

DKPMV: Dense Keypoints Fusion from Multi-View RGB Frames for 6D Pose Estimation of Textureless Objects

Jiahong Chen^{1,2}, Jinghao Wang^{1,2}, Zi Wang^{1,2,*}, Ziwen Wang^{1,2}, Banglei Guan^{1,2} and Qifeng Yu^{1,2}

Abstract—6D pose estimation of textureless objects is valuable for industrial robotic applications, yet remains challenging due to the frequent loss of depth information. Current multi-view methods either rely on depth data or insufficiently exploit multi-view geometric cues, limiting their performance. In this paper, we propose DKPMV, a pipeline that achieves dense keypoint-level fusion using only multi-view RGB images as input. We design a three-stage progressive pose optimization strategy that leverages dense multi-view keypoint geometry information. To enable effective dense keypoint fusion, we enhance the keypoint network with attentional aggregation and symmetry-aware training, improving prediction accuracy and resolving ambiguities on symmetric objects. Extensive experiments on the ROBI dataset demonstrate that DKPMV outperforms state-of-the-art multi-view RGB and RGB-D approaches. The code will be available at <https://github.com/chenjiahongbq/DKPMV>.

I. INTRODUCTION

Textureless objects are commonly encountered in modern industrial scenarios, such as mechanical components and plastic utensils [1]. The lack of distinctive color or texture makes these objects challenging for visual perception, drawing increasing attention in the robotics community [2], as exemplified by benchmarks like the ROBI [3] and XYZ-IBD datasets [4]. Accurate estimation of 6D poses is essential for a wide range of downstream robotic tasks, including autonomous grasping and assembly [5–8].

In recent years, the vast majority of research has focused on addressing the 6D pose estimation problem of textureless objects using depth data [9–11] or RGB-D images [12–14]. While these approaches have significantly improved pose estimation accuracy, they heavily rely on the quality of the depth, which often degrades on specular surfaces [15–17], transparent materials [18, 19], or low-light conditions [20]. Moreover, high-precision depth sensors are costly and operate at low frame rates, limiting their applicability in real-time robotic perception [21]. RGB-based deep learning methods have effectively addressed 6D pose estimation for textureless objects [22]. Nevertheless, the reliance on single-view input in most approaches [23–25] leads to limitations under scale ambiguity, occlusions, and symmetric object [26].

To address the inherent limitations of single RGB-view methods, recent researches have increasingly focused on multi-view frameworks for robust pose estimation of textureless objects. Existing approaches typically fall into two categories: sequential frame processing, as in object-level

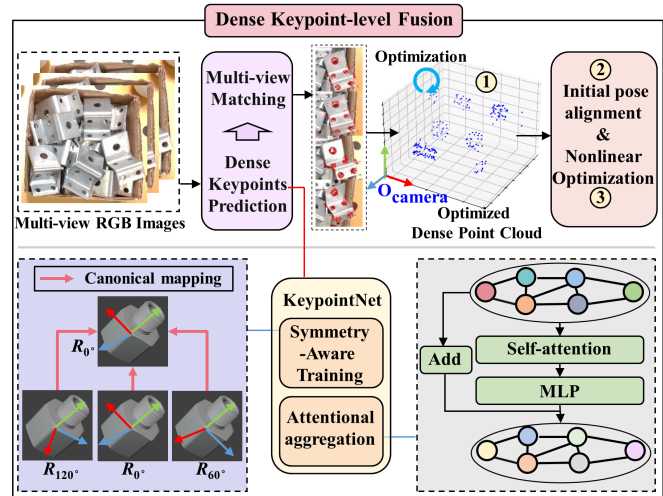


Fig. 1: Brief overview of the DKPMV. A three-stage progressive optimization strategy is proposed, including dense point cloud reconstruction, initial pose alignment, and nonlinear optimization. To enable effective dense keypoint fusion, we enhance the keypoint network with attentional aggregation and symmetry-aware training.

SLAM [27–29], and simultaneous multi-frame input [30–33]. While the former suffers from drift and latency due to frame-by-frame processing, the latter is more suitable for real-time robotic applications [34]. However, most existing multi-frame methods rely on single-view pose initialization followed by global refinement, resulting in pose-level fusion with limited exploitation of multi-view geometric consistency [30, 31, 33]. While MV-3D-KP [32] performs keypoint-level fusion to better utilize multi-view constraints, it integrates only sparse keypoints and remains dependent on depth input, which compromises robustness under occlusion or missing depth, as demonstrated in [33].

Motivated by these challenges, we propose the DKPMV that performs dense keypoint-level multi-view fusion using only RGB images, as illustrated in Fig. 1. To resolve ambiguities in keypoint prediction for symmetric objects, we adopt a symmetry-aware training strategy (SAT) [35]. Moreover, we also design an attentional aggregation module within the keypoint network to capture geometric constraints among dense keypoints and improve prediction accuracy. Furthermore, we propose a three-stage progressive pose estimation strategy that leverages the dense keypoints. These innovative designs enable reliable dense keypoint-level multi-view fusion and significantly improve the robustness of pose estimation in challenging scenarios. Extensive exper-

*Corresponding author: Zi Wang (wangzi16@nudt.edu.cn).

¹College of Aerospace Science and Engineering, National University of Defense Technology; ²Hunan Provincial Key Laboratory of Image Measurement and Vision Navigation, Changsha 410073, China.

iments on the challenging ROBI dataset [3] demonstrate that DKPMV outperforms state-of-the-art RGB and RGB-D based approaches.

The main contributions of this work are as follows:

- We propose the DKPMV, a 6D pose estimation framework that achieves dense keypoint-level multi-view fusion using only multiple RGB images as input.
- We employ SAT to resolve prediction ambiguities and design an attentional aggregation module to enhance keypoint accuracy and ensure effective fusion.
- We design a three-stage progressive pose optimization pipeline that integrates multi-view geometry with dense keypoint cues.

II. RELATED WORK

A. 6D Pose Estimation for Textureless Objects

Traditional methods improve pose estimation accuracy for textureless objects by leveraging depth data [9–11] or RGB-D images [12–14]. Due to depth unreliability on textureless surfaces, RGB-based approaches provide an efficient solution for 6D pose estimation by exploiting visual appearance cues [22]. Template matching methods [36–38] infer poses by comparing images with pre-rendered templates and are suitable for textureless objects. Regression-based [39, 40] methods estimate poses directly from global image features, offering efficiency. However, both approaches exhibit deteriorated accuracy when facing large viewpoint variations, occlusions, or domain shifts, due to the lack of explicit geometrical modeling. Coordinates-based methods [41, 42] aggregate pixel-wise object coordinates via RANSAC [43], ensuring robustness in occlusion. Nonetheless, per-pixel coordinate regression takes vast computational burden, preventing their applications in multi-view and real-time scenarios.

Semantic keypoint-based methods [41, 44, 45] detect a set of 2D semantic keypoints and establish 2D–3D correspondences with predefined points on CAD models, from which the 6D pose is estimated via Perspective-n-Point (PnP). This formulation transforms pose estimation into a structured and interpretable task, enabling consistent exploitation of multi-view geometric constraints [32]. Compared to sparse keypoints, dense keypoint-based methods [24, 25, 46, 47] capture richer scene information and exhibit greater robustness to occlusion. Furthermore, integrating graph neural networks (GNN) enables geometric interactions among keypoints, further enhancing prediction accuracy [25]. However, most existing methods utilize only single-view keypoint predictions without multi-view fusion, leading to depth loss and scale ambiguity. Moreover, they lack dedicated mechanisms to resolve keypoint ambiguities arising from object rotational symmetries [35].

B. Multi-View 6D Object Pose Estimation from RGB Images

Multi-view 6D pose estimation has been shown to be effective in addressing the depth and scale ambiguities inherent in single-view settings. According to the input, these approaches can be categorized into sequential-frame based ones (e.g., object-level SLAM [27–29]) and simultaneous

multi-frame based ones [30–33]. Sequential frame methods are susceptible to cumulative drift and processing latency, whereas simultaneous multi-frame method process multiple views in parallel, delivering the responsiveness required for real-time robotic tasks [34]. However, most existing multi-frame methods perform pose-level fusion either through global optimization of individually estimated single-view poses [30, 31] or by decoupling translation and rotation estimation through multi-view center keypoint fusion [33]. Such strategies are prone to accumulated errors and limited in fully exploiting cross-view consistency. Keypoint-based methods offer a more interpretable and geometric formulation, enabling more geometry-aware multi-view integration [32]. Nevertheless, current multi-view keypoint approaches are typically limited to sparse keypoints while requiring depth input, suffering from occlusion and degraded depth quality [32].

To address these challenges, we achieve dense keypoint fusion and introduce a three-stage progressive optimization pipeline that directly estimates 6D pose from RGB images, enabling robust and accurate keypoint-level multi-view fusion.

III. METHOD

Given a set of RGB images $\{\mathcal{I}_j\}_{j=1}^V$ and a collection of object instances, the objective is to estimate the 6D pose of each object, defined by its rotation $\mathbf{R} \in \text{SO}(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$, w.r.t. a global coordinate frame. We assume that the relative camera poses between views and the 3D geometry of each object (i.e., CAD models) are known. A set of N 3D keypoints $\mathbf{P}_i \in \mathbb{R}^3$ is uniformly sampled from each CAD model using farthest point sampling (FPS).

As illustrated in Fig. 2, the proposed method consists of several stages, with the key innovation lying in the dense keypoint-level fusion across multiple views and a dedicated three-stage progressive pose optimization pipeline. We begin with 2D bounding box detection for each object instance in every image using an off-the-shelf YOLOv11. These bounding boxes are used to crop and resize the input images. The patches are then processed by the KeypointNet-SAT network to generate dense keypoint predictions. A multi-view matching module is then employed to establish consistent keypoint correspondences across views for each object instance. Finally, the matched dense keypoints are passed to the pose estimation module to recover the final 6D pose.

A. Single-view Dense Keypoints Estimation

The dense keypoint prediction network is built upon CheckerPose [25]. It integrates a GNN to capture geometric relations among dense keypoints and a CNN to extract visual features from RGB inputs. In addition to 2D keypoint coordinates, the network outputs a binary visibility code \mathbf{b}_v for each keypoint, indicating whether it lies within the object’s region of interest (ROI). This visibility cue is used to filter out unreliable keypoints and improve the overall prediction quality.

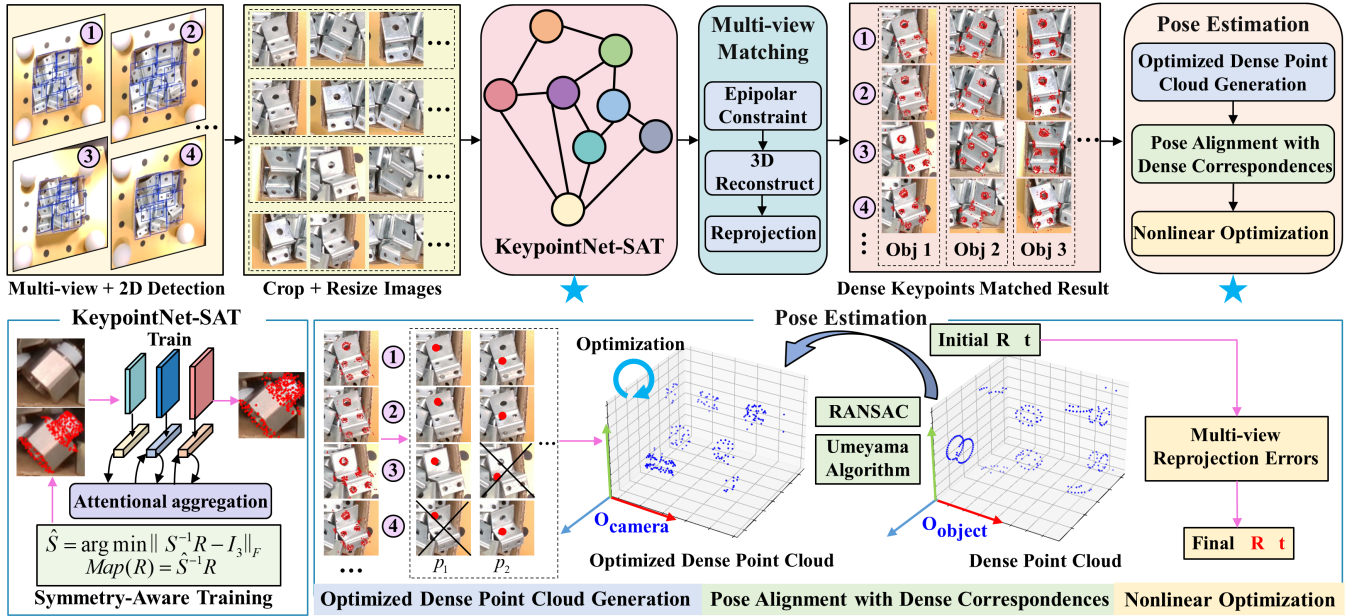


Fig. 2: Structural illustration of the DKPMV. Given multi-view RGB images and corresponding 2D bounding boxes, we perform dense keypoint detection for each object in each view. The detected keypoints are matched across views and subsequently fed into a three-stage progressive pose estimation module.

To ensure effective keypoint fusion and stable network training, we adopt the SAT [35] to resolve keypoint prediction ambiguities on symmetric objects by transforming all equivalent poses into a unified canonical representation. Specifically, given a proper symmetry group $\mathcal{M}(O)$ for object O , which defines the set of transformations that leave the object appearance unchanged:

$$\mathcal{M} = \{\mathbf{m} \in \text{SE}(3) \mid \forall \mathbf{p} \in \text{SE}(3), \mathcal{R}(O, \mathbf{p}) = \mathcal{R}(O, \mathbf{m} \cdot \mathbf{p})\} \quad (1)$$

where $\mathcal{R}(O, \mathbf{p})$ is the image of Object O under pose \mathbf{p} (ignoring lighting effects), \mathbf{m} is a rigid motion related to the symmetry, and $\mathbf{m} \cdot \mathbf{p}$ is the pose after applying motion \mathbf{m} . A corresponding operator $\text{Map}(\cdot)$ on $\text{SO}(3)$ is employed to map symmetric equivalent poses to a consistent canonical form:

$$\text{Map}(\mathbf{R}) = \hat{S}^{-1} \mathbf{R}, \quad \forall \mathbf{R} \in \text{SO}(3) \quad (2)$$

where $\text{Map}(\cdot)$ ensures consistent keypoint supervision across all symmetric-equivalent poses during training. And \hat{S} is the optimal rotation matrix that best aligns the input \mathbf{R} :

$$\hat{S} = \arg \min_{S \in \mathcal{M}(O)} \|S^{-1} \mathbf{R} - I_3\|_F \quad (3)$$

By this means, we significantly improve pose estimation performance for rotationally symmetric objects, as discussed in Section IV.

Furthermore, inspired by the SuperGlue [48], we replace the max-based node update scheme used in the EdgeConv [49] module of CheckerPose with an attentional aggregation (Att), enabling more effective modeling of geometric relationships among dense keypoints. Specifically, each keypoint \mathbf{x}_i aggregates features from its $k = 20$ nearest

neighbors $\{\mathbf{x}_j\}_{j \in \mathcal{N}(i)}$, as defined in CheckerPose [25]. The message $\mathbf{m}_{\mathcal{E} \rightarrow i}$ is computed by attentional aggregation over all connected keypoints $\{j : (i, j) \in \mathcal{E}\}$. Given a query \mathbf{q}_i derived from the keypoint feature \mathbf{x}_i^ℓ at the ℓ -th layer, and key and value $\mathbf{k}_j, \mathbf{v}_j$ derived from each neighboring keypoint feature \mathbf{x}_j^ℓ , the message is computed as a weighted average of the values:

$$\mathbf{m}_{\mathcal{E} \rightarrow i} = \sum_{j: (i, j) \in \mathcal{E}} \alpha_{ij} \mathbf{v}_j, \quad \alpha_{ij} = \text{Softmax}_j(\mathbf{q}_i^\top \mathbf{k}_j) \quad (4)$$

and the updated keypoint feature at the $\ell+1$ -th layer, denoted as $\mathbf{x}_i^{\ell+1}$, is computed as:

$$\mathbf{x}_i^{\ell+1} = \mathbf{x}_i^\ell + \text{MLP}(\text{Concat}[\mathbf{m}_{\mathcal{E} \rightarrow i}, \mathbf{x}_i^\ell]) \quad (5)$$

This node update strategy enhances keypoint localization accuracy, thereby improving the final pose estimation performance, as discussed in Section IV.

B. Estimating Object Pose Using Matched Keypoints

We accomplish multi-view keypoint matching by first pairing the two views with the closest poses, followed by keypoint reconstruction and projection of the reconstructed 3D points onto the remaining views. Implementation details of keypoint matching can be found in Appendix.

After matching instance-level keypoints across multiple views, we obtain a set of 2D predictions $\tilde{\mathbf{p}}_i^j \in \mathbb{R}^2$ for each instance, where i indexes the keypoint and j denotes the view. The theoretical projection of each keypoint $\mathbf{p}_i^j \in \mathbb{R}^2$ can be computed as:

$$\lambda_i^j \begin{bmatrix} \mathbf{p}_i^j \\ 1 \end{bmatrix} = \mathbf{K}_j (\mathbf{R}_j (\mathbf{R} \mathbf{P}_i + \mathbf{t}) + \mathbf{t}_j) \quad (6)$$

where λ_i^j denotes the depth value determined by the known camera intrinsics \mathbf{K}_j and extrinsics $(\mathbf{R}_j, \mathbf{t}_j)$, defined relative to the first camera. Based on the Gaussian assumption and Bayes' theorem, the optimal pose estimation can be formulated as:

$$\mathbf{R}^*, \mathbf{t}^* = \arg \min_{\mathbf{R}, \mathbf{t}} \sum_{i=1}^{\tilde{N}} \sum_{j=1}^V \left\| \tilde{\mathbf{p}}_i^j - \mathbf{p}_i^j \right\|_2 \quad (7)$$

where \tilde{N} denotes the number of valid points that are simultaneously visible across all views, i.e., those with $\mathbf{b}_v = 1$, and thus satisfies $\tilde{N} \leq N$. Detailed derivation for Eq. 7 is referred to Appendix.

Based on Eq. 6, \tilde{N} intermediate variables $\tilde{\mathbf{P}}_i$ are introduced, which represent the reconstructed 3D keypoints and satisfies:

$$\tilde{\mathbf{P}}_i = \mathbf{R}\mathbf{P}_i + \mathbf{t}, \quad i \in \{1, \dots, \tilde{N}\} \quad (8)$$

Then, the optimization in Eq. 7 can be equivalently decomposed into two subproblems:

$$\tilde{\mathbf{P}}_i^* = \arg \min_{\tilde{\mathbf{P}}_i} \sum_{j=1}^V \left\| \tilde{\mathbf{p}}_i^j - \mathbf{p}_i^j \right\|_2, \quad i = \{1, \dots, \tilde{N}\} \quad (9)$$

$$\mathbf{R}^*, \mathbf{t}^* = \arg \min_{\mathbf{R}, \mathbf{t}} \sum_{i=1}^{\tilde{N}} \left\| \mathbf{R}\mathbf{P}_i + \mathbf{t} - \tilde{\mathbf{P}}_i^* \right\|_2 \quad (10)$$

Motivated by Eqs. 7, 9, and 10, we design a three-stage progressive optimization strategy for pose estimation.

Optimized dense point cloud generation. We perform multiple iterations of RANSAC [43], where a threshold τ_1 is applied, and select the candidate set of $\tilde{\mathbf{p}}_i^j$ with the best combined score of inlier count and reprojection error. For each 3D point $\tilde{\mathbf{P}}_i$, we perform multi-view reconstruction using the valid view set \mathcal{V}_i via singular value decomposition (SVD), resulting in the globally optimal estimate $\tilde{\mathbf{P}}_i^*$, as defined in Eq. 9.

Pose alignment with dense correspondence. We estimate the initial pose $(\mathbf{R}^*, \mathbf{t}^*)$ by aligning the reconstructed dense point cloud $\mathcal{P} = \{\tilde{\mathbf{P}}_i^* \in \mathbb{R}^3\}_{i=1}^{\tilde{N}}$ with the reference 3D keypoints $\mathcal{P}_o = \{\mathbf{P}_i \in \mathbb{R}^3\}_{i=1}^{\tilde{N}}$. Specifically, we solve Eq. 10 using the Umeyama algorithm [50], combined with RANSAC [43], where a threshold τ_2 is imposed to suppress outliers by selecting the solution with the maximum number of inliers from \mathcal{P} .

Nonlinear optimization. Given the initial pose estimates $(\mathbf{R}^*, \mathbf{t}^*)$, we further refine the solution by performing global nonlinear optimization (NO), leveraging multi-view keypoint predictions.

To mitigate the influence of outliers, only valid 2D keypoints are included in the optimization, and the cost function is defined as:

$$\arg \min_{\mathbf{R}^*, \mathbf{t}^*} \sum_{i=1}^{N_{\text{inlier}}^{3d}} \sum_{j \in \mathcal{V}_i} \sigma \left(\left\| \mathbf{p}_i^j - \tilde{\mathbf{p}}_i^j \right\|_2 \right) \quad (11)$$

where $\sigma(\cdot)$ denotes a robust loss function to reduce the impact of heavy-tailed noise.

Please refer to Appendix for more detailed information on the three-stage progressive pose estimation method.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

We evaluate our method on the ROBI dataset [3], which provides multi-view RGB images and ground-truth 6D poses for 7 textureless industrial objects. The dataset comprises images acquired using a high-end Ensenso and a low-cost RealSense sensor. To balance speed and accuracy, we adopt 256 keypoints as our primary configuration. Following [17, 33], we train our keypoint network with only synthetic images provided by the ROBI dataset. Evaluation is conducted on the real-world test sets captured by both Ensenso and RealSense sensors for all objects.

Following [33], we evaluate pose estimation performance using the common average recall (AR) under two adopted metrics: the average distance (ADD) and the 5-mm/10-degree (5mm, 10°) metric. For symmetric objects, we report the minimum error over all equivalent ground-truth poses under symmetry, based on either the ADD or the (5mm, 10°) metric. In our evaluation, a ground-truth pose is considered valid only if its visibility score is larger than 75%, consistent with the rule in [33].

B. Comparison with State-of-the-Art Methods

Following [33], we conduct experiments on the ROBI dataset using varying numbers of views (4 and 8). We compare our method with SOTA multi-view approaches, including both RGB and RGB-D methods: CosyPose [30], MV-3D-KP [32], and Jun's method [33]. Tables I and II report the pose estimation results for seven highly reflective textureless objects on the Ensenso and RealSense test sets, respectively.

On the **Ensenso** test set, our method achieves the highest AR under the 4-view setting, reaching 93.4% and 89.8% on the ADD and (5mm, 10°) metrics respectively, surpassing the RGB-D-based MV-3D-KP [32] method. With 8 input views, the performance of our method is further bolstered, yielding 96.8% and 92.6% across the two metrics, respectively. Notably, DKPMV consistently surpasses the MV-3D-KP [32] baseline under these configurations.

On the **RealSense** test set, the performance of RGB-D methods drops significantly due to the reduced quality of depth data, whereas our method demonstrates a substantial advantage. Its performance significantly surpasses both RGB and RGB-D methods, especially under the (5mm, 10°) metric, where it outperforms the second-best method by a large margin of 13.0% and 13.5% in the 4-view and 8-view settings respectively.

C. Ablation Studies

Extensive ablation studies are conducted to evaluate each component, including SAT, attentional aggregation (Att) node update, keypoint density, and pose estimation method. A stricter metric (2mm, 3°) is introduced for more intuitive comparison of pose estimation performance.

Objects		4 Views					8 Views				
		CosyPose + LINE2D		MV-3D-KP	Jun's	Ours	CosyPose + LINE2D		MV-3D-KP	Jun's	Ours
Input Modality		RGB	RGBD	RGBD	RGB	RGB	RGB	RGBD	RGBD	RGB	RGB
Tube*	ADD	32.5	74.8	94.0	89.4	95.4	50.3	91.4	96.0	94.0	98.0
	Fitting† (5, 10)	45.7	76.2	95.4	88.1	88.1	71.5	94.7	96.0	92.0	98.0
Chrome*	ADD	55.7	73.0	90.8	86.7	97.6	70.1	88.5	91.9	93.7	98.2
	Screw† (5, 10)	63.2	78.2	88.5	85.1	94.1	78.7	90.2	90.8	90.2	97.6
Eye Bolt*†	ADD	35.1	85.1	93.2	93.2	87.8	79.7	93.2	94.6	94.6	91.9
	(5, 10)	27.0	78.4	87.8	67.6	81.1	64.9	83.8	85.1	75.8	85.1
Gear*	ADD	25.9	80.2	85.2	91.4	98.8	43.2	88.9	93.8	97.5	100.0
	(5, 10)	29.6	79.0	85.2	85.2	98.8	45.7	92.6	91.4	93.8	97.5
Zigzag	ADD	65.5	87.9	96.6	94.8	100.0	77.6	96.6	96.6	98.3	100.0
	(5, 10)	37.9	75.9	93.1	89.7	100.0	63.8	93.1	96.6	93.1	100.0
DIN Connector	ADD	15.6	57.8	90.6	69.5	92.1	24.2	64.1	93.8	73.4	94.4
	(5, 10)	12.5	46.1	84.4	53.9	90.5	23.4	51.6	93.0	59.4	91.3
D-Sub* Connector†	ADD	9.9	55.3	92.5	79.5	82.0	15.5	63.3	95.7	84.5	95.0
	(5, 10)	11.2	39.1	83.2	47.2	76.4	11.2	41.6	91.3	55.9	78.9
ALL	ADD	34.3	73.4	91.8	86.4	93.4	51.5	83.7	94.6	90.9	96.8
	(5, 10)	32.4	67.6	88.2	73.8	89.8	51.3	78.2	92.0	80.0	92.6

Table I: The AR (%) of 6D object pose estimation on **Ensenso** test set, evaluated with the metrics of ADD and (5mm, 10°). There are a total of nine scenes for each object. (†) indicates the use of the SAT strategy, and (*) denotes symmetric objects, following [33].

Objects		4 Views					8 Views				
		CosyPose + LINE2D		MV-3D-KP	Jun's	Ours	CosyPose + LINE2D		MV-3D-KP	Jun's	Ours
Input Modality		RGB	RGBD	RGBD	RGB	RGB	RGB	RGBD	RGBD	RGB	RGB
Tube*	ADD	27.9	70.6	79.4	86.8	91.2	69.1	83.9	86.8	85.3	92.6
	Fitting† (5, 10)	48.5	72.1	76.5	79.4	77.9	82.3	85.3	83.8	91.2	92.6
Chrome*	ADD	58.6	68.5	87.1	92.9	94.3	77.1	80.0	92.9	92.9	94.3
	Screw† (5, 10)	64.3	82.9	84.3	77.1	91.4	85.7	94.3	91.4	87.1	92.9
Eye Bolt*†	ADD	58.8	76.5	85.3	94.1	85.3	73.5	94.1	85.3	94.1	91.2
	(5, 10)	41.2	67.6	82.3	55.9	76.5	61.8	91.2	85.3	76.5	88.2
Gear*	ADD	36.1	83.3	88.8	94.4	97.2	55.6	97.2	91.7	97.2	97.2
	(5, 10)	38.9	77.8	61.1	86.1	94.4	58.3	94.4	63.9	88.9	94.4
Zigzag	ADD	42.9	78.6	96.4	89.3	92.9	71.4	92.9	92.9	96.4	92.9
	(5, 10)	21.4	71.4	96.4	85.7	92.9	64.3	92.9	92.9	92.9	92.9
DIN Connector	ADD	3.8	36.5	82.7	51.9	98.1	15.1	51.9	84.6	82.7	98.1
	(5, 10)	1.9	30.8	75.0	32.7	96.2	9.6	34.6	84.6	57.7	96.2
D-Sub* Connector†	ADD	6.9	40.3	88.9	70.8	97.2	9.7	45.8	90.3	81.9	98.6
	(5, 10)	6.9	18.1	44.4	31.9	81.9	8.3	33.3	55.6	43.1	94.4
ALL	ADD	33.6	64.9	86.7	82.9	93.7	53.1	78.0	89.2	90.1	95.0
	(5, 10)	31.9	60.1	74.3	64.1	87.3	52.9	75.1	79.6	76.8	93.1

Table II: The AR (%) of 6D object pose estimation on **RealSense** test set, evaluated with the metrics of ADD and (5mm, 10°). There are a total of four scenes for each object. (†) indicates the use of the SAT strategy, and (*) denotes symmetric objects, following [33].

SAT strategy. As illustrated in Fig. 3 (a), SAT strategy produces well-structured keypoint distributions with preserved geometric constraints, while its absence results in scattered and inconsistent predictions. Furthermore, the results presented in Table III confirm that incorporating the SAT strategy significantly improves pose estimation performance. The gains become more pronounced as the number of keypoints increases, since more accurate keypoint predictions enable more effective multi-view fusion. Notably, the largest performance gain is achieved when the number of keypoints increases to 512, yielding improvements of 43.1% under the (2mm, 3°) metrics. This highlights the importance of the SAT strategy for robust multi-view dense keypoint fusion.

Attentional aggregation. As illustrated in Fig. 3 (b), the

Point num	(5, 10)			(2, 3)		
	w/o SAT	w SAT	Δ	w/o SAT	w SAT	Δ
32	48.3	81.5	33.2	21.2	51.0	29.8
64	48.3	85.4	37.1	25.2	62.9	37.7
128	54.3	88.7	34.4	24.5	62.3	37.8
256	57.6	88.1	30.5	28.5	66.9	38.4
512	62.9	90.7	27.8	27.8	70.9	43.1

Table III: Comparison of pose estimation performance with and without the SAT strategy on the Tube-Fitting object from the Ensenso test set.

original CheckerPose [25] exhibits large deviations in the prediction of certain keypoints. In contrast, our keypoint network with Att node update more effectively captures the

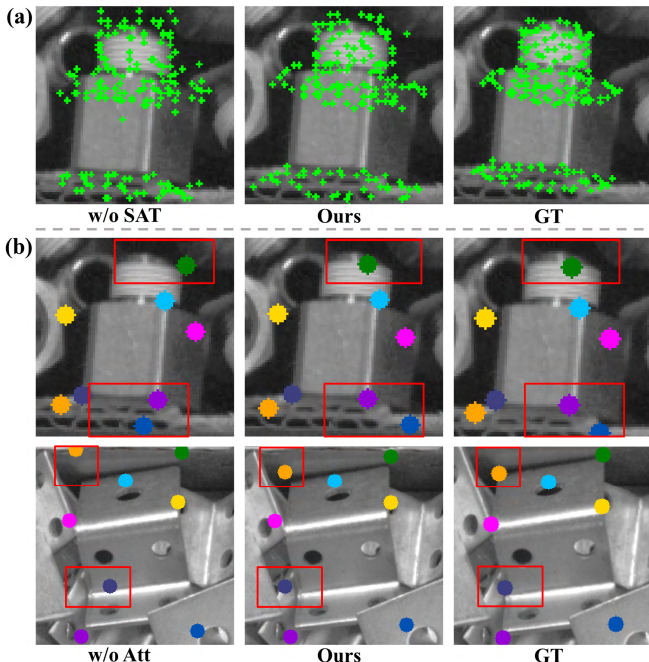


Fig. 3: (a) Comparison of overall keypoint distributions with and without the SAT strategy. (b) Comparison of local keypoint localization accuracy with and without the Att. Please refer to the appendix for additional results.

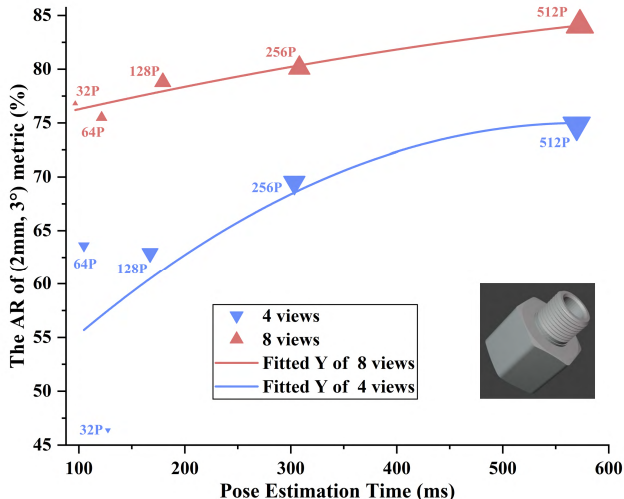


Fig. 4: Pose estimation accuracy and runtime (including keypoint prediction and pose estimation) on the Tube-Fitting object under varying numbers of keypoints.

geometric constraints among neighboring dense keypoints, leading to more geometry-aware and accurate keypoint localization. Meanwhile, the results presented in Table IV demonstrate that our Att node update significantly improves pose estimation performance on most objects. Notably, under the $(2\text{mm}, 3^\circ)$ metric, our method achieves improvements of 4.6% and 2.1% in the 4-view and 8-view settings respectively, confirming the effectiveness of our design.

Number of keypoints. Fig. 4 shows the pose accuracy under the strict $(2\text{mm}, 3^\circ)$ metric and the runtime, tested

object	4 views				8 views			
	(5, 10)		(2, 3)		(5, 10)		(2, 3)	
	Ori	Att	Ori	Att	Ori	Att	Ori	Att
Tube-Fitting	86.8	88.1	64.9	66.9	98.0	98.0	82.1	80.8
Chrome-Screw	92.3	94.1	43.8	50.3	94.7	97.6	55.6	54.4
Eye-Bolt	82.4	81.1	40.5	41.9	86.5	85.1	52.7	55.4
Gear	92.6	98.8	76.5	81.5	96.3	97.5	84.0	84.0
Zigzag	100.0	100.0	93.1	94.8	100.0	100.0	91.4	96.6
DIN-Connector	88.1	90.5	46.0	61.1	90.5	91.3	59.5	68.3
DSub-Connector	73.9	76.4	17.4	18.0	81.3	78.9	23.6	23.6
ALL	88.0	89.8	54.6	59.2	92.5	92.6	64.1	66.2

Table IV: Comparison of pose estimation performance between our Att node and the original (Ori) on the Ensenso test set.

on an Intel® Core™ Ultra 9 285K CPU and Nvidia 5090 GPU. As expected, pose estimation accuracy exhibits a generally increasing trend with more keypoints in both 4-view and 8-view settings, validating the effectiveness of the DKPMV. The observed variations in accuracy and runtime at 32, 64, and 128 keypoints stem from insufficient keypoint density, where prediction noise disrupts the effectiveness of keypoint fusion and increases the required number of RANSAC iterations during optimization. This further confirms the importance of using an appropriate number of dense keypoints for effective fusion.

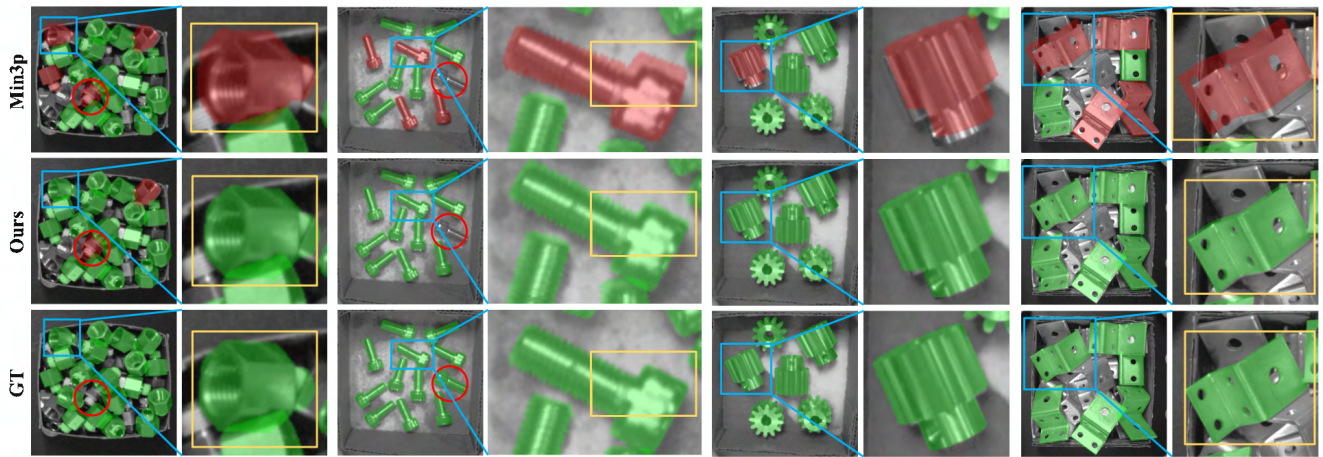
Pose estimation method. As shown in Table V, we compare our pose estimation method with the minimal three-point solver (Min3P) [51] with RANSAC [43]. Our method consistently outperforms Min3P [51] across both metrics, with especially notable gains under the strict $(2\text{mm}, 3^\circ)$ metric, achieving improvements of 12.0% and 9.2% in the 4-view and 8-view settings, respectively. These results demonstrate the effectiveness of our pose estimation pipeline. Meanwhile, NO improves the performance of both methods, demonstrating its effectiveness. Although Min3P requires only three points to compute the pose, it fails to fully leverage the information from multi-view keypoints, resulting in lower pose estimation accuracy. In contrast, our method achieves high-precision pose estimation by performing only the first two stages of optimization. Fig. 5 illustrates the pose estimation visualizations for both our method and Min3P.

	Min3p	Ours (w/o NO)	NO	4 views		8 views	
				(5, 10)	(2, 3)	(5, 10)	(2, 3)
1	✓			65.3	13.4	80.6	23.8
2	✓		✓	88.5	47.2	90.1	57.0
3		✓	✓	89.8	58.3	92.6	66.2
4		✓	✓	89.8	59.2	92.6	66.2

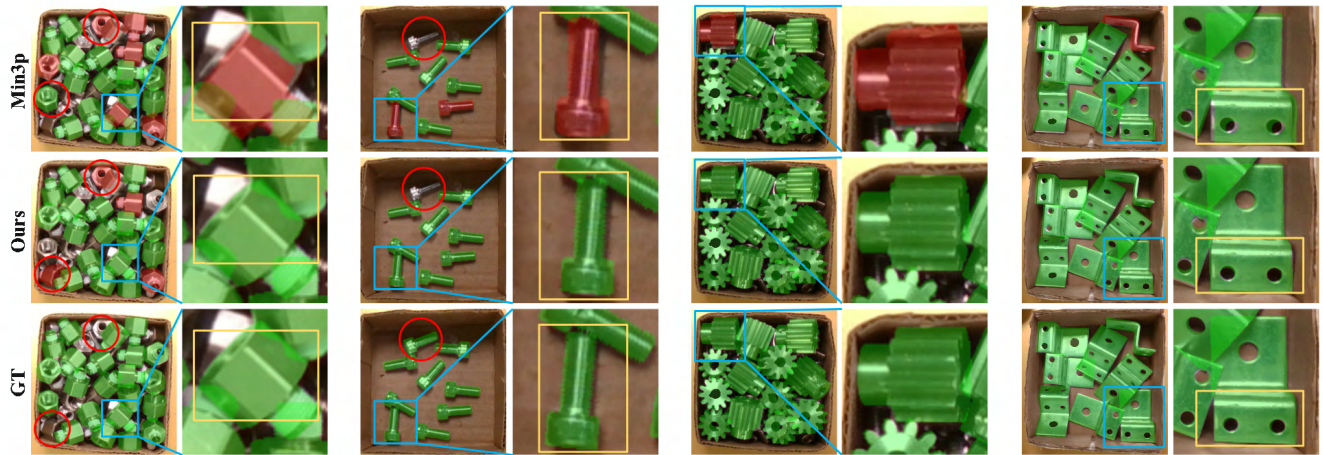
Table V: Comparing our method with Min3P [51] on Ensenso test set for all objects. We adopt RANSAC followed by nonlinear optimization (NO) for Min3P to suppress outliers and refine the estimated poses, thereby ensuring a fair comparison.

V. CONCLUSIONS

This paper proposes DKPMV, a novel multi-view RGB pipeline for 6D pose estimation of textureless objects, which achieves dense keypoint-level fusion without relying on



(a) Object pose estimation results on Ensenso test set



(b) Object pose estimation results on RealSense test set

Fig. 5: Visualizations of the 6D pose estimation results on (a) Ensenso and (b) RealSense test sets using 4-view input. The green or red renderings indicate whether the predicted poses satisfy the $(5\text{mm}, 10^\circ)$ metric, while red circles highlight missed or falsely detected object. The enlarged views further highlight the fine-grained differences in pose estimation accuracy. More examples are in the appendix.

depth input. Leveraging three-stage progressive optimization, Att, and SAT, our method effectively captures multi-view geometric cues. Extensive evaluations on the ROBI dataset confirm that our approach sets a new SOTA among RGB-based methods and even surpasses RGB-D methods in most scenarios. Future work will focus on modeling dense key-point uncertainty to enhance multi-view geometric fusion.

Acknowledgment. We gratefully acknowledge Dr. Yang for his technical guidance and assistance with the ROBI [3] dataset, which significantly facilitated our research. This work was supported in part by the National Natural Science Foundation of China under Grants 12302252 and 12372189, the Research Program of National University of Defense Technology under Grant ZK24-31, and the Hunan Provincial Natural Science Foundation of China under Grant 2025JJ60054.

REFERENCES

- [1] Manuel Stoiber, Martin Sundermeyer, and Rudolph Triebel. “Iterative corresponding geometry: Fusing region and depth for highly efficient 3d tracking of textureless objects”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 6855–6865.
- [2] Li Jin et al. “Online hand-eye calibration with decoupling by 3d textureless object tracking”. In: *Proceedings of 2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11453–11460.
- [3] Jun Yang et al. “Robi: A multi-view dataset for reflective objects in robotic bin-picking”. In: *Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 9788–9795.
- [4] Junwen Huang et al. “XYZ-IBD: High-precision Bin-picking Dataset for Object 6D Pose Estimation Capturing Real-world Industrial Complexity”. In: *arXiv preprint arXiv:2506.00599* (2025).
- [5] Jun Shi et al. “Asgrasp: Generalizable transparent object reconstruction and 6-dof grasp detection from rgb-d active stereo camera”. In: *Proceedings of 2024 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2024, pp. 5441–5447.
- [6] Liming Zhang et al. “Instance-level 6D pose estimation based on multi-task parameter sharing for robotic grasping”. In: *Scientific Reports* 14.1 (2024), p. 7801.
- [7] Boyoung Kim and Junhong Min. “Sim-to-real Object Pose Estimation for Random Bin Picking”. In: *Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 10749–10756.
- [8] Jianqiu Chen et al. “ZeroBP: Learning Position-Aware Correspondence for Zero-shot 6D Pose Estimation in Bin-Picking”. In: *arXiv preprint arXiv:2502.01004* (2025).
- [9] Ge Gao et al. “6D Object Pose Regression via Supervised Learning on Point Clouds”. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. 2020, pp. 3643–3649.

- [10] Ge Gao et al. "Cloudaae: Learning 6d object pose regression with on-line data synthesis on point clouds". In: *Proceedings of 2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11081–11087.
- [11] Dingding Cai, Janne Heikkilä, and Esa Rahtu. "Ove6d: Object viewpoint encoding for depth-based 6d object pose estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 6803–6813.
- [12] Chen Wang et al. "Densefusion: 6d object pose estimation by iterative dense fusion". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 3343–3352.
- [13] Yisheng He et al. "PVN3D: A Deep Point-Wise 3D Keypoints Voting Network for 6DoF Pose Estimation". In: *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 11629–11638.
- [14] Lounes Saadi et al. "Optimizing rgb-d fusion for accurate 6dof pose estimation". In: *IEEE Robotics and Automation Letters* 6.2 (2021), pp. 2413–2420.
- [15] Peiyuan Ni et al. "Reasoning and Learning a Perceptual Metric for Self-Training of Reflective Objects in Bin-Picking with a Low-cost Camera". In: *IEEE Robotics and Automation Letters* (2025), pp. 1–8.
- [16] Jun Yang and Steven L Waslander. "Next-best-view prediction for active stereo cameras and highly reflective objects". In: *Proceedings of 2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 3684–3690.
- [17] Jun Yang et al. "6D Pose Estimation for Textureless Objects on RGB Frames using Multi-View Optimization". In: *Proceedings of 2023 IEEE International Conference on Robotics and Automation (ICRA)*. 2023, pp. 2905–2912.
- [18] Huijie Zhang et al. "Transnet: Category-level transparent object pose estimation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 148–164.
- [19] Xingyu Liu et al. "Keypose: Multi-view 3d labeling and keypoint estimation for transparent objects". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 11602–11610.
- [20] Chun-Yu Chai, Yu-Po Wu, and Shiao-Li Tsao. "Deep Depth Fusion for Black, Transparent, Reflective and Texture-Less Objects". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. 2020, pp. 6766–6772.
- [21] Konrad P Cop et al. "New metrics for industrial depth sensors evaluation for precise robotic applications". In: *Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 5350–5356.
- [22] Jian Liu et al. "Deep learning-based object pose estimation: A comprehensive survey". In: *arXiv preprint arXiv:2405.07801* (2024).
- [23] Kilian Kleeberger and Marco F Huber. "Single shot 6d object pose estimation". In: *Proceedings of 2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6239–6245.
- [24] Yongzhi Su et al. "Zerapose: Coarse to fine surface encoding for 6dof object pose estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 6738–6748.
- [25] Ruyi Lian and Haibin Ling. "Checkerpose: Progressive dense keypoint localization for object pose estimation with graph neural network". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 14022–14033.
- [26] Dominik Bauer et al. "Challenges for monocular 6d object pose estimation in robotics". In: *IEEE Transactions on Robotics* (2024).
- [27] Jiahui Fu et al. "A multi-hypothesis approach to pose ambiguity in object-based slam". In: *Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 7639–7646.
- [28] Nathaniel Merrill et al. "Symmetry and uncertainty-aware object slam for 6dof object pose estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 14901–14910.
- [29] Kevis-Kokitsi Maninis et al. "Vid2cad: Cad model alignment using multi-view constraints from videos". In: *IEEE transactions on Pattern Analysis and Machine Intelligence* 45.1 (2022), pp. 1320–1327.
- [30] Yann Labbé et al. "Cosypose: Consistent multi-view multi-object 6d pose estimation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 574–591.
- [31] Ivan Shugurov et al. "Multi-view object pose refinement with differentiable renderer". In: *IEEE Robotics and Automation Letters* 6.2 (2021), pp. 2579–2586.
- [32] Alan Li and Angela P. Schoellig. "Multi-View Keypoints for Reliable 6D Object Pose Estimation". In: *Proceedings of 2023 IEEE International Conference on Robotics and Automation (ICRA)*. 2023, pp. 6988–6994.
- [33] Jun Yang et al. "Active 6D Pose Estimation for Textureless Objects using Multi-View RGB Frames". In: *arXiv preprint arXiv:2503.03726* (2025).
- [34] Rohan Choudhury, Kris M Kitani, and László A Jeni. "Tempo: Efficient multi-view pose estimation, tracking, and forecasting". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 14750–14760.
- [35] Giorgia Pitteri et al. "On object symmetries and 6d pose estimation from images". In: *Proceedings of 2019 International Conference on 3D Vision (3DV)*. IEEE, 2019, pp. 614–622.
- [36] Xinke Deng et al. "PoseRBPF: A Rao-Blackwellized particle filter for 6-D object pose tracking". In: *IEEE Transactions on Robotics* 37.5 (2021), pp. 1328–1342.
- [37] Martin Sundermeyer et al. "Implicit 3d orientation learning for 6d object detection from rgb images". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 699–715.
- [38] Zhigang Li and Xiangyang Ji. "Pose-guided auto-encoder and feature-based refinement for 6-dof object pose regression". In: *Proceedings of 2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 8397–8403.
- [39] Yinlin Hu et al. "Single-stage 6d object pose estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 2930–2939.
- [40] Gu Wang et al. "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 16611–16621.
- [41] Li Xu et al. "6d-diff: A keypoint diffusion framework for 6d object pose estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 9676–9686.
- [42] Xingyu Liu, Shun Iwase, and Kris M Kitani. "Kdfnet: Learning keypoint distance field for 6d object pose estimation". In: *Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4631–4638.
- [43] Martin A Fischler and Robert C Bolles. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography". In: *Communications of the ACM* 24.6 (1981), pp. 381–395.
- [44] Georgios Pavlakos et al. "6-dof object pose from semantic keypoints". In: *Proceedings of 2017 IEEE international Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 2011–2018.
- [45] Penglei Liu et al. "MFPN-6D: Real-time one-stage pose estimation of objects on RGB images". In: *Proceedings of 2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 12939–12945.
- [46] Hansheng Chen et al. "Epro-ppn: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 2781–2790.
- [47] Yan Di et al. "So-pose: Exploiting self-occlusion for direct 6d pose estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 12396–12405.
- [48] Paul-Edouard Sarlin et al. "Superglue: Learning feature matching with graph neural networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 4938–4947.
- [49] Yue Wang et al. "Dynamic graph cnn for learning on point clouds". In: *ACM Transactions on Graphics (TOG)* 38.5 (2019), pp. 1–12.
- [50] Shinji Umeyama. "Least-squares estimation of transformation parameters between two point patterns". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13.4 (2002), pp. 376–380.
- [51] Gim Hee Lee et al. "Minimal solutions for the multi-camera pose estimation problem". In: *The International Journal of Robotics Research* 34.7 (2015), pp. 837–848.