generate GT keypoints through heuristic approaches and then regress and align the keypoint locations in different spaces. 2) In this paper, we found that the pre-defined keypoint method outperforms heuristic algorithms. The core idea is that pre-defined keypoints make the radius differences, which are originally uneven across different sets of keypoint, more uniform, thereby reducing prediction bias.

Traditionally, the 6D object pose estimation problem has been addressed by matching points between 3D models and images. In recent years, the advent of low-cost RGB-D sensors has made it possible to infer poses for low-texture objects. In low-light environments, RGB-D methods [Wang *et al.*, 2019] have demonstrated higher accuracy compared to pure RGB methods. Furthermore, as an alternative to directly regressing keypoint coordinates, keypoint-based methods have proven to be highly effective for pose estimation [Li *et al.*, 2021], and can be seamlessly applied to multi-modal input for 6D object pose estimation. Despite progress, category-level articulated pose estimation remains a challenging and critical task, as existing methods still encounter several unresolved issues:

(i) Multi-Modal Fusion. 3D object models provide geometric information, while RGB images offer texture and color details. Traditional fusion methods (e.g., [Wang *et al.*, 2019]) integrate features from these modalities but often fail to address their distinct data distributions and feature representations. This oversight can result in information redundancy or mismatches, leading to learning difficulties or convergence failures. Although subsequent works, such as [Shivakumar *et al.*, 2019], attempt to address this issue using depth maps generated from projections, these maps lack the rich color and texture information of RGB images, as they only represent 2D plane projections. Similarly, RGB images lack depth information, capturing only 2D color distributions. Multi-modal fusion thus remains a significant challenge, further complicated by factors like viewpoint changes, background noise, and varying lighting conditions.

(ii) Keypoint Estimation Manner. Keypoint-based methods are widely used in domains like human and animal pose estimation. Recent advances include heatmap-based [Sun *et al.*, 2020] and sequence-based [Lin *et al.*, 2022] techniques, which, despite differing mechanisms, belong to the regression paradigm. However, these methods face similar challenges: 1) The distribution of distances between visible points and keypoints is overly scattered, resulting in nearby points having a stronger sensitivity to positional features, making them easier to regress. 2) Dependency on global features: Reliance on global features makes them vulnerable to noise and occlusion, leading to inaccurate keypoint predictions.

To address the first issue, we propose a geometry-guided, image-enhanced fusion module that aligns point cloud and image features. Point clouds provide geometric structure, while images offer rich texture. By projecting point clouds into depth maps to locate regions of interest in image features, we extract pose-sensitive features via interpolation and selection. A tailored fusion module then filters redundancy and noise, ensuring semantic consistency across modalities.

To address the second issue, we reformulate the task as proposal integration and scoring (see Fig. 1). The network regresses a radius per visible point, denoting its Euclidean distance to the keypoint, and maps it into a 3D accumulator space to generate robust, independent proposals per grid, even under occlusion. Additionally, pre-defined keypoints enhance performance by yielding more uniform radius distributions compared to heuristically generated ones.

In summary, to address the category-level articulated object pose estimation problem under multi-modal representation, we propose a **P**re-defined **K**eypoint multi-modal **A**li**G**nm**E**nt framework, named **PAGE**. The core of our method lies in a novel geometric-color alignment process, pre-defined keypoint estimation mechanism, and proposal integration scoring strategy. In practice, to validate the effectiveness of our method, we conduct extensive experiments on multiple datasets, including synthetic, semi-synthetic, and real-world datasets. We believe these experimental results demonstrate the superior performance and robustness of our PAGE. Our main contributions can be concluded as threefold:

- We propose the **PAGE**, a novel **P**re-defined Keypoint multi-modal **A**li**G**nm**E**nt framework for category-level articulation pose estimation.
- In PAGE, Color features are leveraged to complement geometric information, forming the enhanced geometric-color alignment. To avoid the weakness of traditional heuristics and regressing methods, we propose the customized Keypoint estimator and proposal integration scoring scheme.
- The experimental results from synthetic (ArtImage), semi-synthetic (ReArtMix), and real-world datasets (RobotArm) all demonstrate the effectiveness and robustness of the proposed PAGE.

2 Related Work

2.1 Pose Estimation from Different Modalities

Pose estimation methods can be categorized by input data type: 1) *RGB-Based*: These methods estimate 6D poses from color or monochrome images, typically via 2D detection or keypoint localization [Baek *et al.*, 2019]. 2) *Depth-Based*: Depth images are converted into partial point clouds using camera parameters. Approaches include reinforcement learning [Liu *et al.*, 2023] and regression models [Li *et al.*, 2020]. 3) *RGB-D-Based*: These methods fuse depth and RGB images through various levels and paradigms. For example, DenseFusion [Wang *et al.*, 2019] employs dense pixel-wise feature embedding.

In summary, RGB-only methods are sensitive to lighting, causing performance drops in low illumination, while depth-only methods struggle to differentiate objects with similar shapes but different colors, leading to ambiguities. To overcome these issues, we adopt an RGB-D manner. Instead of traditional ways, we emphasize feature alignment to reduce redundancy and improve downstream task performance.

2.2 Keypoint Based Pose Estimation

Keypoint-based methods are widely applied in various fields, such as human [McNally *et al.*, 2022], animal [Gong *et al.*, 2022], and object estimation [Xue *et al.*, 2021]. The general workflow of these methods typically begins with a keypoint regression network, where the core task is to estimate the keypoint locations of an object in different spaces. The object's pose is then inferred based on the correspondences of these keypoints. The key advantage of these methods lies in their heuristic algorithm (e.g., FPS, BBox) design, aiming to generate geometrically dispersed keypoints. To enhance robustness, this process often integrates classical algorithms, such as RANSAC [Besl and McKay, 1992; Horn *et al.*, 1988; Fischler and Bolles, 1981]. Keypoint-based approaches have been proven to be one of the most accurate solutions for 6D pose estimation, and recent research has further advanced the development of these methods. Generally speaking, it can be divided into the following paradigms: 1) Detection. i.e., detect the position of key points through images or sensor data (such as [Qiao *et al.*, 2017], [Nguyen *et al.*, 2022]). 2) Matching. i.e., match key points between different images or frames for tracking or alignment (e.g., [DeTone *et al.*, 2018], [Rockwell *et al.*, 2024]). 3) Generation. i.e., Generate the location or heatmap of key points ([Zhu and Ye, 2024]).

Although previous methods vary in their implementation, most can be regarded as regression methods. These meth-

ods rely on clear and well-defined input-output mappings and are heavily dependent on high-quality data and robust global features. Furthermore, this paradigm is highly sensitive to outliers, which can lead to biased network convergence and a lack of robustness in noisy and occluded scenarios.

3 Problem Statement

In this paper, our goal is to address the **C**ategory-level **A**rticulation **P**ose Estimation task based on RGB-D modalities (**CAPE** task). The core idea is to predict the pre-defined keypoints through a scoring mechanism to determine the articulated pose. To this end, we define a new paradigm for the CAPE task, named **PAGE**. Specifically speaking, given the partial point cloud $\mathcal{P}$ of an articulated object $A = \{\delta_k\}_{k=1}^{K}$ (where $\{\delta_k\}$ represents the k-th rigid part) and its corresponding 2D observed image I as the **input**, the **output** objectives of our PAGE are as follows: (i) per-part 3D rotation $R^{(k)} \in SO(3)$; (ii) per-part 3D translation $\mathbf{t}^{(k)}$. Together, the rotation and translation parameters form the 6D pose estimation result, denoted as $T = \{R^{(k)}, \mathbf{t}^{(k)}\}_{k=1}^{K} \in SE(3)$.

The proposed PAGE framework operates as follows: point clouds within the intra-category are processed by a backbone optimized using geometry-sensitive loss functions to generate pre-defined keypoints GT. Then the partial point cloud $\mathcal{P} \in \mathbb{R}^{N\times 3} = \{p_i\}_{i=1}^{N}$ is projected into a depth image using the camera mapping parameter M, establishing a 1-to-1 mapping with the 2D image I. This filters redundant information, yielding point-wise color features $\mathcal{F}_C$, which are processed by the GC-Fuser to produce re-modulated features $\mathcal{F}_{CG}$. These features are used to either decode the articulation segmentation mask (optional) or directly predict the radius r_i for each *visible point* p_i. Per-part keypoints $\mathcal{K}^{(k)} = \{\kappa_i\}_{i=1}^{3}$ are predicted as follows: if the error between the predicted radius and the corresponding grid's distance is within a threshold η, the corresponding proposal count is incremented by 1. Then each grid's score S_i is computed, and the highest-scoring grid's center is selected as the keypoint. Finally, the ICP algorithm outputs the estimated 6D pose $\{R^{(k)}, \mathbf{t}^{(k)}\}_{k=1}^{K}$.

4 Methodology

4.1 Pre-defined Keypoint Estimation

As highlighted in Sec. 2, the core of traditional keypoint-based methods lies in their heuristic algorithms, which aim to generate geometrically scattered keypoints, typically placed near the object's surface. These keypoints exhibit a nonlinear distribution, meaning they aren't arranged along a single line but are scattered in multiple directions, aiding in the subsequent pose transformation recovery through correspondences between keypoints and image points. However, we observe that the distance distribution between heuristic keypoints and visible points follows a distinct divergent pattern, a phenomenon that has not been adequately addressed in previous studies. Therefore, this work shifts its focus to the generation of pre-defined keypoints. We find that by training a Graph Network, a set of geometrically sensitive scattered keypoints can be selected, which significantly enhances the accuracy and efficiency of object pose estimation. Compared to previous heuristic keypoint algorithms, this method provides a more precise prediction of articulated object poses.

As shown in Fig. 2, our goal is to train the network such that the radius (i.e., the Euclidean distance between visible points and keypoints) distribution of each keypoint set is as similar as possible. To achieve this, we draw inspiration from the design of [Qiu *et al.*, 2020] and employ GraphNet. Note that the input module consists of the complete point clouds of the same category, while the output is a keypoint distribution that is spatially almost consistent. This approach leads to a pre-defined keypoint generator that is specific to the current category. High-quality 3D keypoints should possess geometric sensitivity in order to more accurately capture the pose of articulated objects. Therefore, during the GT keypoint estimation process, we consider the following key factors: 1) **Minimum Distribution Difference**: Heuristically generated keypoints often exhibit a large variance in the distance distribution to visible points. This causes the model to develop a preference for utilizing point cloud regions near the keypoints (where features are denser) and avoiding regions farther away (where features are sparser), leading to local optimality or significant prediction errors. 2) **Divergent Layout**: Keypoints should avoid clustering in close or identical locations. Instead, they should be distributed as evenly as possible to more comprehensively represent the geometric features of the object. 3) **Geometric Inclusiveness**: The distribution of keypoints should not be too far removed from the object, as this would lead to a lack of shape meaning and may degrade the object representation to mere points.

Thus, we propose the **KP-Estimator**, which conducts customized design and constraints as follows: 1) We aim for the generated keypoints to have similar radius clusters (i.e., the Euclidean distance clusters between each keypoint and each visible point), meaning that the radius distribution for each pair of keypoints should be as close as possible. To achieve this, we introduce the Wasserstein loss $\mathcal{L}_{wass}$ [Wu *et al.*, 2024], which goes beyond WassGAN-GP [Gulrajani *et al.*, 2017] and in the same spirit inspired by involving both a critic loss and gradient penalty. The Wasserstein loss $\mathcal{L}_{wass}$ between two keypoint radius sets R^i and R^j is defined as:

$$\mathcal{L}_{wass} = \frac{1}{K}\sum_{k=1}^{K}\Big\{\sum_{\kappa_i}^{\mathcal{K}^{(k)}}\sum_{\kappa_j}^{\mathcal{K}^{(k)}}\big[D(R_i) - D(R_j) + \lambda\left(\|\nabla_R D(R)\| - 1\right)^2\big]\Big\} \tag{1}$$

Where $D(R_i)$ and $D(R_j)$ represent the histogram distributions of radius, R_i and R_j represent the radius group of the respective keypoints κ_i and κ_j. $D(R)$ denotes the joint distribution of R_i and R_j, while ∇_R is the gradient calculated from R, and λ is the gradient penalty hyperparameter. Unlike WassGAN-GP [Gulrajani *et al.*, 2017], which applies the gradient penalty only when the generator learns to imitate the GT, we apply the gradient penalty to all sets of radius values. This helps to enhance the smoothness and stability of the predictions, thereby optimizing overall performance.

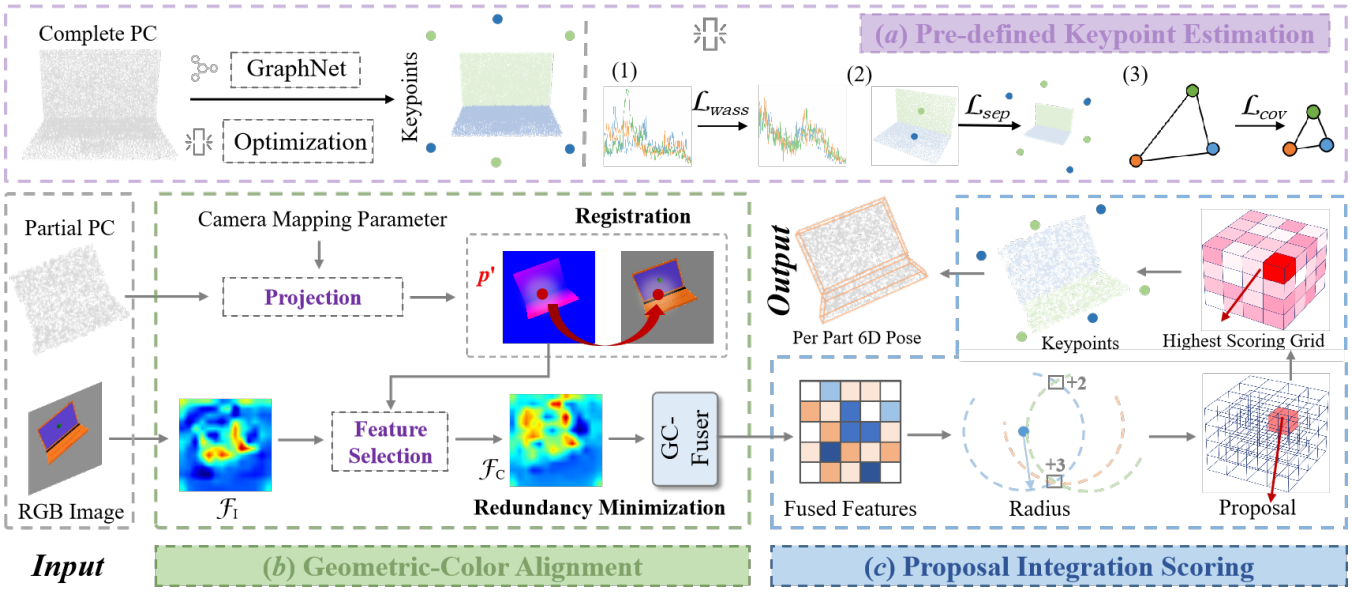

Figure 2: The Pipeline of the Proposed PAGE. It consists of the following components: (a) Pre-defined Keypoint Estimation. We use the geometric-sensitive loss to get the pre-defined keypoint GT. (Sec. 4.1) (b) Geometric-Color Alignment. This module takes partial PC and RGB image as input, and outputs the fused feature after aligning. (Sec. 4.2) (c) Proposal Integration Scoring. The predicted radii will be used for generating proposals and scoring for keypoints. (Sec. 4.3).

2) To prevent keypoints collapse, we introduce the per-part separation loss, $\mathcal{L}_{sep}$, which ensures that multiple keypoints do not converge at the same location.

$$\mathcal{L}_{sep} = \frac{1}{K}\sum_{k=1}^{K}\Big\{\sum_{\kappa_i}^{\mathcal{K}^{(k)}}\sum_{\kappa_j}^{\mathcal{K}^{(k)}}\exp(\zeta^2 - \|\kappa_i - \kappa_j\|_2)\Big\} \quad (2)$$

where ζ is the threshold. Note that ζ is slightly larger than the distance between keypoints sampled by FPS, because our key points are not distributed on the surface of the object, but have a certain distance.

3) Finally, to ensure comprehensive coverage of the rigid part's shape geometry by the learned keypoints, we introduce the per-part coverage loss, $\mathcal{L}_{cov}$. This loss is based on the difference between the volume $\mathbf{V}(\cdot)$ of the keypoints, $P^{(k)}$, and the input point cloud, $X^{(k)}$, while also penalizing keypoints that are distant from $X^{(k)}$:

$$\mathcal{L}_{cov} = \frac{1}{K}\sum_{k=1}^{K}(\|\mathbf{V}(P^{(k)}) - \mathbf{V}(X^{(k)})\|_2 + \sum_{\kappa_i}^{\mathcal{K}}\|\kappa_i - X^{(k)}\|_2) \quad (3)$$

The total loss, $\mathcal{L}_{ttl}$, for KP-Estimator, which estimates self-supervised per-part 3D keypoints, is defined as the weighted sum of the Wasserstein loss $\mathcal{L}_{wass}$, separation loss $\mathcal{L}_{sep}$, and coverage loss $\mathcal{L}_{cov}$, with corresponding weights $\lambda_1 = 0.6$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.2$, as follows:

$$\mathcal{L}_{ttl} = \lambda_1\mathcal{L}_{wass} + \lambda_2\mathcal{L}_{sep} + \lambda_3\mathcal{L}_{cov} \quad (4)$$

Through this approach, KP-Estimator is trained to identify a set of dispersed keypoints with uniformly distributed radii. This results in a unified set of keypoints applicable to all objects in the dataset, which can subsequently be utilized in keypoint-based 6D pose estimation methods. Experimental results show that the keypoints selected by KP-Estimator improve the accuracy of most evaluation metrics (details can be found in the ablation study).

4.2 Geometric-Color Alignment

Feature alignment and fusion is not a trivial task, as many existing methods directly perform fusion while neglecting the importance of proper alignment. To address this, we propose a novel multi-modal alignment and fusion method, called GCA (**G**eometric-**C**olor **A**lignment). Specifically, the proposed GCA consists of three steps: (1) Depth-RGB Image Registration, (2) Redundancy Minimization (i.e., feature selection), and (3) Geometry-Color Fusion (GC-Fuser).

(1) Depth-RGB Image Registration. Firstly, the partial PC will be projected into a 2D depth image with the help of the camera mapping matrix (denoted as M). Afterward, we conduct the registration by bilinear interpolation, and then the 1-1 correspondence between depth and RGB images can be easily established at different resolutions. Mathematically, for a specific point p in the point cloud, we can calculate its corresponding position $p'(u, v)$ in the depth image. This process can be expressed by the projection equation $p' = M \times p$.

(2) Redundancy Minimization. With the 1-1 registration between Depth and RGB images, our target is to extract RGB features strongly correlated with each visible point. Formally, taking the sampled position p' and the image feature map $\mathcal{F}_I$ as inputs, generating point-wise color feature representations $\mathcal{F}_C$ for each sampled position. Thus, we filter out the redundant mutual information. In practice, since the sampling position may lie between adjacent pixels, we employ interpolation to obtain the color features at continuous coordinates.

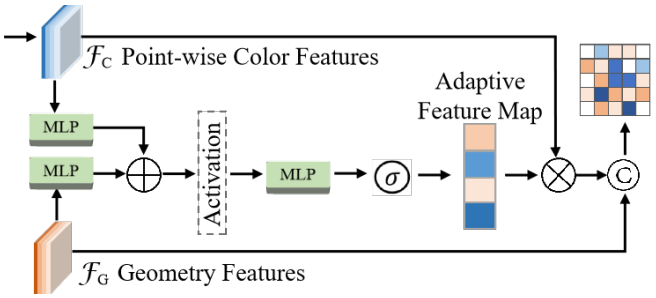

Figure 3: Illustration of the GC-Fuser. The geometry features are applied to guide the fusion procedure with point-wise color features.

(3) Geometry-guided Color Fuser. To address the aforementioned fusion issue, we propose the **GC-Fuser** (**G**eometry-guided **C**olor **Fuser**) that adaptively estimates the importance of color features on a per-point basis, guided by geometry features. The specific process is outlined in Fig. 3: first, geometry features $\mathcal{F}_G$ and per-point color features $\mathcal{F}_C$ are projected into modality-aware feature spaces independently. Then, we further conduct mutual information minimization between them to explicitly promote the learning of complementary information. Next, these two branches are fused into a compact feature representation, which is then compressed through another fully connected layer, ultimately generating an adaptive weight map $\mathbf{w}$. We normalize this weight map to the range [0,1] using the softmax function. This procedure can be formulated as:

$$\mathbf{w} = \mathcal{S}(\mathcal{W} \cdot \boldsymbol{H}(\mathrm{Add}(\mathcal{U}F_P, \mathcal{V}F_I))) \tag{5}$$

where $\mathcal{W}$, $\mathcal{U}$, $\mathcal{V}$ denote the learnable weight matrices in our GC-Fuser. $\mathcal{S}()$ represents Softmax and $\boldsymbol{H}()$ represents the output activation.

After obtaining the adaptive weight map $\mathbf{w}$, we combine the geometry features $\mathcal{F}_G \in \mathbb{R}^{N\times C_1}$ and semantic-related color features $\mathbf{w}\mathcal{F}_C \in \mathbb{R}^{N\times C_2}$ in a concatenation manner. The specific process can be expressed as:

$$\mathcal{F}_{CG} = \mathop{||}_{i=1}^{C} (\mathbf{w}\mathcal{F}_C, \mathcal{F}_G) \tag{6}$$

where $||$ represents concatenation, $C = C_1 + C_2$.

4.3 Proposal Integration Scoring

The distance voting method [Wu *et al.*, 2022] has been proven to be an effective position inference algorithm. By adhering to the inclusion similarity between the geometric structure of visible points and the complete point cloud structure, it leverages local structures to infer global pose information. Building on this, we further propose a novel 3D proposal integration scoring method. The core idea is: in the 3D accumulator space, the number of valid proposals from visible points is counted for each grid, and the confidence score is calculated by aggregating these proposals.

For each rigid part, we output a set of keypoint proposals along with their corresponding aggregation scores independently. To more intuitively illustrate the proposal integration scoring mechanism, we draw an analogy between each visible point in the point cloud and a Member of Parliament (MP), whose primary task is to raise proposals for each grid in the 3D accumulator space. In this procedure, considering the partial point cloud $\mathcal{P} \in \mathbb{R}^{N\times 3}$ (input), we utilize the fused features $\mathcal{F}_{CG}$ to output $3N$ channels, which represents the predicted distances (radii) from each visible point to the three pre-defined keypoints. Furthermore, we define the resolution of the 3D accumulator space as ρ, which corresponds to the edge length of the smallest unit grid voxel. If the absolute error between the computed radial distance and the estimated keypoint radius is less than or equal to ρ, the proposal is considered valid, and the proposal count for the corresponding grid is incremented by 1; otherwise, the proposal is deemed invalid, with the count remaining at 0. We adopted the confidence function from [Tekin *et al.*, 2018] and applied it to our proposal function, as expressed mathematically below:

$$P_i = \mathbb{1}\{(1 - \frac{|d_i - \hat{r}_i|}{\sigma}) \geq 0\} \tag{7}$$

Where $\mathbb{1}(\cdot)$ is the indicator function.

Subsequently, the proposal count for each grid is aggregated to compute the confidence score. Here, the confidence score is defined as the ratio of the current grid's proposal count to the total proposal count. The grid with the highest score is selected as the final result (Eq. 8), and its center point is considered as the elected keypoint $\hat{\mathcal{K}}$ (Eq. 9).

$$S_i = \frac{P_i}{P_{total}} \times 100\% \tag{8}$$

$$\hat{\mathcal{K}} = \mathrm{argmax}_i\{S_i\} \tag{9}$$

Overall, when the network is sufficiently trained, the majority of the proposals generated by local feature points are able to accurately localize the target keypoint, and the final center point estimation converges to the ground truth. Even in the presence of a small amount of noisy votes, no significant deviation occurs, which demonstrates the robustness of our method. From another perspective, our aggregation step effectively estimates the probability density of the center point position in the spatial domain (i.e., the keypoint) and determines its location accordingly. Notably, this process is invariant to rigid transformations (i.e., SE(3) invariance), ensuring stability even when the object pose changes, and maintaining a prominent peak in the distribution.

5 Experiments

Datasets, Baselines, and Metrics. We evaluate our PAGE framework on the ArtImage [Xue *et al.*, 2021], ReArtMix [Liu *et al.*, 2022], and RobotArm [Liu *et al.*, 2022] datasets, covering synthetic, semi-synthetic and real-world scenarios. For performance comparison, we benchmark against four RGB+Depth-based methods: A-NCSH [Li *et al.*, 2020], Densefusion [Wang *et al.*, 2019], ASMM [Zhang *et al.*, 2025]. The evaluation metrics include degree error for 3D rotation, distance error for 3D translation, and 3D Intersection over Union (IoU) for measuring 3D scale.

Implementation Details. During the pre-processing, the input RGB images are scaled into 224×224 resolution and the

Category	Method	Per-part Pose		
		rotation error (°) ↓	translation error (m) ↓	3D IoU (%) ↑
Laptop	A-NCSH [Li *et al.*, 2020]	5.3, 5.4	0.054, 0.043	56.7,40.2
	Densefusion [Wang *et al.*, 2019]	5.4, 4.3	0.062, 0.061	43.5,24.1
	ASMM [Zhang *et al.*, 2025]	4.9, 4.4	0.048, 0.042	58.1,43.5
	PAGE (Ours)	**4.0**, **1.7**	**0.014**, **0.019**	**64.6**,**50.4**
Eyeglasses	A-NCSH [Li *et al.*, 2020]	3.5, 18.3, 18.2	0.043, 0.286, 0.283	52.5,40.2,39.6
	Densefusion [Wang *et al.*, 2019]	4.9, 7.5, 7.5	0.062, 0.103, 0.104	46.8,41.5,38.4
	ASMM [Zhang *et al.*, 2025]	3.5, 6.1, 6.4	0.041, 0.235, 0.236	51.2,43.1,41.5
	PAGE (Ours)	**3.2**, **5.2**, **5.1**	**0.027**, **0.075**, **0.071**	**58.6**,**46.5**,**51.7**
Dishwasher	A-NCSH [Li *et al.*, 2020]	4.0, 4.8	0.059, 0.123	84.3,56.2
	Densefusion [Wang *et al.*, 2019]	6.0, 6.2	0.104, 0.142	66.5,38.9
	ASMM [Zhang *et al.*, 2025]	12.5, 4.6	0.146, 0.184	43.8,28.6
	PAGE (Ours)	**3.9**,**4.3**	**0.055**,**0.079**	**89.3**,**67.6**
Scissors	A-NCSH [Li *et al.*, 2020]	2.0, **2.6**	0.035, **0.021**	45.8,44.8
	Densefusion [Wang *et al.*, 2019]	3.9, 3.4	0.048, 0.039	35.6,34.5
	ASMM [Zhang *et al.*, 2025]	3.6, 4.7	0.047, 0.060	38.4,29.0
	PAGE (Ours)	**1.9**, 5.4	**0.013**, 0.032	**47.9**, **48.6**
Drawer	A-NCSH [Li *et al.*, 2020]	2.8, 3.3, 3.5, 2.7	0.041, 0.145, 0.137, 0.072	90.5,82.1,79.4,83.7
	Densefusion [Wang *et al.*, 2019]	4.4, 4.4, 4.4, 4.4	0.111, 0.143, 0.144, 0.115	75.8,73.4,70.2,71.3
	ASMM [Zhang *et al.*, 2025]	3.2, 3.6, 3.5, 3.8	0.124, 0.178, 0.175, 0.121	80.3,74.2,75.7,74.4
	PAGE (Ours)	**2.8**, **2.8**, **2.8**, **2.8**	**0.010**, **0.017**, **0.015**, **0.013**	**91.8**,**87.6**,**85.0**,**86.2**

Table 1: Comparison with State-of-the-arts on ArtImage Dataset. The categories laptop, eyeglasses, dishwasher and scissors contain only free joint and revolute joints, and the drawer category contains free joint and prismatic joints.

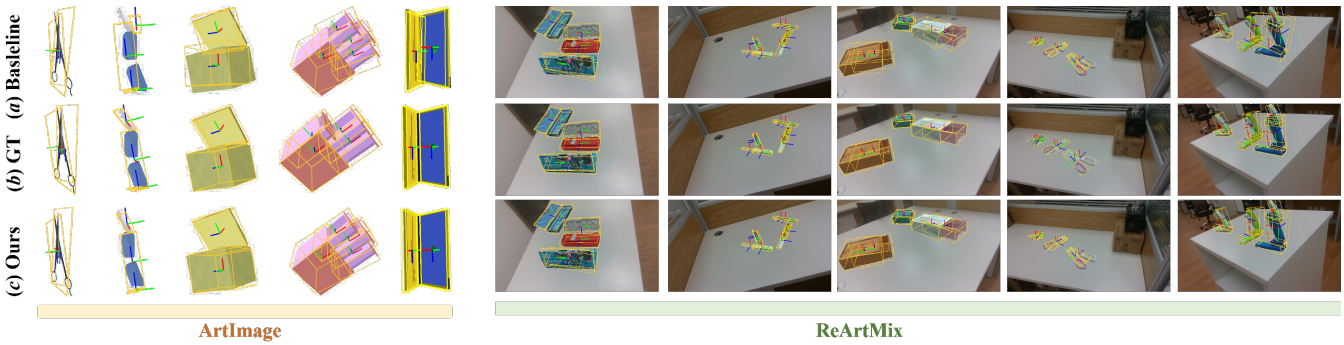

Figure 4: Qualitative Results on ArtImage and ReArtMix. The left is a synthetic dataset and the right is the semi-synthetic scenario.

point clouds are downsampled into 2,048 points before being fed into the network. Please note that the input modality of all methods have been adjusted to depth + color, and re-training was conducted under the same experimental settings compared to ours. When training the backbone (PointNet++ [Qi *et al.*, 2017]), the Adam optimizer was employed with an initial learning rate of 0.001 and a weight decay of 0.0001. The learning rate decayed by a factor of 0.5 every 20 epochs. All experiments were conducted on four NVIDIA GeForce RTX 4090 GPUs, each with 24GB of memory.

5.1 Comparison with the SOTA Methods

In this section, we conduct comparative experiments in the synthetic dataset (i.e., ArtImage) with the classical methods, aiming to verify the effectiveness of our PAGE. Tab. 1 demonstrates the quantitative results. 1) When considered as a whole, our method has comprehensively refreshed the SOTA results, which proves the effectiveness of the coordination of each proposed module. 2) When considered separately, we get the best pose estimation result lies in category *laptop*, with $(\mathbf{4.0}^\circ, \mathbf{1.7}^\circ)$, $(\mathbf{0.014}m, \mathbf{0.019}m)$ for rotation error and translation error. This can be attributed to our keypoint generation strategy, which can better generalize to objects with uniform geometric scales. Considering the 3D IoU metric, our prediction errors are significantly better at each part compared to the baseline model (A-NCSH), with the average value **57.5%** vs. **48.5%**. Qualitative results in Fig. 4 (left) further highlight that our predictions closely align with GT, confirming the robustness and precision of our approach.

5.2 Ablation Study

Pre-defined Keypoints. As mentioned in Sec. 4.1, we propose that methods of pre-defined keypoints can outperform the heuristic methods. Therefore, we conduct detailed ablation experiments in this section. Results are shown in Tab. 2 (I - III). It is noted that the FPS and BBox based keypoint generation methods are adopted from [Zhao *et al.*, 2020]. From the quantitative results, we can conclude that: our method significantly outperforms the heuristic methods (i.e., FPS and

Configuration	Keypoints Numbers	Per-part Pose	
		rotation error (°) ↓	translation error (m) ↓
I	FPS	11.2, 12.6	0.086, 0.112
II	BBox	9.8, 10.5	0.042, 0.085
III (Ours)	Pre-defined	1.9, 5.4	0.013, 0.032
Configuration	Fusion	Per-part Pose	
IV	Concat	2.6, 7.1	0.025, 0.056
V	Dense	2.3, 6.5	0.020, 0.045
VI (Ours)	-	1.9, 5.4	0.013, 0.032

Table 2: Ablation Study. It is noted that experiments are conducted on the category *Scissors*.

Category	Method	Per-part Pose	
		rotation error (°) ↓	translation error (m) ↓
Box	A-NCSH	**4.1**, 3.5	0.023, 0.034
	PAGE	4.4, **1.3**	**0.015**, **0.017**
Stapler	A-NCSH	5.1, 6.4	0.034, 0.041
	PAGE	**2.6**, **3.1**	**0.027**, **0.031**
Cutter	A-NCSH	3.1, 3.4	0.017, 0.021
	PAGE	**1.4**, **1.4**	**0.015**, **0.016**
Scissors	A-NCSH	5.7, 5.4	**0.013**, 0.015
	PAGE	**1.6**, **1.3**	0.018, **0.012**
Drawer	A-NCSH	3.4, 3.6	0.022, 0.021
	PAGE	**1.2**, **1.2**	**0.013**, **0.017**

Table 3: Pose Estimating Results on ReArtMix Dataset.

BBox based methods) by a large margin, which reported (1.9° and 5.4°) and (0.013m, 0.032m) with our methods. All in all, the proposed method not only addresses the distribution of keypoints but also takes into account the geometric and structural relationships of the articulation, optimizing through three geometry-sensitive loss functions (see Sec. 4.1). This is crucial for modeling the pose of objects based on keypoints.

Fusion Manner. In this work, we propose a novel multi-modal learning method coined geometric-color alignment as detailed in Sec. 4.2. To verify the effectiveness of the proposed module, we conduct the ablation experiments in Tab. 2 (IV - VI). Note that configuration IV represents the direct concatenation of depth and RGB features. Configuration V represents the method from densefusion [Wang *et al.*, 2019]. It can be inferred that our method achieves state-of-the-art performance, while the traditional method (i.e., direct concatenation) suffers from severe performance degradation due to noisy feature alignment.

5.3 Generalization Capacity

Experiments on Semi-Synthetic Scenarios. ReArtMix dataset is used to evaluate our method, which incorporates semi-synthetic scenarios. The detailed results are presented in Tab. 3. Our approach achieves the best performance on the *Drawer* category, with rotation error of only **1.2**°, and translation error of **0.013**m and **0.017**m. This demonstrates that our method is equally effective in Semi-Synthetic Scenarios, showcasing its robustness and adaptability across diverse object categories and scenes. Qualitative results for the five categories are illustrated in Fig. 4 (Right).

Experiments on Real-world Scenarios. We further train and evaluate PAGE on the 7-part RobotArm dataset in real-world scenarios. As shown in the quantitative results (Tab. 4), our method achieves superior performance in per-part pose estimation compared to the baseline A-NCSH. Specifically, PAGE significantly reduces both rotation and translation errors. For rotation errors, PAGE achieves notable improvements across parts 1 to 7, with average rotation errors reduced to **7.5**° compared to **12.1**° of A-NCSH. Similarly, for translation errors, PAGE maintains an average error of **0.044m**, outperforming the baseline's **0.126m**. While accumulative errors are observed in deeper kinematic structures (5th, 6th, and 7th parts), our method demonstrates greater robustness against these challenges compared to the baseline, which exhibits more significant accumulative errors in both rotation and translation. Qualitative results are illustrated in Fig. 5.

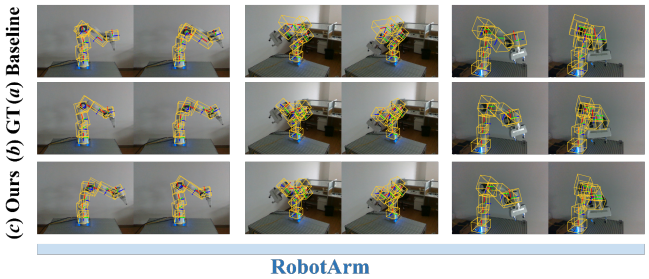

Figure 5: Qualitative Results on 7-part RobotArm dataset.

Per-part Rotation Error (°)							
Part ID	1	2	3	4	5	6	7
A-NCSH	7.6	7.8	10.1	10.3	10.8	15.7	22.3
PAGE	**3.7**	**3.4**	**3.5**	**4.6**	**5.5**	**12.1**	**19.8**
Per-part Translation Error (m)							
Part ID	1	2	3	4	5	6	7
A-NCSH	0.011	0.042	0.066	0.060	0.075	0.232	0.399
PAGE	**0.003**	**0.006**	**0.011**	**0.013**	**0.018**	**0.089**	**0.167**

Table 4: Quantitative Results on RobotArm Dataset.

6 Conclusion

In this work, we propose a novel framework, PAGE, to address category-level articulation pose estimation through multi-modal alignment. Our approach focuses on sequentially aligning geometry and color features, followed by effective feature fusion. To overcome the limitations of heuristic keypoints, we introduce a tailored pre-defined keypoint estimation method that enhances pose estimation performance. Additionally, a proposal integration scoring strategy is employed to accurately determine the 6D pose of articulated objects. Experimental results demonstrate that PAGE achieves state-of-the-art performance on the synthetic ArtImage dataset and exhibits strong generalization capabilities on the semi-synthetic ReArtMix dataset and real-world multi-hinged articulated object datasets (e.g., RobotArm). These results underscore the robustness and adaptability of our method across diverse scenarios.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grant 62302143 and Anhui Provincial Natural Science Foundation under Grant 2308085QF207

References

[Amin and Govilkar, 2015] Dhiraj Amin and Sharvari Govilkar. Comparative study of augmented reality sdks. *International Journal on Computational Science & Applications*, 5(1):11–26, 2015.

[Baek *et al.*, 2019] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1067–1076, 2019.

[Besl and McKay, 1992] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992.

[DeTone *et al.*, 2018] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018.

[Fischler and Bolles, 1981] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[Gong *et al.*, 2022] Caili Gong, Yong Zhang, Yongfeng Wei, Xinyu Du, Lide Su, and Zhi Weng. Multicow pose estimation based on keypoint extraction. *PloS one*, 17(6):e0269259, 2022.

[Gulrajani *et al.*, 2017] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[Guo *et al.*, 2024] Dan Guo, Kun Li, Bin Hu, Yan Zhang, and Meng Wang. Benchmarking micro-action recognition: Dataset, methods, and applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6238–6252, 2024.

[Horn *et al.*, 1988] Berthold KP Horn, Hugh M Hilden, and Shahriar Negahdaripour. Closed-form solution of absolute orientation using orthonormal matrices. *JOSA A*, 5(7):1127–1135, 1988.

[Li *et al.*, 2020] Xiaolong Li, He Wang, Li Yi, Leonidas J Guibas, A Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3706–3715, 2020.

[Li *et al.*, 2021] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 11313–11322, 2021.

[Lin *et al.*, 2022] Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A Vela, and Stan Birchfield. Keypoint-based category-level object pose tracking from an rgb sequence with uncertainty estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 1258–1264. IEEE, 2022.

[Liu *et al.*, 2021] Yiming Liu, Chunki Yiu, Huiling Jia, Tszhung Wong, Kuanming Yao, Ya Huang, Jingkun Zhou, Xingcan Huang, Ling Zhao, Dengfeng Li, et al. Thin, soft, garment-integrated triboelectric nanogenerators for energy harvesting and human machine interfaces. *EcoMat*, 3(4):e12123, 2021.

[Liu *et al.*, 2022] Liu Liu, Han Xue, Wenqiang Xu, Haoyuan Fu, and Cewu Lu. Toward real-world category-level articulation pose estimation. *IEEE Transactions on Image Processing*, 31:1072–1083, 2022.

[Liu *et al.*, 2023] Liu Liu, Jianming Du, Hao Wu, Xun Yang, Zhenguang Liu, Richang Hong, and Meng Wang. Category-level articulated object 9d pose estimation via reinforcement learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 728–736, 2023.

[McNally *et al.*, 2022] William McNally, Kanav Vats, Alexander Wong, and John McPhee. Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation. In *European Conference on Computer Vision*, pages 37–54. Springer, 2022.

[Nguyen *et al.*, 2022] Hung-Cuong Nguyen, Thi-Hao Nguyen, Jakub Nowak, Aleksander Byrski, Agnieszka Siwocha, and Van-Hung Le. Combined yolov5 and hrnet for high accuracy 2d keypoint and human pose estimation. *Journal of Artificial Intelligence and Soft Computing Research*, 12(4):281–298, 2022.

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

[Qiao *et al.*, 2017] Sen Qiao, Yilin Wang, and Jian Li. Real-time human gesture grading based on openpose. In *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–6. IEEE, 2017.

[Qiu *et al.*, 2020] Zhongwei Qiu, Kai Qiu, Jianlong Fu, and Dongmei Fu. Dgcn: Dynamic graph convolutional network for efficient multi-person pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11924–11931, 2020.

[Rockwell *et al.*, 2024] Chris Rockwell, Nilesh Kulkarni, Linyi Jin, Jeong Joon Park, Justin Johnson, and David F Fouhey. Far: Flexible accurate and robust 6dof relative camera pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19854–19864, 2024.

[Shivakumar *et al.*, 2019] Shreyas S Shivakumar, Ty Nguyen, Ian D Miller, Steven W Chen, Vijay

Kumar, and Camillo J Taylor. Dfusenet: Deep fusion of rgb and sparse depth information for image guided dense depth completion. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 13–20. IEEE, 2019.

[Sun *et al.*, 2020] Ke Sun, Zigang Geng, Depu Meng, Bin Xiao, Dong Liu, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation by ranking heatmap-guided adaptive keypoint estimates. *arXiv preprint arXiv:2006.15480*, 2020.

[Tekin *et al.*, 2018] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 292–301, 2018.

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

[Wu *et al.*, 2022] Yangzheng Wu, Mohsen Zand, Ali Etemad, and Michael Greenspan. Vote from the center: 6 dof pose estimation in rgb-d images by radial keypoint voting. In *European Conference on Computer Vision*, pages 335–352. Springer, 2022.

[Wu *et al.*, 2024] Ruihai Wu, Kai Cheng, Yan Zhao, Chuanruo Ning, Guanqi Zhan, and Hao Dong. Learning environment-aware affordance for 3d articulated object manipulation under occlusions. *Advances in Neural Information Processing Systems*, 36, 2024.

[Xiong *et al.*, 2023] Haoyu Xiong, Haoyuan Fu, Jieyi Zhang, Chen Bao, Qiang Zhang, Yongxi Huang, Wenqiang Xu, Animesh Garg, and Cewu Lu. Robotube: Learning household manipulation from human videos with simulated twin environments. In *Conference on Robot Learning*, pages 1–10. PMLR, 2023.

[Xue *et al.*, 2021] Han Xue, Liu Liu, Wenqiang Xu, Haoyuan Fu, and Cewu Lu. Omad: Object model with articulated deformations for pose estimation and retrieval. *arXiv preprint arXiv:2112.07334*, 2021.

[Yu *et al.*, 2023] Qiaojun Yu, Junbo Wang, Wenhai Liu, Ce Hao, Liu Liu, Lin Shao, Weiming Wang, and Cewu Lu. Gamma: Generalizable articulation modeling and manipulation for articulated objects. *arXiv preprint arXiv:2309.16264*, 2023.

[Zhang *et al.*, 2025] Hong-Bo Zhang, Jia-Xin Hong, Jing-Hua Liu, Qing Lei, and Ji-Xiang Du. Images, normal maps and point clouds fusion decoder for 6d pose estimation. *Information Fusion*, page 102907, 2025.

[Zhao *et al.*, 2020] Wanqing Zhao, Shaobo Zhang, Ziyu Guan, Wei Zhao, Jinye Peng, and Jianping Fan. Learning deep network for detecting 3d object keypoints and 6d poses. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 14134–14142, 2020.

[Zhu and Ye, 2024] Xicheng Zhu and Xinchen Ye. Ganbodypose: Real-time 3d human body pose data key point detection and quality assessment assisted by generative adversarial network. *Image and Vision Computing*, page 105144, 2024.