Pre-Defined Keypoints Worth It: Multi-Modal Learning for Category-Level Articulated Objects Pose Estimation

Li Zhang, Wenbo Xu, Liu Liu, Haonan Jiang, Rujing Wang, Xue Wang, and Yan Zhong

***Abstract*—Articulated objects play a vital role in daily interactions, but traditional RGB-based pose estimation methods often face challenges such as lighting variations and shadows. To address these limitations, we introduce PAGE, a novel Pre-defined keypoint-based framework for category-level articulation pose estimation via multi-modal AliGnmEnt. Our approach is motivated by the observation that the distance distribution between heuristically generated keypoints and visible points exhibits a divergent pattern, a phenomenon previously overlooked. To tackle this, we propose a customized unsupervised keypoint estimation method that enhances the stability and robustness of model predictions. Furthermore, to minimize mutual information redundancy between point clouds and RGB images, we design a geometry-color alignment module that fuses features after aligning the two modalities. This is followed by decoding the radius for each visible point and applying our proposal integration scoring strategy to predict keypoints. The framework ultimately outputs the per-part 6D pose of the articulated object. We conduct extensive experiments across diverse datasets, ranging from synthetic to real-world scenarios, demonstrating the robustness and superior performance of PAGE. This work holds significant promise for applications in robotics, embodied intelligence, and augmented reality. Codes and datasets are available at the website: https://sites.google.com/view/pageforart**

***Note to Practitioners*—This work was motivated by the problem of articulation pose estimation. Articulated objects, such as cabinets, laptops, and scissors, are ubiquitous in both household and industrial settings. Traditional RGB-based methods for pose estimation often struggle with challenges such as lighting variations, occlusions, and the complexity of multi-part objects. This paper introduces PAGE (Pre-defined Keypoint multi-modal AliGnmEnt framework), a novel approach designed to enhance the robustness and accuracy of category-level articulation pose estimation. Specifically, 1) Pre-defined Keypoints: Unlike heuristic methods that generate keypoints based on empirical rules, PAGE employs pre-defined keypoints that are uniformly distributed and geometrically sensitive. 2) Multi-Modal Fusion: PAGE integrates RGB and depth data through a geometric-color alignment module. This fusion minimizes redundancy and ensures that the model leverages both geometric and texture information effectively. 3) Proposal Integration Scoring: The framework uses a scoring mechanism to aggregate proposals from visible points, enhancing the model's robustness against occlusions and noise. Extensive experiments validate the superiority of PAGE over existing SOTA methods. Our approach outperforms others not only on synthetic datasets but also demonstrates generalization capabilities for unseen object instances. We believe that accurately estimating the pose of articulations has the potential to be applied in many fields including robotics, embodied intelligence, and augmented reality.**

***Index Terms*—Articulation, pose estimation, multi-modal.**

I. Introduction

ARTICULATED objects, which include everyday items such as scissors and cabinets, as well as office equipment like laptops, are prevalent in daily life. These objects are characterized by their assembly of multiple rigid parts connected by joints, distinguishing them from conventional rigid objects due to their kinematic constraints. This distinctive configuration adds complexity to the estimation of articulation poses, presenting greater challenges compared to rigid objects. The precise and swift estimation of these poses is a significant obstacle in numerous applications, especially in the realms of robotic manipulation [1], [2], [3], interactions between humans and objects [4], [5], [6], 3D scene understanding [7], [8], [9], [10], and augmented reality [5], [11], [12].

Traditionally, the issue of 6D object pose estimation has been approached by aligning corresponding points between 3D models and images. In recent years, the proliferation of affordable RGB-D sensors has enabled the inference of poses for objects with limited texture. In environments with dim lighting, RGB-D techniques [13] have demonstrated greater accuracy compared to methods relying solely on RGB. Moreover, as an alternative to directly predicting keypoint coordinates, keypoint-based approaches have proven to be highly effective for pose estimation [14], [15] and can be readily adapted to multi-modal inputs for 6D object pose estimation.

Despite advancements, category-level articulated pose estimation remains a challenging and critical task, as existing methods still face several unresolved issues:

Received 7 February 2025; revised 22 June 2025, 6 October 2025, and 7 January 2026; accepted 31 March 2026. Date of publication 9 April 2026; date of current version 22 April 2026. This article was recommended for publication by Associate Editor P. Zheng and Editor D. Song upon evaluation of the reviewers' comments. This work was supported in part by the National Natural Science Foundation of China under Grant 62302143. *(Corresponding author: Liu Liu.)*

Li Zhang is with Hefei University of Technology, Hefei 230026, China, and also with the University of Science and Technology of China, Hefei 230026, China (e-mail: zanly20@mail.ustc.edu.cn).

Wenbo Xu and Liu Liu are with Hefei University of Technology, Hefei 230026, China (e-mail: 2023170714@mail.hfut.edu.cn; liuliu@hfut.edu.cn).

Haonan Jiang is with Zhejiang University of Technology, Hangzhou 310000, China (e-mail: 211122030127@zjut.edu.cn).

Rujing Wang and Xue Wang are with the Institute of Intelligent Machines and Hefei Institute of Physical Science, Chinese Academy of Sciences, Hefei 230031, China (e-mail: rjwang@iim.ac.cn; xwang@iim.ac.cn).

Yan Zhong is with Peking University, Beijing 100000, China (e-mail: zhongyan@stu.pku.edu.cn).

Digital Object Identifier 10.1109/TASE.2026.3682387

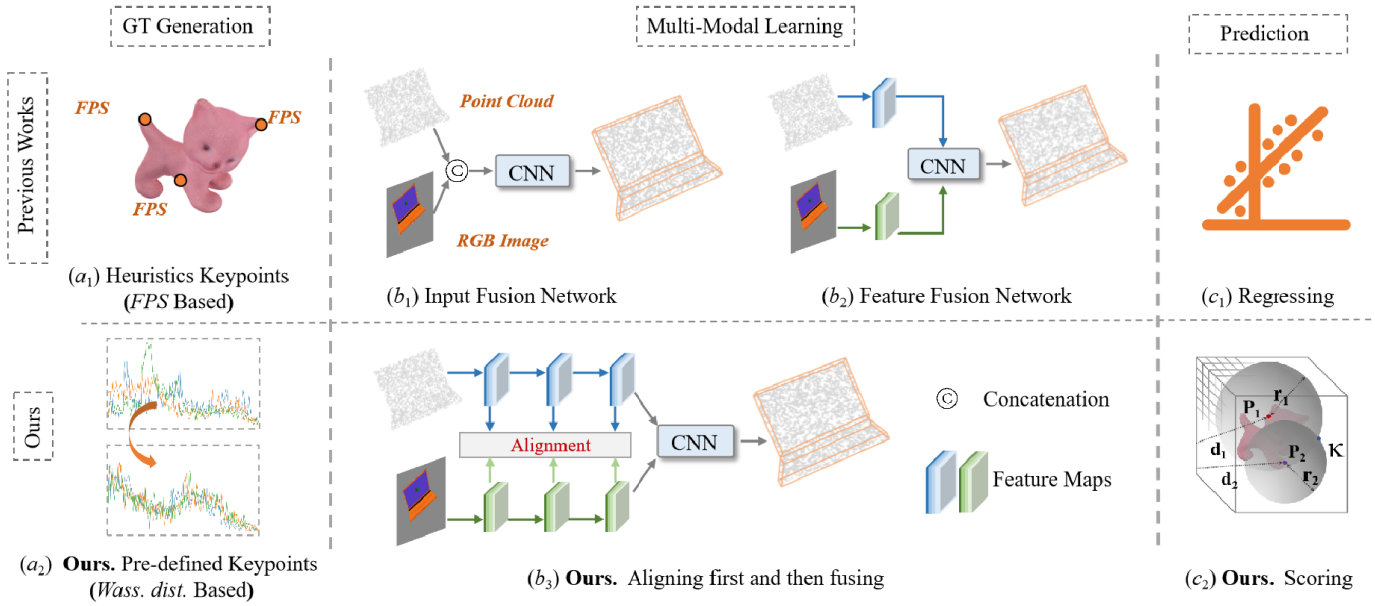

Fig. 1. **Novelty** of this work. **a) GT Generation.** Traditional methods generate keypoint ground truth (GT) heuristically and then predict and align the keypoint locations in different spaces. This approach relies on heuristic rules, which may cause the generated keypoint positions to be constrained by the accuracy of the empirical rules. **b) Multi-Modal Learning.** The categorization of existing multi-modal learning methods. Traditional methods tend to direct fusion while ignoring the importance of feature alignment. **c) Prediction.** In this paper, we found that the pre-defined keypoint method outperforms heuristic algorithms. The fundamental principle is that pre-defined keypoints reduce the radius differences, which are initially uneven across different groups, making them more uniform and thus reducing prediction bias. Based on this, we employ multiple scoring-based predictions to accurately estimate the keypoint locations.

(I) Keypoint Estimation Approach. Techniques centered on keypoints are extensively utilized in fields such as the estimation of human and animal postures. Innovations in this area encompass methodologies based on heatmaps [16] and sequences [17], which, while differing in their operational frameworks, fall under the regression-based paradigm (Fig. 1-c_1)). Nonetheless, these approaches encounter analogous difficulties: 1) The intra-cluster variance of Euclidean distances between heuristic keypoints and their corresponding visible point sets (as opposed to occluded or invisible points) is typically large (Fig. 1-a_1)), reflecting the non-uniform spatial distribution of the point cloud. This uneven distribution leads to a biased feature utilization between keypoint-adjacent regions (feature-dense) and more distant areas (feature-sparse), which can result in suboptimal local minima or significant pose estimation errors. 2) Susceptibility to global features: Their dependence on global features renders them prone to disturbances and obstructions, culminating in imprecise keypoint estimations.

(II) Multi-Modal Fusion. In general, 3D object models provide geometric information, while RGB images capture texture and color details. Traditional methods fuse features from these two modalities using various strategies (e.g., [13], [18]); however, they often fail to account for the differences in data distribution and feature representation, leading to information redundancy or mismatches. This, in turn, can cause learning difficulties or hinder convergence. Specifically, existing fusion techniques largely follow traditional approaches, which can be classified into the following categories: early fusion (data-level, i.e., model input, Fig. 1-b_1) and deep fusion (feature-level, Fig. 1-b_2), which overlooks the specificity of geometric and color information in the feature space. Besides, RGB images lack depth information and merely depict color distributions on a 2D plane. As a result, multi-modal fusion remains a significant challenge, influenced by factors such as viewpoint variations, background noise, and lighting conditions.

To address the first issue, our approach fundamentally reformulates pose estimation as a proposal integration and scoring process. Specifically, the network regresses a geometric distance value for each visible point, representing its spatial distance to the corresponding mesh. These distances are then mapped into an accumulator space to generate independent proposals for each mesh. The independent accumulation of proposals endows the method with strong robustness in complex scenarios involving occlusions. Furthermore, we observe that replacing heuristic keypoints with pre-defined keypoints (**obtained via an unsupervised algorithm that integrates geometric and semantic constraints**) significantly improves model performance. The distance distribution between pre-defined keypoints and visible points is more uniform (Refer to Fig. 1-a_2), reducing the network's learning difficulty and accelerating convergence, thereby enhancing overall accuracy and stability. This strategy substantially strengthens the model's generalization capability, particularly under high occlusion and complex object surface conditions.

To resolve the second issue, our approach can be succinctly summarized as a geometry-guided image-enhanced fusion module (Refer to Fig. 1-b_3). Point clouds and images serve as mutual complements in depicting an object's structure and

surface details: point clouds delineate the geometric form and spatial arrangement, whereas images furnish intricate texture and visual cues. Expanding on this premise, we transform the point cloud into a depth map, utilizing its spatial data to pinpoint areas of significance within the image features. We then employ interpolation and selection techniques to distill features that are sensitive to pose, thereby harmonizing the features of point clouds and RGB images within the semantic domain. A customized fusion module is subsequently employed to sift through extraneous data and disturbances, maintaining a harmonious relationship between point cloud and image features throughout the fusion process.

In conclusion, to tackle the challenge of category-level articulated object pose estimation within a multi-modal representation context, we introduce a **P**re-defined **K**eypoint multi-modal **A**li**G**nm**E**nt framework, designated as **PAGE**. The essence of our approach is encapsulated in an innovative geometric-color alignment procedure, a pre-defined keypoint estimation mechanism, and a strategy for integrating and scoring proposals. In practical terms, to verify the efficacy of our method, we have undertaken comprehensive experiments across a variety of datasets, encompassing synthetic, semi-synthetic, and real-world data. We are confident that the results of these experiments highlight the exceptional performance and resilience of our PAGE framework. Our principal contributions are concluded in the following threefold:

- We have developed **PAGE**, an innovative **P**re-defined Keypoint multi-modal **A**li**G**nm**E**nt framework aimed at category-level articulation pose estimation.
- Within PAGE, color features are utilized to augment geometric data, thereby creating an enhanced geometric-color alignment. To avoid the limitations inherent in traditional heuristic and regression methods, we have devised a customized keypoint estimator and a proposal integration scoring methodology.
- The extensive results derived from synthetic (ArtImage), semi-synthetic (ReArtMix), and real-world datasets (RobotArm) collectively confirm the efficacy and robustness of the proposed PAGE framework.

II. Related Work

A. Multi-Modal Representation Learning

Feature fusion remains challenging due to sensor noise, inefficient information utilization, and cross-modal misalignment, which collectively limit performance. Although fusion is commonly categorized into early, deep, and late paradigms, most existing methods symmetrically fuse point clouds and RGB images, overlooking the distinct roles of geometric and appearance cues and thereby oversimplifying multimodal integration. Moreover, many earlier approaches rely on handcrafted features or surrogate objectives [19], [20], [21], [22] rather than directly optimizing for 6D pose estimation.

To address these limitations, we exploit the complementary strengths of RGB and depth modalities through a parallel-branch fusion framework. Inspired by DenseFusion [13], our method establishes cross-modal feature correspondences based on feature similarity, spatial proximity, and geometric consistency, and progressively refines them by suppressing redundant signals while enhancing complementary cues across network layers. A dedicated fusion module then aggregates the refined features for pose prediction. Extensive experiments demonstrate that our local feature fusion strategy consistently outperforms DenseFusion and representative fusion baselines.

B. Pose Estimation From Different Modalities

The geometric properties (e.g., shape and contours) and appearance cues (e.g., lighting and shading) of 3D objects inherently encode pose information. Accordingly, existing 6D pose estimation methods can be broadly categorized by input modality into three groups. 1) RGB-based methods infer poses from color images via 2D detection or keypoint localization [23], [24], [25]. While effective in controlled settings, they are sensitive to illumination changes due to reliance on appearance cues; recent works improve robustness by modeling spatio-temporal dependencies and keypoint uncertainty [26], [27]. 2) Depth-based methods convert depth maps into partial point clouds and exploit geometric structure for pose estimation. Approaches based on reinforcement learning [28], voting [29], [30], and regression [31] exhibit strong robustness to visual disturbances, while geometry-regularized designs such as ArtPoint [32] further enhance rotational stability. 3) RGB-D methods jointly leverage appearance and geometry to improve accuracy, with DenseFusion [13] performing pixel-wise fusion and UDA-COPE [33] mitigating the synthetic-to-real gap via domain adaptation. Beyond rigid objects, deformation-aware approaches (e.g., ARAPReg [34]) enforce geometric consistency and offer insights for articulated object modeling.

In essence, methods relying solely on RGB are prone to variations in lighting, which can degrade performance under poor illumination, whereas depth-only methods may fail to distinguish between objects of similar shapes but differing colors, resulting in confusion. To address these limitations, we have chosen an RGB-D approach. Rather than employing conventional direct fusion, ***we focus on feature alignment to minimize redundancy and enhance the performance of subsequent tasks.***

C. Keypoint Based Pose Estimation

Keypoint-based approaches are widely used in human [35], animal [36], and object pose estimation [37], as well as in voting-based paradigms such as 2D voting [38], 3D bounding-box voting [39], and circle voting [40]. These methods typically follow a two-stage pipeline, where a network first predicts keypoint locations and then estimates object pose via geometric correspondences among the predicted keypoints. Their effectiveness largely stems from heuristic design choices (e.g., Farthest Point Sampling and bounding box-based selection) that promote geometric coverage, often combined with robust estimators such as RANSAC [41], [42], [43] during pose recovery. Keypoint-based strategies remain among the most accurate solutions for 6D pose estimation and continue to be actively refined. Existing methods can be broadly categorized into three classes: (1) Detection, which localizes keypoints from images or sensor data [44], [45]; (2) Matching,

which establishes correspondences across views or time steps [46], [47]; and (3) Generation, which directly predicts keypoint coordinates or heatmaps [48].

While earlier methodologies differ in their execution tactics, the majority can be classified as ***regression techniques****. These techniques depend on explicit and precise input-output correlations and are significantly reliant on superior data quality and strong global features. Moreover, this approach is particularly vulnerable to anomalies, which may result in skewed network convergence and diminished robustness in environments with noise and obstructions.*

III. Preliminary

This section formally defines the mathematical representations of 3D rotations and rigid transformations essential for articulated object pose estimation. We adopt the standard Lie group formulations prevalent in robotics and computer vision literature [49]. SO (3) defines the space of 3D rotations:

$$\mathrm{SO}(3) = \left\{ R \in \mathbb{R}^{3\times 3} \mid R^{\top} R = \mathbf{I}, \det(R) = 1 \right\}, \tag{1}$$

where the orthogonal constraint preserves vector norms and angles, while $\det(R) = 1$ ensures right-handed coordinate systems. SE (3) extends this to full rigid transformations:

$$\mathrm{SE}(3) = \left\{ T = \begin{bmatrix} R & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} \mid R \in \mathrm{SO}(3), \mathbf{t} \in \mathbb{R}^3 \right\}, \tag{2}$$

with $\mathbf{t}$ denoting translation. This homogeneous representation enables efficient composition of transformations. For articulated objects comprising K rigid parts, the complete pose is $T = \{T^{(k)}\}_{k=1}^{K}$ where each $T^{(k)} \in \mathrm{SE}(3)$ defines a 6-degree-of-freedom (6-DoF) transformation. Category-level pose estimation aims to predict T without instance-specific models by leveraging these group-theoretic foundations.

IV. 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 articulation 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 points of the k-th rigid part.) and its corresponding 2D observed RGB 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 pipeline of the proposed PAGE framework can be formulated as follows: First, all point clouds of the same category are input into a graph convolutional network (GCN) and optimized using our designed geometry-sensitivity loss function to automatically generate the ground truth (GT) for pre-defined keypoints. It is noteworthy that these keypoint annotations are not manually labeled but are obtained through unsupervised learning by the proposed KP estimator. By incorporating constraints such as geometric consistency, spatial dispersion, and coverage, the method ensures the rationality and representativeness of the keypoints. Then, the partial point cloud $\mathcal{P} = \{p_i\}_{i=1}^{N} \in \mathbb{R}^{N\times 3}$ will be projected into a depth image based on the camera mapping parameters M and establish the 1-1 mapping relationship with the 2D image I. Based on this, the redundant information is filtered out and we obtain point-wise color features $\mathcal{F}_C$, which will be fed into the GC-Fuser and get the re-modulated feature $\mathcal{F}_{CG}$. This feature can be used either to decode the segmentation mask of the articulation (optional) or directly to output the radius r_i for each *visible point* p_i. Finally, we perform per-part keypoints $\mathcal{K}^{(k)} = \{\kappa_i^{(k)}\}_{i=1}^{3}$ prediction: if the error between the predicted radius and the corresponding grid's distance falls within the threshold η, the corresponding proposal is incremented by 1. Then each grid's score S_i is computed, and the center of grid with the highest score is identified as the keypoint. We use the ICP algorithm to output the estimated 6D pose $\{R^{(k)}, \mathbf{t}^{(k)}\}_{k=1}^{K}$. **To avoid ambiguity and better clarify the variables, we use the symbol * to indicate the GT variables, and ˆ to indicate the predicted variables.**

V. Methodology

A. Pre-Defined Keypoint Estimation

As emphasized in Sec. II, the key idea of conventional keypoint-based approaches lies in their heuristic algorithms, designed to produce keypoints that are geometrically dispersed, often located close to the object's surface. These keypoints display a nonlinear distribution, indicating they are not aligned on a single line but are instead spread out in various directions, facilitating the subsequent recovery of pose transformation through the correspondences between keypoints and image points. Nonetheless, we note that the distribution of distances between heuristic keypoints and visible points follows a unique divergent pattern, a phenomenon that prior research has not sufficiently explored. Drawing from this insight, our study reorients its attention towards the generation of pre-defined keypoints. We discover that by training a Graph Network, a set of geometrically sensitive scattered keypoints can be identified, which markedly improves the precision and efficiency of object pose estimation. In contrast to earlier heuristic keypoint algorithms, this technique offers a more accurate forecast of articulated object poses (Refer to the ablation study in Sec. VI).

As depicted in Fig. 2, our objective is to train the network to ensure that the radius (i.e., the Euclidean distance between visible points and keypoints) distribution for each keypoint set is as uniform as possible. To accomplish this, we take inspiration from the architecture of work [50] and utilize GraphNet. It is important to note that the input module is composed of the complete point clouds from the same category, while the output is a keypoint distribution that is nearly consistent spatially. This methodology results in a pre-defined keypoint generator tailored to the current category. High-quality 3D keypoints should exhibit geometric sensitivity to more precisely capture the pose of articulated objects. Consequently, during the GT keypoint estimation process, we take into account the following critical factors: 1) **Minimum**

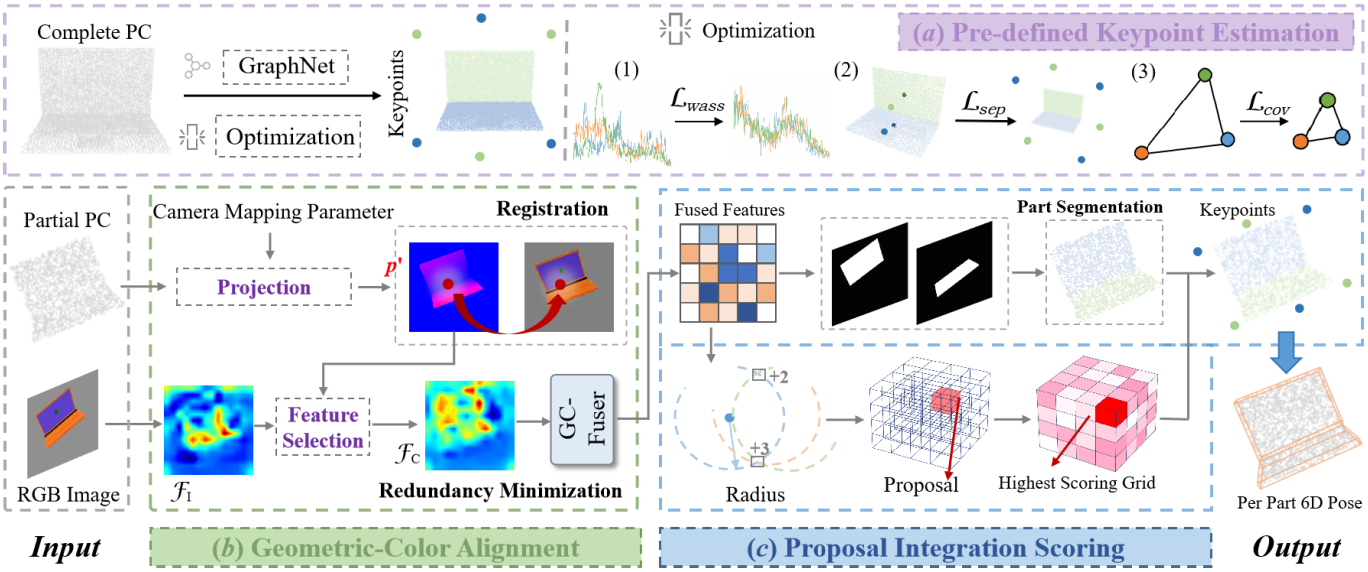

Fig. 2. **The pipeline of the proposed PAGE.** It consists of the following components: **(a) Pre-defined Keypoint Estimation.** We use the geometric-sensitive loss to get the pre-defined keypoint GT. **(b) Geometric-Color Alignment.** This module takes partial PC and RGB image as input, and outputs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

Distribution Difference: Keypoints generated heuristically often show a significant variance in the distance distribution to visible points. This can cause the model to develop a preference for utilizing point cloud regions near the keypoints (where features are more concentrated) and to avoid more distant regions (where features are more dispersed), resulting in local optimality or substantial prediction inaccuracies. 2) **Divergent Layout**: Keypoints should be arranged to prevent clustering in close or identical locations. Instead, they should be as evenly distributed as possible to more thoroughly represent the geometric features of the object. 3) **Geometric Inclusiveness**: The distribution of keypoints should not be excessively distant from the object, as this could result in a loss of shape significance and potentially reduce the object representation to few points.

Consequently, we introduce the **KP-Estimator**, which implements tailored design and constraints as follows: 1) Our objective is for the generated keypoints to exhibit similar radius clusters (i.e., the Euclidean distance clusters between each keypoint and each visible point), implying that the radius distribution for each pair of keypoints should be as similar as possible. To realize this, we incorporate the Wasserstein loss $\mathcal{L}_{wass}$, which extends beyond WassGAN-GP [51] and is motivated by the inclusion of both a critic loss and a gradient penalty. The Wasserstein loss $\mathcal{L}_{wass}$ between two keypoint radius sets R^i and R^j is defined as:

$$\mathcal{L}_{wass} = \frac{1}{K}\sum_{k=1}^{K}\Bigg\{\sum_{\kappa_i}^{\mathcal{K}^{(k)}}\sum_{\kappa_j}^{\mathcal{K}^{(k)}}\big[D_m(R_i) - D_m(R_j) + \lambda\left(\left\|\nabla_R D_j(R)\right\| - 1\right)^2\big]\Bigg\} \tag{3}$$

Here, $D_m(R_i)$ and $D_m(R_j)$ signify the normalized histogram results via bins of the radius, with R_i and R_j representing the radius groups corresponding to the keypoints κ_i and κ_j. $D_j(R)$ refers to the joint distribution histogram result of R_i and R_j (i.e., $D(R) = \mathrm{Cat}(D(R_i), D(R_j))$), while ∇_R is the gradient derived from R, and λ is the hyperparameter for the gradient penalty. In contrast to WassGAN-GP [51], which imposes the gradient penalty solely when the generator is learning to replicate the GT, we enforce the gradient penalty across all sets of radius values. This approach aids in augmenting the smoothness and stability of the predictions, thereby refining the overall performance.

2) To prevent keypoints from collapsing, we introduce the per-part separation loss, $\mathcal{L}_{sep}$, which guarantees that multiple keypoints remain distinct and do not merge.

$$\mathcal{L}_{sep} = \frac{1}{K}\sum_{k=1}^{K}\left\{\sum_{\kappa_i}^{\mathcal{K}^{(k)}}\sum_{\kappa_j}^{\mathcal{K}^{(k)}}\exp(\zeta^2 - \|\kappa_i - \kappa_j\|_2)\right\} \tag{4}$$

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.

Lastly, to guarantee that the learned keypoints thoroughly encompass the geometric shape of the rigid part, we implement the per-part coverage loss, $\mathcal{L}_{cov}$. This loss is derived from the discrepancy between the volume $\mathbf{V}(\cdot)$ of the keypoints, $P^{(k)}$, and the input point cloud, $X^{(k)}$, while also imposing a penalty on keypoints that are far removed from $X^{(k)}$:

$$\mathcal{L}_{cov} = \frac{1}{K}\sum_{k=1}^{K}\left(\|\mathbf{V}(P^{(k)}) - \mathbf{V}(X^{(k)})\|_2 + \sum_{\kappa_i}^{\mathcal{K}^{(k)}}\|\kappa_i - X^{(k)}\|_2\right) \tag{5}$$

The total loss, $\mathcal{L}_{ttl}$, for the KP-Estimator, which calculates self-supervised per-part 3D keypoints, is formulated as the weighted combination of the Wasserstein loss $\mathcal{L}_{wass}$, separation loss $\mathcal{L}_{sep}$, and coverage loss $\mathcal{L}_{cov}$, with respective weights $\lambda_1 = 0.6$ and $\lambda_2 = 0.2$, as delineated below:

$$\mathcal{L}_{ttl} = \lambda_1\mathcal{L}_{wass} + \lambda_2\mathcal{L}_{sep} + (1 - \lambda_1 - \lambda_2)\mathcal{L}_{cov} \tag{6}$$

By employing this methodology, the KP-Estimator is trained to identify a set of scattered keypoints characterized by uniformly allocated radii. This yields a consistent set of keypoints that can be applied to all objects within the dataset, which can then be employed in keypoint-based 6D pose estimation techniques. Experimental results demonstrate that the keypoints chosen by the KP-Estimator enhance the precision of the majority of evaluation metrics (specifics are available in the ablation study).

B. Geometric-Color Alignment

The alignment and fusion of multi-modal features remain a significant challenge, as many existing approaches tend to perform direct fusion without adequately addressing the importance of precise spatial correspondence and modality-specific characteristics. To overcome this limitation, we propose a novel multi-modal alignment and fusion module, termed **GCA** (**G**eometric-**C**olor **A**lignment). The GCA framework is designed to incorporate RGB information in a geometry-aware and asymmetrically guided manner, rather than relying on naive concatenation or symmetric attention mechanisms. Specifically, GCA comprises three sequential stages: (1) *Depth-to-RGB Image Registration*, which establishes pixel-wise correspondences between 3D points and their projected RGB counterparts; (2) *Redundancy Minimization*, which filters interpolated RGB features by assessing their geometric relevance to suppress noise and misalignment; and (3) *Geometry-Color Fusion (GC-Fuser)*, which adaptively integrates the retained RGB features into the geometric backbone using a learned gating mechanism. As illustrated in Fig. 2, we now describe each component in detail.

(1) Depth-RGB Image Registration. Initially, the partial point cloud is projected onto a 2D depth image utilizing the camera mapping matrix (represented as M). Subsequently, we perform the registration through bilinear interpolation, thereby facilitating a 1-1 correspondence between depth and RGB images across various resolutions. Mathematically, for a given point p within the point cloud, we can determine its corresponding position $p'(u, v)$ in the depth image. This operation can be formulated by the projection equation:

$$p' = M \times p. \tag{7}$$

where p is a 3D point in the point cloud, represented as $p(x, y, z)$, p' is the projected 2D point on the camera image plane, represented as $p'(x', y')$. M is a 3×4 mapping matrix that transforms the 3D point to the 2D image plane. The detailed steps are as follows:

i) To enable matrix multiplication, the 3D point $p(x, y, z)$ is converted into homogeneous coordinates, expressed as:

$$p = \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}. \tag{8}$$

Similarly, the 2D point $p'(u, v)$ is also represented in homogeneous coordinates:

$$p' = \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}. \tag{9}$$

ii) Mapping Matrix M The mapping matrix M is a 3×4 matrix, typically derived from the extrinsic and intrinsic parameters of the camera. It can be decomposed as:

$$M = K \times [R|t], \tag{10}$$

where K is the camera intrinsic matrix (size 3×3), which encodes the internal parameters of the camera, such as focal length and principal point. $[R|t]$ is the extrinsic matrix (size 3×4), where R is the rotation matrix and t is the translation vector, describing the relative pose.

iii) Substituting the matrices into the projection formula, we obtain:

$$p' = K \times [R|t] \times p. \tag{11}$$

Expanding this, we have:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \times \begin{bmatrix} R_{11} & R_{12} & R_{13} & t_1 \\ R_{21} & R_{22} & R_{23} & t_2 \\ R_{31} & R_{32} & R_{33} & t_3 \end{bmatrix} \times \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}.$$

In this way, we get $p' = (u, v)$ finally.

(2) Redundancy Minimization. Given the 1-1 registration between Depth and RGB images, our objective is to derive RGB features that are highly correlated with each visible point. Formally, by inputting the sampled position p' and the image feature map $\mathcal{F}_I$, we generate point-wise color feature representations $\mathcal{F}_C$ for each sampled position, thereby eliminating redundant mutual information. In practice, as the sampling position may fall between neighboring pixels, we utilize bilinear interpolation to acquire the color features at continuous coordinates. This procedure can be expressed as:

$$\mathcal{F}_C = \mathcal{B}(\mathcal{F}_{\mathbf{N}(p')}) \tag{12}$$

where $\mathcal{F}_C$ is the point-wise color feature for point p, $\mathcal{B}$ denotes the bilinear interpolation function, and $F_{(\mathcal{N}(p'))}$ represents the image features $\mathcal{F}_I$ of the neighboring pixels $\mathbf{N}(p')$ for the sampling position p'.

(3) Geometry-guided Color Fuser. To tackle the previously mentioned fusion challenge, we introduce the **GC-Fuser** (**G**eometry-guided **C**olor **Fuser**), which dynamically assesses the significance of color features for each individual visible point, under the guidance of geometry features. The detailed procedure is illustrated in Fig. 3: initially, geometry features $\mathcal{F}_G$ (extracted from the partial point cloud $\mathcal{P}$ using PointNet++) and per-point color features $\mathcal{F}_C$ are independently projected into modality-aware feature spaces. Subsequently, we perform mutual information minimization between them to explicitly encourage the acquisition of complementary information. Following this, these two branches are merged into a concise feature representation, which is then condensed through another fully connected layer, ultimately producing an adaptive weight map $\mathbf{w}$. We normalize this weight map to the interval [0,1] using the gumbel-softmax function. This process can be described as:

$$\mathbf{w} = \mathcal{GS}(\mathcal{W} \cdot \boldsymbol{H}(\text{Add}(\mathcal{U}\mathcal{F}_P, \mathcal{V}\mathcal{F}_I))) \tag{13}$$

where $\mathcal{W}$, $\mathcal{U}$, $\mathcal{V}$ denote the learnable weight matrices in our GC-Fuser. $\mathcal{F}_P$ and $\mathcal{F}_I$ are point cloud features and image

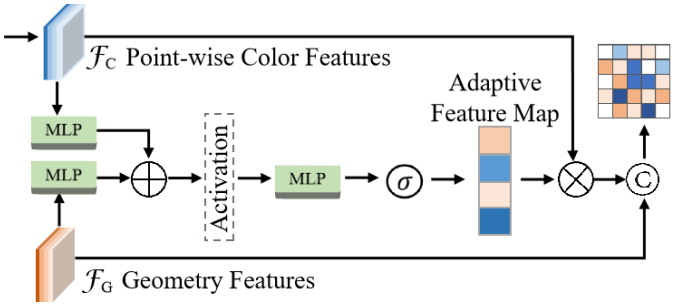

Fig. 3. **Illustration of the GC-Fuser.** The geometry features are applied to guide the fusion procedure with point-wise color features.

feature map, respectively. $\mathcal{GS}()$ represents Gumbel-Softmax and $\boldsymbol{H}()$ represents the output activation. Add$(\cdot)$ means the element-wise addition.

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:

$$\mathcal{F}_{CG} = \mathop{\Vert}_{i=1}^{C} (\mathbf{w}\mathcal{F}_C, \mathcal{F}_G) \tag{14}$$

where $\Vert$ represents channel-wise concatenation, $C = C_1 + C_2$.

C. Proposal Integration Scoring

The distance voting technique has demonstrated its efficacy as a location inference algorithm. It capitalizes on the inclusion similarity between the geometric configuration of visible points and the overall point cloud structure, utilizing local structures to deduce global pose information. Expanding upon this, we introduce an innovative 3D proposal integration scoring method. The fundamental concept is: within the 3D accumulator space, the quantity of valid proposals from visible points is tallied for each grid, and the confidence score is determined by aggregating these proposals.

Considering this, we first conduct the part segmentation. Given the fused feature $\mathcal{F}_{CG}$, we employ an MLP as the decoder to generate K masks, $\{m^{(k)} \in \mathbb{R}^{N\times 1} | k = 1, \ldots, K\}$. The part-level point cloud is then obtained using Eq. 15:

$$\mathcal{P}^{(k)} = \mathcal{P} \cdot m^{(k)}, \text{ where } m^{(k)} = \mathbb{1}(\delta_k = k) \tag{15}$$

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

In this way, we divide the articulation into K rigid parts. Afterward, for every rigid part, we independently produce a collection of keypoint proposals accompanied by their respective aggregation scores. To more vividly depict the proposal integration scoring mechanism, we draw an analogy between each visible point in the point cloud and a Member of Parliament (MP), whose main duty is to submit proposals for each grid within the 3D accumulator space. In this process, we employ the fused features $\mathcal{F}_{CG}$ followed by an MLP to generate $3N$ channels, signifying the predicted distances (radii) from each visible point to the three pre-defined keypoints. Additionally, 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

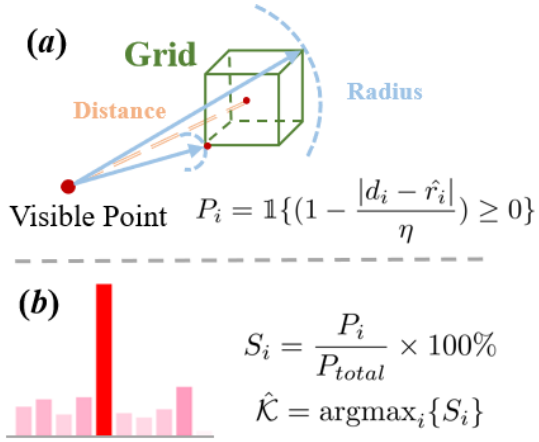

Fig. 4. **Illustration of our proposal integration scoring scheme**. (a) The principle of proposal generation. (b) Score aggregation and ranking, the grid with the highest score is excelled.

calculated radial distance and the estimated keypoint radius is less than or equal to ρ, the proposal is regarded as valid, and the proposal count for the corresponding grid is increased by 1; otherwise, the proposal is considered invalid, with the count staying at 0 (Fig. 4-a). We have adopted the confidence function from paper [52] and applied it to our proposal function, as mathematically represented below:

$$P_i = \mathbb{1}\left\{\left(1 - \frac{|d_i - \hat{r}_i|}{\eta}\right) \geq 0\right\} \tag{16}$$

Where $\mathbb{1}(\cdot)$ is the indicator function, d_i is the closest Euclidean distance from the visible point p_i to the grid, $\hat{r}_i$ is the predicted radius from p_i to the keypoint (output by the network), η is the body diagonal length of the grid voxel.

Following this, the proposal count for each grid is compiled to determine the confidence score (Fig. 4-b). In this context, the confidence score is characterized as the proportion of the current grid's proposal count relative to the total proposal count. The grid achieving the highest score is chosen as the definitive result (Eq. 17), and its central point is recognized as the elected keypoint $\hat{\mathcal{K}}$ (Eq. 18).

$$S_i = \frac{P_i}{P_{total}} \times 100\% \tag{17}$$

$$\hat{\mathcal{K}} = \text{argmax}_i\{S_i\} \tag{18}$$

The intricate procedure is delineated in Algorithm 1. In summary, once the network is adequately trained, the bulk of the proposals produced by local feature points can precisely pinpoint the target keypoint, and the final estimation of the center point aligns with the ground truth. Even when a minor number of noisy votes are present, no substantial deviation is observed, underscoring the robustness of our approach. From an alternative viewpoint, our aggregation step proficiently estimates the probability density of the center point's position within the spatial domain (i.e., the keypoint) and ascertains its location accordingly. It is noteworthy that this process remains invariant to rigid transformations (i.e., SE (3) invariance), ensuring steadfastness even as the object's pose alters, and preserving a dominant peak in the distribution.

Algorithm 1 Proposal Integration Scoring Scheme:

Input Predicted radius of each visible point: $\{r_n\}_{n=1}^{n=N}$.
Parameter: Distance threshold η.
Output: The predicted keypoint $\mathcal{K}$.

```
Divide the 3D space into multiple grids;
Initialize the proposal pool P = {} for each grid in 3D accumulator space.
for K parts do
    for N visible point {p_n}_{n=1}^{n=N} do
        Compute the Euclidean distance d_n from the visible point p_n to each grid.
    end for
end for
if r_n − d_n < η and r_n − d_n > 0 then
    the proposal belongs to the grid increases by one (Eq. 16).
end if
Compute the total scores by Eq. 17.
Output the center of the grid that has the highest score as the predicted keypoint: K = argmax(S) by Eq. 18.
```

D. Loss Functions

In the training process, our target is to segment the articulation into K rigid parts (optional) and conduct the radius prediction. Specifically, we use the Cross-Entropy loss $\mathcal{L}_{seg}$ for part segmentation and MSE loss $\mathcal{L}_{rad}$ for radius prediction. Overall, the final loss of training process $\mathcal{L}_{tr}$ can be summarized as:

$$\mathcal{L}_{tr} = \tau_1 \mathcal{L}_{seg} + \tau_2 \mathcal{L}_{rad} \tag{19}$$

$$\text{where } \mathcal{L}_{seg} = \frac{1}{K} \Sigma_{k=1}^{K} (m^{(k),*}, \hat{m}^{(k)}), \tag{20}$$

$$\mathcal{L}_{rad} = \frac{1}{K} \Sigma_{k=1}^{K} \left\{ \frac{1}{N} \Sigma_{i=1}^{N} (r_i^{(k),*}, \hat{r}_i^{(k)}) \right\} \tag{21}$$

where $m^{(k)}$ is k-th part predicted mask, which can be referred to in Eqn. (15), $r_i^{(k)}$ is k-th part i-th visible point's radius. We use a convex combination of these losses, where τ_1, τ_2 are 0.25 and 0.75, respectively.

It is noted the above process and the training of the KP-Estimator (its loss functions are detailed in Sec. V-A) are independent of each other.

VI. Experiments

Datasets, Baselines, and Metrics: Our PAGE framework is assessed using the ArtImage [37], ReArtMix [55], and RobotArm [55] datasets, encompassing synthetic, semi-synthetic, and real-world environments. For comparative analysis, we measure our performance against four RGB+Depth-based methodologies: A-NCSH [53], DenseFusion [13], ASMM [54]. Additionally, more voting-based methods [38], [39], [40] are adopted for comparison. The metrics for evaluation comprise the degree error for 3D rotation, the distance error for 3D translation, and the 3D Intersection over Union (IoU) for assessing 3D scale.

Implementation Details: In the pre-processing phase, the input RGB images are resized to a resolution of 224 × 224, and the point clouds are reduced to 2,048 points before being input into the network. It is important to note that the input modality for all methods has been standardized to depth + color, and re-training was performed under identical experimental conditions as ours. For training the backbone (PointNet++ [56]), the Adam optimizer was utilized with an initial learning rate of 0.001 and a weight decay of 0.0001. The learning rate was reduced by a factor of 0.5 every 20 epochs. All experiments were carried out on four NVIDIA GeForce RTX 4090 GPUs, each equipped with 24GB of memory.

For the main network training, we adopt a category-level learning paradigm, consistent with [53]. Our proposed PAGE framework is trained and evaluated on three representative datasets: ArtImage (synthetic), ReArtMix (semi-synthetic), and RobotArm (real-world), covering a diverse range of data domains.

The KP-Estimator is trained independently for each object category, using the same dataset splits as the main pose estimation network. For every category, we generate a consistent set of spatially distributed and geometrically meaningful keypoints from the corresponding point clouds. These keypoints serve as fixed anchors for downstream 6D pose inference. As detailed in Section IV-A, the KP-Estimator is trained in a self-supervised fashion using customized geometric loss functions—namely, Wasserstein loss, separation loss, and coverage loss—to learn category-specific pre-defined keypoints. Notably, this process does not require any manual keypoint annotations.

A. Comparison With the SOTA Methods

In this section, we perform comparative experiments within the synthetic dataset (namely, ArtImage) alongside classical methods, with the objective of validating the efficacy of our PAGE. Tab. I presents the quantitative results. 1) Overall, our approach has consistently surpassed the SOTA benchmarks, demonstrating the effectiveness of the integration of each proposed module. 2) Individually, we achieve the most accurate pose estimation in the *eyeglasses* category, with rotation and translation errors of ($\mathbf{3.2}^\circ, \mathbf{5.2}^\circ, \mathbf{5.1}^\circ$), (**0.027**m, **0.075**m, **0.071**m), respectively. This success can be credited to our keypoint generation strategy, which excels in adapting to objects with consistent geometric scales. Regarding the 3D IoU metric, our prediction errors are markedly superior for each component when compared to the baseline model (A-NCSH), with the average value **52.3%** vs. **44.1%**. Compared with other voting-based methods, our method is also more competitive, which we believe can be attributed to the synergy of the customized modules proposed. Qualitative results in Fig. 5 (left) further underscore that our predictions are in tight concordance with GT, affirming the robustness and accuracy of our methodology.

B. Ablation Study

1) Pre-Defined Keypoints: As outlined in Sec. V-A, we propose that pre-defined keypoint methods can surpass heuristic approaches. Consequently, we undertake comprehensive ablation experiments in this section. The results are presented

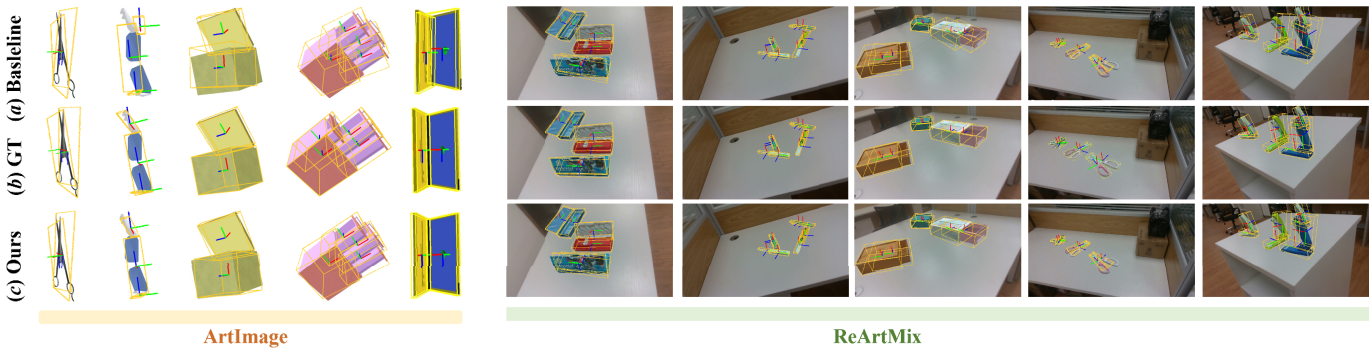

Fig. 5. **Qualitative results on artimage and ReArtMix.** The left is a synthetic dataset and the right is the semi-synthetic scenario.

TABLE I

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

Category	Method	Per-part Pose		
		rotation error (°) ↓	translation error (m) ↓	3D IoU (%) ↑
Laptop	A-NCSH [53]	5.3, 5.4	0.054, 0.043	56.7, 40.2
	DenseFusion [13]	5.4, 4.3	0.062, 0.061	43.5, 24.1
	PoseCNN [38]	5.1, 4.7	0.047, 0.056	49.2, 36.7
	6-PACK [39]	5.0, 4.5	0.046, 0.037	52.3, 41.1
	CPPF [40]	4.7, 4.5	0.043, 0.032	59.2, 44.3
	ASMM [54]	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 [53]	3.5, 18.3, 18.2	0.043, 0.286, 0.283	52.5, 40.2, 39.6
	DenseFusion [13]	4.9, 7.5, 7.5	0.062, 0.103, 0.104	46.8, 41.5, 38.4
	PoseCNN [38]	4.2, 8.2, 9.3	0.058, 0.142, 0.158	43.2, 39.8, 36.5
	6-PACK [39]	3.7, 6.8, 7.4	0.046, 0.098, 0.101	47.2, 45.1, 39.9
	CPPF [40]	3.3, 5.8, 5.9	0.038, 0.086, 0.125	54.1, 42.8, 43.7
	ASMM [54]	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 [53]	4.0, 4.8	0.059, 0.123	84.3,56.2
	DenseFusion [13]	6.0, 6.2	0.104, 0.142	66.5,38.9
	PoseCNN [38]	6.5, 7.1	0.128, 0.153	61.7, 49.0
	6-PACK [39]	9.3, 8.7	0.135, 0.146	55.4, 35.1
	CPPF [40]	5.8, 5.4	0.078, 0.095	80.1, 62.4
	ASMM [54]	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 [53]	2.0, **2.6**	0.035, **0.021**	45.8,44.8
	DenseFusion [13]	3.9, 3.4	0.048, 0.039	35.6,34.5
	PoseCNN [38]	4.1, 4.2	0.055, 0.046	33.1, 32.6
	6-PACK [39]	3.8, 6.2	0.047, 0.066	37.2, 34.5
	CPPF [40]	3.2, 5.7	0.024, 0.045	44.9, 38.2
	ASMM [54]	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 [53]	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 [13]	4.4, 4.4, 4.4, 4.4	0.111, 0.143, 0.144, 0.115	75.8, 73.4, 70.2, 71.3
	PoseCNN [38]	3.7, 3.7, 4.3, 4.2	0.058, 0.122, 0.131, 0.087	78.2, 71.8, 69.5, 70.3
	6-PACK [39]	3.2, 5.2, 4.7, 4.9	0.122, 0.103, 0.110, 0.138	81.9, 71.2, 73.4, 70.7
	CPPF [40]	2.9, 4.3, 4.4, 4.0	0.076, 0.243, 0.136, 0.152	83.5, 72.8, 70.6, 74.5
	ASMM [54]	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**

in Tab. II (I-III). It is observed that the FPS and BBox based keypoint generation techniques are adapted from [57]. The quantitative results can be concluded that: our approach exceeds the heuristic methods (i.e., FPS and BBox based methods) by a considerable margin, reporting (1.9° and 5.4°) and (0.013m, 0.032m) with our methods. In summary, the proposed method not only resolves the keypoint distribution but also considers the geometric and structural relationships of

TABLE II

ABLATION STUDY. IT IS NOTED THAT EXPERIMENTS ARE CONDUCTED ON THE CATEGORY SCISSORS

Configuration	Keypoints Generation	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
Configuration	Numbers	Per-part Pose	
VII (Ours)	3	1.9, 5.4	0.013, 0.032
VIII	4	1.8, 5.2	0.011, 0.031
IX	5	1.7, 5.1	0.010, 0.027
Configuration	Prediction	Per-part Pose	
X	Direct Regression	10.5, 15.4	0.135, 0.237
XI (Ours)	Scoring	1.7, 5.1	0.010, 0.027
Configuration	Voting	Per-part Pose	
X	6-PACK [39]	7.8, 8.5	0.068, 0.064
XI (Ours)	Ours	1.7, 5.1	0.010, 0.027

TABLE III

POSE ESTIMATING RESULTS ON REARTMIX DATASET

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**

the articulation, refined through three geometry-sensitive loss functions (see Sec. V-A). This is essential for keypoint-based object pose modeling.

2) Fusion Manner: In this study, we introduce an innovative multi-modal learning approach termed geometric-color alignment, as elaborated in Sec. V-B. To validate the efficacy of the proposed module, we perform ablation experiments as shown in Tab. II (IV-VI). It should be noted that configuration IV represents the straightforward concatenation of depth and RGB features. Configuration V denotes the method from densefusion [13]. It is evident that our method attains state-of-the-art performance, whereas the conventional method (i.e., direct concatenation) experiences significant performance degradation due to noisy feature alignment.

3) Keypoints Numbers: In this work, we propose to use 3 keypoints for each rigid part modeling. However, we are curious about how the performance of the proposed method varies with changes in the number of keypoints. Therefore, we conducted ablation experiments related to the number of keypoints, quantitative results can be seen in Tab. II (VII-IX). We observed that increasing the number of keypoints did not significantly reduce rotation and translation errors. For instance, when the number of keypoints increased from 3 to 5, the average rotation error only decreased by 0.2°, and the average translation error dropped by 0.04m. Therefore, to balance performance and computational cost, 3 keypoints are selected for each rigid part to model its shape. It is worth emphasizing that at least three non-collinear keypoints are required to uniquely determine a rigid transformation. When the number of keypoints is less than 3, shape modeling becomes ambiguous.

4) Direct Regression versus Scoring: As mentioned before, our method not only proposes a novel pre-defined keypoint generation method but also introduces an innovative algorithm for keypoint-based pose modeling of articulated objects, named *scoring*. In fact, the generated keypoints can also be directly used for regression tasks, and we have conducted relevant ablation experiments in Tab II (X-XI). The experimental results show that our scoring method outperforms direct regression methods (for example, our rotation errors and translation errors are $(1.7°, 5.1°)$ and (0.010m, 0.027m), while the error of the direct regression method is $(10.5°, 15.4°)$ and (0.135m, 0.237m). We conjecture that the performance improvement stems from our scoring method, which integrates the predictions from all points, leading to a more stable and reliable final result.

5) Voting-Based Methods: To further assess the effectiveness of our scoring-based keypoint inference strategy, we perform a comparative ablation study with 6-PACK, a representative 3D voting-based pose estimation method. For a fair comparison, we adapt 6-PACK to the part-level estimation setting, where each articulated object is decomposed into rigid segments, and pose is predicted independently for each part. Both methods are evaluated under identical input settings. As shown in Table II, our method significantly outperforms 6-PACK in both rotation and translation accuracy. We attribute this to two main factors: (1) our geometry-aware scoring formulation provides better robustness to partial observations than anchor-based proposal aggregation, and (2) our predefined keypoints are optimized for spatial consistency and category-level generalization, which 6-PACK does not enforce.

C. Generalization Capacity

1) Experiments on Semi-Synthetic Scenarios: The ReArtMix dataset is utilized to assess our method, which includes semi-synthetic scenarios. The specific results are detailed in Tab. III. Our method attains the highest performance in the *Drawer* category, with a mere **1.2**° rotation error, and translation errors of **0.013**m and **0.017**m. This indicates that our approach is equally potent in Semi-Synthetic Scenarios, highlighting its robustness and versatility across various object categories and environments. Qualitative results for the five categories are depicted in Fig. 5 (Right).

2) 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. V), 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

TABLE IV

EXPERIMENTS ON THE INTEGRATION OF RGB INFORMATION ON ARTIMAGE [37]. NOTE THAT THE CATEGORIES LAPTOP, EYEGLASSES, DISHWASHER, AND SCISSORS CONTAIN ONLY FREE JOINT AND REVOLUTE JOINTS, AND THE DRAWER CATEGORY CONTAINS FREE JOINT AND PRISMATIC JOINTS. THE ACCURACY IS REPORTED FOR THE PERFORMANCE OF PART SEGMENTATION

Category	Input Modality	Part Segmentation (↑)	Per-part Pose	
			rotation error (°) ↓	translation error (m) ↓
Eyeglasses	Depth	0.97	5.3, 8.1, 8.5	0.062, 0.112, 0.123
	Depth + RGB	**0.99**	**3.2**, **5.2**, **5.1**	**0.027**, **0.075**, **0.071**
Scissors	Depth	0.96	3.9, 7.3	0.039, 0.058
	Depth + RGB	**0.98**	**1.9**, **5.4**	**0.013**, **0.032**
Laptop	Depth	0.95	6.7, 4.2	0.072, 0.053
	Depth + RGB	**0.97**	**4.0**, **1.7**	**0.014**, **0.019**
Dishwasher	Depth	0.98	6.1, 7.5	0.086, 0.102
	Depth + RGB	**0.99**	**3.9**, **4.3**	**0.055**, **0.079**
Drawer	Depth	0.94	4.4, 4.4, 4.4, 4.4	0.031, 0.263, 0.272, 0.243
	Depth + RGB	**0.98**	**2.8**, **2.8**, **2.8**, **2.8**	**0.010**, **0.017**, **0.015**, **0.013**

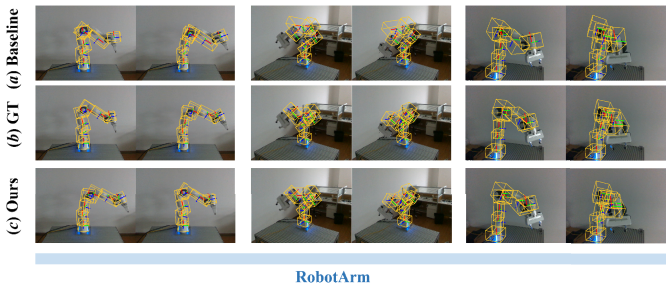

Fig. 6. Qualitative results on 7-part RobotArm dataset.

TABLE V

QUANTITATIVE RESULTS ON 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**

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. 6.

D. Experiments of Hyperparameter Selection

★ (1) τ_1 **and** τ_2**.** Regarding the training loss weights (τ_1, τ_2), the part segmentation module is optional in our framework, while radius prediction plays a decisive role in keypoint estimation accuracy. Therefore, based on empirical observations, we assign a lower weight to segmentation (τ_1 = 0.25) and a higher weight to radius prediction (τ_2 = 0.75).

To validate this weighting strategy, we conducted an ablation study by varying τ_1 and τ_2 to assess their effects on pose estimation accuracy. The results, summarized in Tab. VI, reveal that increasing τ_2 (i.e., emphasizing radius prediction)

TABLE VI

ABLATION STUDY OF τ_1 AND τ_2. IT IS NOTED THAT EXPERIMENTS ARE CONDUCTED ON THE CATEGORY SCISSORS

Configuration	Loss Weight $\mathcal{L}_{seg}$, $\mathcal{L}_{rad}$	Per-part Pose	
		rotation error (°) ↓	translation error (m) ↓
I	0.9, 0.1	8.8, 12.8	0.123, 0.227
II	0.75, 0.25	3.2, 9.8	0.045, 0.086
III	0.5, 0.5	2.1, 5.9	0.023, 0.040
IV (Ours)	0.25, 0.75	1.9, 5.4	0.013, 0.032
V	0.1, 0.9	1.8, 5.3	0.013, 0.030

TABLE VII

ABLATION STUDY OF λ_1 AND λ_2. IT IS NOTED THAT EXPERIMENTS ARE CONDUCTED ON THE CATEGORY SCISSORS

Configuration	Loss Weight $\mathcal{L}_{wass}$, $\mathcal{L}_{sep}$, $\mathcal{L}_{cov}$	Per-part Pose	
		rotation error (°) ↓	translation error (m) ↓
I	0.2, 0.4, 0.4	3.8, 8.5	0.068, 0.142
II	0.4, 0.3, 0.3	2.1, 5.8	0.019, 0.077
III (Ours)	0.6, 0.2, 0.2	1.9, 5.4	0.013, 0.032
IV	0.8, 0.1, 0.1	1.9, 5.3	0.014, 0.032

TABLE VIII

SELF-OCCLUSION ANALYSIS. POSE ERRORS FOR 'DRAWER' CATEGORY WITH VARYING OCCLUSION LEVELS

Occlusion Level (Visibility)	Rotation Error(°)	Translation Error(m)
0%-40%	2.2	0.013
40%-80%	2.7	0.013
80%-100%	3.1	0.016

consistently enhances performance—particularly in rotation accuracy—highlighting the importance of the radius regression task. However, when τ_2 exceeds 0.75, the performance gain saturates, and excessive down-weighting of the segmentation loss weakens geometric regularization, leading to reduced generalization in complex scenes.

Consequently, the configuration of τ_1 = 0.25 and τ_2 = 0.75 achieves a ***well-balanced trade-off*** between the two objectives and has been empirically verified to ensure stable and effective convergence.

★ (2) λ_1 and λ_2. For the KP-Estimator loss weights (λ_1, λ_2), the Wasserstein loss serves as the primary component, enforcing consistent and uniform radius distributions across different object parts and instances—an essential factor for reliable keypoint generation. Consequently, it is assigned a

TABLE IX

ROBUSTNESS EVALUATION. NOTE THAT THE CATEGORY LAPTOP IS USED FOR EVALUATION

Noise Setting	Method	Per-part Pose	
		rotation error (°) ↓	translation error (m) ↓
Gaussian	CPPF	4.9, 3.6	0.039, 0.066
	Ours	**4.3**, **1.9**	**0.019**, **0.030**
Scaling	CPPF	4.9, 4.5	0.036, 0.036
	Ours	**4.1**, **1.7**	**0.016**, **0.028**
Offset	CPPF	7.8, 4.3	0.089, 0.053
	Ours	**4.4**, **1.8**	**0.015**, **0.023**

TABLE X

THE RADII COMPARISON OF DIFFERENT KEYPOINT ESTIMATION METHODS. NOTE THAT THE CATEGORY LAPTOP IS USED FOR DEMONSTRATION. THE RADIUS UNIT IS (m)

Pre-defined Keypoints			Heuristic Method (FPS)		
κ_1	κ_2	κ_3	κ_1	κ_2	κ_3
0.634	0.634	0.635	0.036	0.061	0.057
0.634	0.634	0.638	0.048	0.068	0.062
0.670	0.663	0.665	0.048	0.070	0.075
0.671	0.664	0.668	0.058	0.089	0.083
...	...	...	...	...	...
0.695	0.664	0.697	0.066	0.102	0.094
0.697	0.690	0.701	0.070	0.124	0.097
0.697	0.696	0.708	0.086	0.132	0.140
0.708	0.718	0.724	0.118	0.157	0.173
Variance of Radius Cluster					
0.039	0.063	0.027	0.193	0.115	0.104

relatively high weight ($\lambda_1 = 0.6$). The separation loss ($\lambda_2 = 0.2$) mitigates spatial redundancy among keypoints, while the remaining weight is allocated to the coverage loss to promote uniform keypoint distribution across the visible surface. To validate this design, we conducted an ablation study by varying λ_1 and λ_2 and examining their influence on pose estimation accuracy. The results, summarized in Table X of the revised manuscript, are detailed in Tab VII.

When λ_1 decreases below 0.4, performance drops markedly—particularly in translation accuracy—as the model loses distributional supervision and begins to resemble heuristic keypoint baselines. Conversely, increasing λ_1 beyond 0.6 yields marginal gains, indicating that the performance improvement saturates once sufficient distributional uniformity is achieved.

These findings empirically confirm that the configuration $\lambda_1 = 0.6$ and $\lambda_2 = 0.2$ provides a well-balanced trade-off among distributional consistency, spatial separation, and surface coverage, ultimately leading to more accurate and stable pose estimation.

E. Robustness Experiment

1) Self-Occlusion Analysis: To further assess the robustness of our PAGE under self-occlusion, we divide the test samples of the *Drawer* category into three subsets based on the level of occlusion, as self-occlusion is more prevalent in drawers. The occlusion level is defined as the ratio of visible points compared to the total number of points on the part, similar to the A-NCSH [53]. Specifically, we use three subsets with varying visibility: (0%, 40%), (40%, 80%), and (80%, 100%). For evaluation, we employ mean rotation error and mean translation error as performance metrics for pose estimation. The experimental results are presented in Tab. VIII. As the occlusion level increases, the rotation error remains relatively stable, thanks to our multi-modal learning manner. Our method demonstrates minimal translation error even at occlusion levels up to 80%. In extreme cases, with occlusion rates ranging from 80% to 100%, our method still delivers favorable results, with a translation error of only 0.016 m. These results demonstrate that our method effectively addresses the self-occlusion problem.

2) Depth Only: In this paper, we combine color and geometric information for pose estimation of articulated objects. In contrast, using only depth information as input results in performance degradation across all categories. For instance, in the *Laptop* category, our method demonstrates a significant performance improvement. Specifically, for pose estimation, the results using only depth information are $(6.7°, 4.2°)$ and (0.072m, 0.053m), while the results using RGB+depth information are $(4.0°, 1.7°)$ and (0.014m, 0.019m). For part segmentation, RGB information also plays a key role in this task, which suppresses the methods that only depth information is involved (e.g., 0.99 accuracy of Depth+RGB vs., 0.97 of Depth). These experimental results validate the effectiveness of our proposed geometric-color alignment. Our framework explicitly encourages the interaction between geometric and color features and facilitates complementary learning, thus enhancing overall performance.

3) Noise: To more systematically analyze the robustness of our model, we introduce perturbations into the input point clouds. Three types of synthetic noise are applied independently and jointly: 1) Gaussian Positional Noise: Random zero-mean Gaussian noise with standard deviation $\sigma \in [0.01, 0.03]$ m is added to each point coordinate. This simulates sensor jitter and frame-to-frame variation. 2) Scaling Noise: A global scale factor is sampled from the range $[0.95, 1.05]$ and applied to all 3D points, mimicking depth distortion due to calibration drift or focal length error. 3) Offset Noise: A uniform additive offset of $\pm$1.5mm to $\pm$5.0mm is introduced along each axis to simulate systemic sensor bias or spatial drift.

We evaluate the performance under these perturbations on the Laptop category. Results are summarized in Table IX. Compared with voting-based baselines such as CPPF, our method shows significantly less performance degradation. For instance, under Gaussian noise with $\sigma = 0.03$m, our method retains a translation error below 0.025m and a rotation error under 3.1°, while CPPF's error increases substantially (translation is 0.52m, rotation is 4.2°). The improved robustness of our method can be attributed to two key factors: 1) Geometrically-consistent scoring: Unlike traditional vote counting or pairwise correspondence approaches, our proposal integration strategy evaluates radius alignment across all observable points in a continuous 3D accumulator space. This reduces the influence of noisy local patterns and stabilizes keypoint inference. 2) Category-level keypoint representation: Our pre-defined keypoints are learned to be consistent across instances and

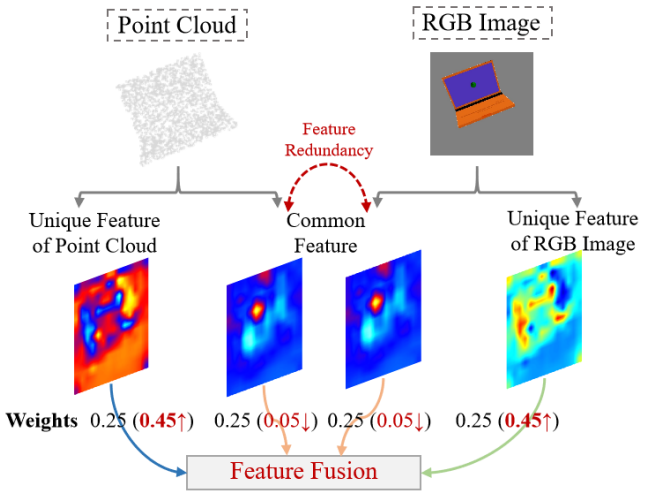

Fig. 7. **The decomposed components of the point cloud features and RGB image features.** Note that we use the "0.25 (0.45 ↑)" mode to represent the weights from the previous method and our method.

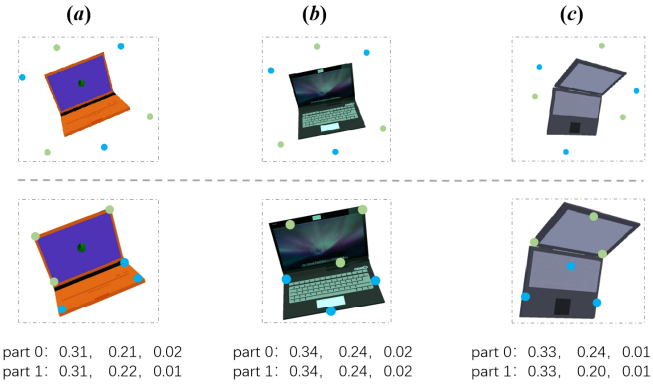

Fig. 8. **Visualization results of keypoints generated by different methods.** The values are Per-part scale (m).

TABLE XI
COMPARISON OF INFERENCE SPEED

Method	DenseFusion	ASMM	PAGE (Ours)
FPS	4.54	5.13	6.25

structurally distributed across object parts, making them more resistant to partial observations and spatial perturbations. These results confirm that the proposed method is not only accurate but also robust to various input degradations, thereby supporting its deployment in real-world applications involving articulated objects and noisy RGB-D inputs.

F. Analysis and Discussion

1) Pre-Defined Keypoint Estimation, Extended: In this paper, we observed that keypoints generated using heuristic algorithms exhibit significant differences (i.e., variance of Euclidean distances) in the distribution of radius clusters. For example, refer to the Tab. X. The distance clusters of the three keypoints generated using the FPS algorithm have variances of 0.193, 0.115, and 0.104, respectively. Therefore, we adopted a customized method to generate pre-defined keypoints, and the results show that the variance significantly decreased to 0.039, 0.063, and 0.027, while the intra-cluster differences between the radius clusters were also much smaller. This indicates that our method can better control the distribution of radius clusters when generating keypoints, thereby enhancing the stability and accuracy of the model.

2) Geometric-Color Alignment, Extended: Geometric and color features capture the same scene through different modalities, resulting in both shared information and unique features, as illustrated in Fig. 7. Existing advanced multi-modal learning methods rarely perform explicit alignment and complementary information learning between modalities. Specifically, as shown in the weight illustration in Fig. 7, previous approaches often assign relatively uniform weights to each feature, leading to the presence of redundant information and the learning of replication artifacts [58], [59]. To address these limitations, our work emphasizes feature alignment between different modalities and enhances the learning of complementary features by assigning higher weights to the unique characteristics of each modality. This strategy effectively reduces information redundancy and significantly improves the performance of the articulation pose estimation task.

3) Definition of Observable Point: i.e., visible point, refers to a surface point that is not occluded from the current viewpoint and can be captured by a depth sensor. In contrast, invisible points are either located in self-occluded regions blocked by foreground structures or lie outside the sensor's field of view.

G. Visualization

In contrast to instance-level object pose estimation, category-level articulated object estimators aim to predict the rotation, translation, and scale of rigid parts of observed articulated objects without the need for known CAD models. Classical methods typically achieve accurate 6D pose estimation by predicting shape descriptors for objects within the same category. For instance, a prominent approach leverages a normalized object coordinate space [60] and its variants [53], [61], where the coordinates of each pixel in the normalized space are predicted to establish a unified representation for objects of the same category.

We present a visualization of different keypoint generation methods in Fig. 8. As shown, our method produces keypoints that are stable across different instances within the same category, enabling generalization at the category level. In contrast, traditional methods exhibit scale sensitivity, leading to significant intra-category discrepancy regarding keypoint distribution. *In essence, traditional methods are limited to instance-level pose estimation, whereas our approach extends to category-level generalization. We define this property as* ***inter-category stability****, which mitigates the instance discrepancies that prior methods struggled with.*

VII. INFERENCE SPEED

Based on an NVIDIA GeForce RTX 4090 GPU (batch size is set to 1), our method takes about 15ms to conduct the geometric-color alignment module, and 115ms to go through the proposal integration scoring module. which achieves about

160 ± 10ms to finish a single prediction, while DenseFusion takes about 220 ± 10ms and ASMM takes 195 ± 10ms under the same setting. Our method achieves competitive inference speed as well as state-of-the-arts. The FPS results comparison is shown in Tab. XI.

VIII. Conclusion

In this work, we propose a novel framework coined PAGE, aiming to cope with category-level articulation pose estimation via multi-modal alignment. To deal with multi-modal learning problems, our key idea is first aligning (geometry and color features) and then fusing. Considering the weakness of heuristics keypoints, the proposed pre-defined keypoint estimation is tailored to help improve the performance of articulation pose estimation. Finally, we elaborate on the proposal integration scoring strategy to determine 6D pose of the articulated object. Experimental results show that our method achieves state-of-the-art performance in pose estimation on the synthetic ArtImage dataset. Additionally, it demonstrates strong generalization capabilities on the semi-synthetic ReArtMix dataset and real-world multi-hinged articulated object datasets (RobotArm). This highlights the robustness and versatility of our approach across diverse scenarios.

References

[1] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, p. 8414, Jun. 2019.

[2] M. T. Mason, *Mechanics of Robotic Manipulation*. Cambridge, MA, USA: MIT Press, 2001.

[3] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Proc. Conf. Robot Learn.*, 2022, pp. 894–906.

[4] Y. Liu et al., "Thin, soft, garment-integrated triboelectric nanogenerators for energy harvesting and human machine interfaces," *EcoMat*, vol. 3, no. 4, p. 12123, Aug. 2021.

[5] D. Amin and S. Govilkar, "Comparative study of augmented reality SDKS," *Int. J. Comput. Sci. Appl.*, vol. 5, no. 1, pp. 11–26, 2015.

[6] E. Kang, H. Qiao, Z. Chen, and J. Gao, "Tracking of uncertain robotic manipulators using event-triggered model predictive control with learning terminal cost," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 4, pp. 2801–2815, Oct. 2022.

[7] J. Hou, B. Graham, M. Nießner, and S. Xie, "Exploring data-efficient 3D scene understanding with contrastive scene contexts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15582–15592.

[8] M. Jaritz, J. Gu, and H. Su, "Multi-view PointNet for 3D scene understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3995–4003.

[9] L. Liu, Z. Min, J. Wang, R. Wang, L. Zhang, and Y. Zhong, "Rethinking 3D convolution in ℓ_p-norm space," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 91012–91035.

[10] T. Yu, X. Lin, S. Wang, W. Sheng, Q. Huang, and J. Yu, "A comprehensive survey of 3D dense captioning: Localizing and describing objects in 3D scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 3, pp. 1322–1338, Mar. 2024.

[11] F. Tang, Y. Wu, X. Hou, and H. Ling, "3D mapping and 6D pose computation for real time augmented reality on cylindrical objects," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 2887–2899, Sep. 2020.

[12] J. Carmigniani and B. Furht, "Augmented reality: An overview," in *Handbook Augmented Reality*. Heidelberg, Germany: Springer, 2011, pp. 3–46.

[13] C. Wang et al., "DenseFusion: 6D object pose estimation by iterative dense fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3343–3352.

[14] Y. Li et al., "Tokenpose: Learning keypoint tokens for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Mar. 2021, pp. 11313–11322.

[15] C. Ye, S. Hong, and A. Tamjidi, "6-DOF pose estimation of a robotic navigation aid by tracking visual and geometric features," *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 4, pp. 1169–1180, Oct. 2015.

[16] K. Sun et al., "Bottom-up human pose estimation by ranking heatmap-guided adaptive keypoint estimates," 2020, *arXiv:2006.15480*.

[17] Y. Lin, J. Tremblay, S. Tyree, P. A. Vela, and S. Birchfield, "Keypoint-based category-level object pose tracking from an RGB sequence with uncertainty estimation," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 1258–1264.

[18] Z. Min, D. Zhu, H. Ren, and M. Q.-H. Meng, "Feature-guided nonrigid 3-D point set registration framework for image-guided liver surgery: From isotropic positional noise to anisotropic positional noise," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 2, pp. 471–483, Apr. 2021.

[19] P. Wohlhart and V. Lepetit, "Learning descriptors for object recognition and 3D pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3109–3118.

[20] A. Tejani, D. Tang, R. Kouskouridas, and T.-K. Kim, "Latent-class Hough forests for 3D object detection and pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 462–477.

[21] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6D object pose estimation using 3D object coordinates," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 536–551.

[22] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1954–1963.

[23] L. Zhang et al., "R^2-art: Category-level articulation pose estimation from single RGB image via cascade render strategy," in *Proc. AAAI Conf. Artif. Intell.*, 2025, vol. 39, no. 9, pp. 9985–9993.

[24] M. Raessa, J. C. Y. Chen, W. Wan, and K. Harada, "Human-in-the-loop robotic manipulation planning for collaborative assembly," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 4, pp. 1800–1813, Oct. 2020.

[25] W. Zhang et al., "Design and development of a new biped robotic system for exoskeleton-structure window cleaning," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 3160–3171, 2025.

[26] J. Zhang, Z. Tu, J. Yang, Y. Chen, and J. Yuan, "MixSTE: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13232–13242.

[27] J. Zhang, Y. Chen, and Z. Tu, "Uncertainty-aware 3D human pose estimation from monocular video," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 5102–5113.

[28] L. Liu et al., "Category-level articulated object 9D pose estimation via reinforcement learning," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 728–736.

[29] L. Zhang et al., "Vocapter: Voting-based pose tracking for category-level articulated object via inter-frame priors," in *Proc. ACM Multimedia*, 2024, pp. 8942–8951.

[30] L. Zhang, "U-cope: Taking a further step to universal 9d category-level object pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2025, pp. 254–270.

[31] L. Zhang et al., "GaPT-DAR: Category-level garments pose tracking via integrated 2D deformation and 3D reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 22638–22647.

[32] R. Wang, Y. Yang, and D. Tao, "ART-point: Improving rotation robustness of point cloud classifiers via adversarial rotation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14351–14360.

[33] T. Lee et al., "UDA-COPE: Unsupervised domain adaptation for category-level object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14891–14900.

[34] Q. Huang, X. Huang, B. Sun, Z. Zhang, J. Jiang, and C. Bajaj, "ARAPReg: An as-rigid-as possible regularization loss for learning deformable shape generators," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5795–5805.

[35] W. McNally, K. Vats, A. Wong, and J. McPhee, "Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 37–54.

[36] C. Gong, Y. Zhang, Y. Wei, X. Du, L. Su, and Z. Weng, "Multicow pose estimation based on keypoint extraction," *PLoS ONE*, vol. 17, no. 6, 2022, Art. no. 0269259.

[37] H. Xue, L. Liu, W. Xu, H. Fu, and C. Lu, "OMAD: Object model with articulated deformations for pose estimation and retrieval," 2021, *arXiv:2112.07334*.

[38] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," 2017, *arXiv:1711.00199*.

[39] C. Wang et al., "6-PACK: Category-level 6D pose tracker with anchor-based keypoints," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 10059–10066.

[40] Y. You, R. Shi, W. Wang, and C. Lu, "CPPF: Towards robust category-level 9D pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6856–6865.

[41] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," *Proc. SPIE*, vol. 1611, pp. 586–606, Apr. 1992.

[42] B. K. P. Horn, H. M. Hilden, and S. Negahdaripour, "Closed-form solution of absolute orientation using orthonormal matrices," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 5, no. 7, pp. 1127–1135, 1988.

[43] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[44] S. Qiao, Y. Wang, and J. Li, "Real-time human gesture grading based on OpenPose," in *Proc. 10th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2017, pp. 1–6.

[45] H.-C. Nguyen, T.-H. Nguyen, J. Nowak, A. Byrski, A. Siwocha, and V.-H. Le, "Combined YOLOv5 and HRNet for high accuracy 2D keypoint and human pose estimation," *J. Artif. Intell. Soft Comput. Res.*, vol. 12, no. 4, pp. 281–298, Oct. 2022.

[46] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Mali, Jun. 2018, pp. 224–236.

[47] C. Rockwell, N. Kulkarni, L. Jin, J. J. Park, J. Johnson, and D. F. Fouhey, "FAR: Flexible, accurate and robust 6DoF relative camera pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 19854–19864.

[48] X. Zhu and X. Ye, "GAN-BodyPose: Real-time 3D human body pose data key point detection and quality assessment assisted by generative adversarial network," *Image Vis. Comput.*, vol. 149, Sep. 2024, Art. no. 105144.

[49] J. Solà, J. Deray, and D. Atchuthan, "A micro lie theory for state estimation in robotics," 2018, *arXiv:1812.01537*.

[50] Z. Qiu, K. Qiu, J. Fu, and D. Fu, "DGCN: Dynamic graph convolutional network for efficient multi-person pose estimation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11924–11931.

[51] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 105–115.

[52] B. Tekin, S. N. Sinha, and P. Fua, "Real-time seamless single shot 6D object pose prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 292–301.

[53] X. Li, H. Wang, L. Yi, L. J. Guibas, A. L. Abbott, and S. Song, "Category-level articulated object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3703–3712.

[54] H.-B. Zhang, J.-X. Hong, J.-H. Liu, Q. Lei, and J.-X. Du, "Images, normal maps and point clouds fusion decoder for 6D pose estimation," *Inf. Fusion*, vol. 117, May 2025, Art. no. 102907.

[55] L. Liu, H. Xue, W. Xu, H. Fu, and C. Lu, "Toward real-world category-level articulation pose estimation," *IEEE Trans. Image Process.*, vol. 31, pp. 1072–1083, 2022.

[56] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5099–5108.

[57] W. Zhao, S. Zhang, Z. Guan, W. Zhao, J. Peng, and J. Fan, "Learning deep network for detecting 3D object keypoints and 6D poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Mar. 2020, pp. 14134–14142.

[58] H. Xu, J. Ma, Z. Shao, H. Zhang, J. Jiang, and X. Guo, "SDPNet: A deep network for pan-sharpening with enhanced information representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4120–4134, May 2021.

[59] X. Deng and P. L. Dragotti, "Deep convolutional neural network for multi-modal image restoration and fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3333–3348, Oct. 2021.

[60] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6D object pose and size estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2642–2651.

[61] Y. Weng et al., "CAPTRA: Category-level pose tracking for rigid and articulated objects from point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13189–13198.