

RefPose: Leveraging Reference Geometric Correspondences for Accurate 6D Pose Estimation of Unseen Objects

Jaeguk Kim¹ Jaewoo Park¹ Keuntek Lee¹ Nam Ik Cho^{1,2}

¹Department of ECE, INMC, Seoul National University, Korea

²IPAI, Seoul National University, Korea

{jaeguk, bjw0611, leekt000, nicho}@snu.ac.kr

Abstract

Estimating the 6D pose of unseen objects from monocular RGB images remains a challenging problem, especially due to the lack of prior object-specific knowledge. To tackle this issue, we propose RefPose, an innovative approach to object pose estimation that leverages a reference image and geometric correspondence as guidance. RefPose first predicts an initial pose by using object templates to render the reference image and establish the geometric correspondence needed for the refinement stage. During the refinement stage, RefPose estimates the geometric correspondence of the query based on the generated references and iteratively refines the pose through a render-and-compare approach. To enhance this estimation, we introduce a correlation volume-guided attention mechanism that effectively captures correlations between the query and reference images. Unlike traditional methods that depend on pre-defined object models, RefPose dynamically adapts to new object shapes by leveraging a reference image and geometric correspondence. This results in robust performance across previously unseen objects. Extensive evaluation on the BOP benchmark datasets shows that RefPose achieves state-of-the-art results while maintaining a competitive runtime.

1. Introduction

6D pose estimation is a key aspect of computer vision and robotics, focusing on accurately predicting an object’s position (3D translation) and orientation (3D rotation) in a given scene. This task is essential for a range of applications, such as autonomous driving [5, 23], augmented reality (AR) [25, 38], and robotic manipulation [2, 36]. Despite significant research efforts in this field, estimating the 6D pose of previously unseen objects remains a considerable challenge. This challenge largely stems from the lack of prior knowledge and the limited generalization capabilities of existing models when faced with new objects [12].

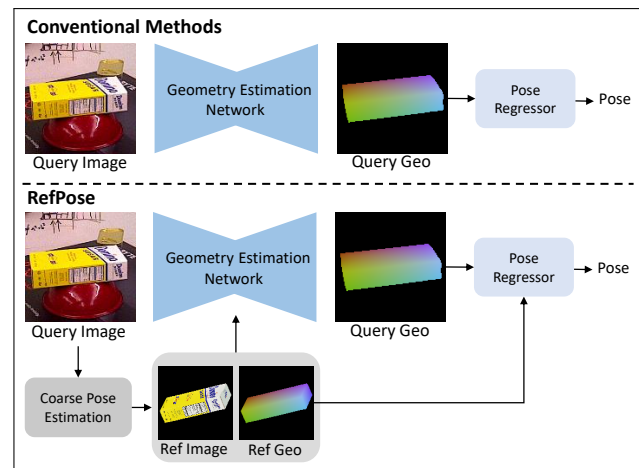


Figure 1. Comparison between conventional methods and proposed method (RefPose). In contrast to conventional methods, RefPose leverages a reference image and geometric correspondence generated from the estimated pose in the coarse pose estimation stage to guide the query’s geometric correspondence and pose estimation.

In instance-level object pose estimation, various methods rely on geometric correspondence as a crucial element for achieving accurate pose estimation [3, 7, 11, 19, 34, 35, 42, 45, 47]. Geometric correspondence refers to identifying the 3D model points that correspond to each pixel in a 2D image, which provides essential information for determining an object’s pose. Typically, this correspondence is estimated using deep learning networks and is subsequently used to infer the object’s pose through methods such as PnP/RANSAC [16] or neural network regression. However, these techniques face challenges in accurately estimating 2D-3D geometric correspondences for unseen objects, mainly due to their reliance on pre-defined object models and the limited generalization capabilities of the correspondence estimation network.

To address these limitations, we propose RefPose, a

novel approach to object pose estimation for unseen objects. As illustrated in Fig. 1, RefPose predicts the geometric correspondence of the query object by leveraging a reference image and geometric correspondence as guidance. This guidance provides crucial geometric information about the target object, allowing the network to avoid reliance on shape priors learned from a fixed set of objects during training. FoundPose [32] also estimates correspondences using pre-rendered templates through patch-wise matching, similar to our approach. However, these pre-rendered templates often lack proper alignment with the query image, leading to inaccurate matches and insufficient geometric information. To overcome this, we perform coarse pose estimation by processing the pre-rendered templates to obtain an initial pose, which is then used to render a reference image closely aligned with the query image, providing more reliable information. Additionally, rather than relying solely on direct matching, we carefully design a network that enhances correspondence estimation by integrating information from both the reference image and geometric correspondence guidance.

Specifically, in the coarse pose estimation stage, we select multiple templates from a set of pre-rendered templates for the target object. This selection is based on the accuracy of optical flow predictions made by the optical flow network [43], which will later help us estimate the geometric correspondence for the query object. We then use these selected templates in a warping-based approach that employs medoid voting to enhance robustness against outliers, yielding a reliable coarse geometric correspondence for the query object. Subsequently, we obtain an initial pose using PnP/RANSAC and render references accordingly. With synthesized reference guidance, we estimate a more reliable and precise geometric correspondence for the query during the refinement stage. To improve this estimation, we introduce a novel attention mechanism that leverages a correlation volume from the optical flow network, effectively integrating reference information. The estimated geometric correspondence for the query then serves as a fixed basis for further refining the pose. We iteratively update the initial pose by estimating the relative pose in comparison to the reference geometric correspondence. At each iteration, we re-render the reference geometric correspondence using the updated pose and compare it to the fixed geometric correspondence of the query. This render-and-compare process is repeated until we achieve an accurate final pose.

We assess our method on seven key datasets from the BOP benchmark [12]. Our findings indicate that RefPose achieves superior accuracy in both coarse pose estimation and final pose accuracy compared to state-of-the-art methods. By optimizing runtime in the pose refinement stage, RefPose not only delivers the highest accuracy across all methods but also maintains competitive speed.

Our contributions are as follows:

- We propose RefPose, a method that leverages a reference image and geometric correspondence to guide the estimation of the query’s geometric correspondence and pose, eliminating the need for shape priors learned from predefined object sets.
- We present a classifier based on optical flow for improved template selection. Additionally, we introduce a warping-based geometry estimation method that utilizes medoid voting to enhance robustness against outliers, leading to more accurate coarse pose estimates.
- We propose a correlation volume-guided attention mechanism, improving the model’s ability to focus on relevant regions in a reference image corresponding to the query image.
- We achieve state-of-the-art results on the BOP benchmark datasets while maintaining competitive runtime.

2. Related work

Geometric correspondence-based pose estimation. In instance-level object pose estimation, where training and testing are performed within a fixed set of objects, a common strategy is to utilize 2D-3D geometric correspondence. Most methods follow a two-step process: they first establish 2D-3D correspondences from an RGB image and then determine the pose using a RANSAC-based PnP algorithm or a neural network. Early studies [37, 44] employed the 3D bounding box corners of the object as keypoints, focusing on identifying the projected positions of these points in the image. For more robust and precise pose estimation, recent research [3, 7, 11, 19, 35, 45, 47] has shifted toward establishing dense correspondence maps rather than relying on sparse points. Consequently, designing and training deep learning networks that can accurately predict geometric correspondence maps is essential. However, these networks often struggle with generalization, especially when applied to unseen objects outside the fixed training set. Some models [20, 42] even require separate network models for each object. Therefore, to effectively estimate the geometric correspondence of the query object, we train the network using not only the query image but also a reference image and its geometric correspondence, providing reliable and valuable contextual information.

Most methods use 3D coordinates as the geometric correspondence; however, some approaches modify these coordinates to achieve finer correspondence estimation. For example, [20, 42] adopts a binary code representation, while [34] applies positional encoding, as seen in NeRF [27], to improve performance. Inspired by these methods, we also employ a positionally encoded representation for geometric correspondence during the refinement stage.

Unseen object pose estimation. Some studies focus on

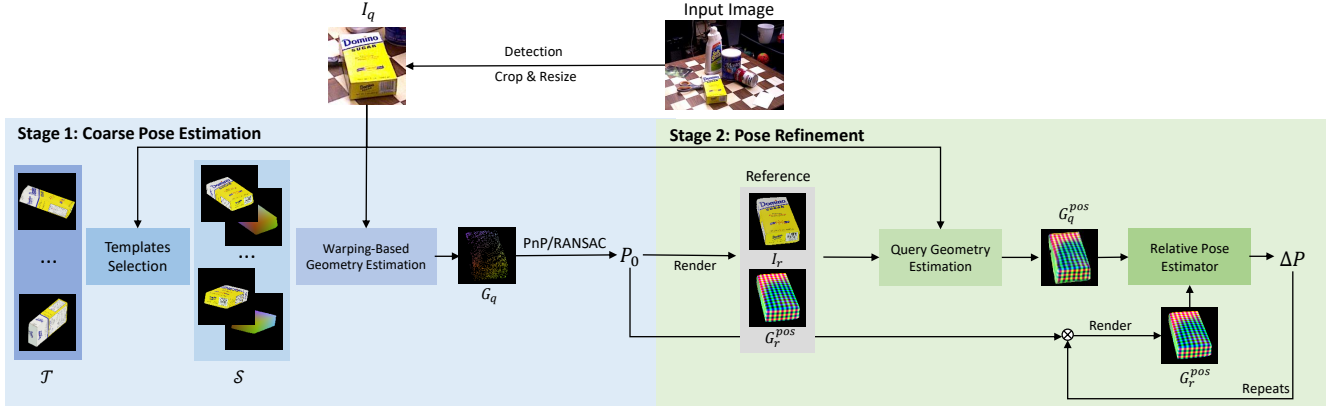


Figure 2. Overview of the RefPose pipeline. Given an input RGB image, the target object is first detected, cropped, and resized to create the query image, I_q . In **Stage 1: Coarse Pose Estimation**, a set of templates, \mathcal{S} , is selected from the pre-rendered template set, \mathcal{T} , to estimate an initial pose, P_0 , for the query object. In **Stage 2: Pose Refinement**, the query’s geometric correspondence, G_q^{pos} , is estimated using the rendered reference image, I_r , and geometric correspondence, G_r^{pos} . The initial pose, P_0 , is iteratively refined by estimating the relative pose, ΔP , between the query and reference. At each iteration, G_r^{pos} is re-rendered to align with the updated pose, leading to an accurate final pose estimate.

category-level pose estimation [6, 17, 24, 46], utilizing shared geometric traits within a category to broaden the range of identifiable target objects. However, these methods face generalization limitations, making it challenging to unseen object categories in real-world settings. To address these limitations, recent research has explored unseen object pose estimation, which aims to accurately predict poses for objects not encountered during training.

MegaPose [15] combines a render-and-compare refiner with a classifier that assesses whether the refiner can correct given pose errors, achieving strong generalization by training on a large synthetic dataset. GigaPose [30] leverages discriminative templates to handle out-of-plane rotations and uses patch correspondences for estimating remaining pose parameters, resulting in improvements in both speed and segmentation robustness. GenFlow [28] addresses the accuracy-scalability trade-off by directly leveraging the target object’s shape through optical flow prediction. GenFlow iteratively refines poses by leveraging a 3D shape constraint alongside a multi-scale, coarse-to-fine refinement process. FoundPose [32] establishes 2D-3D correspondences by matching patch descriptors from a self-supervised DINOv2 [31] model between the image and pre-rendered templates, integrating these descriptors into a bag-of-words model for more efficient template retrieval.

FoundPose is particularly relevant to our approach in that it also aims to estimate correspondences to assist pose estimation. However, FoundPose derives correspondences by performing patch-wise matching with pre-rendered templates, which may result in inaccuracies due to misalignment with the query image. In contrast, our method begins with a coarse pose estimation to establish an initial

pose, which we then refine using a rendered reference image closely aligned with the query object. Additionally, rather than relying solely on direct matching, we design a geometry estimation network that improves correspondence estimation by effectively integrating information from the reference image and geometric correspondence guidance.

3. Method

This section introduces RefPose, a novel approach to object pose estimation for unseen objects. We start with a brief overview (Sec. 3.1), followed by a detailed explanation of the initial pose estimation process used to generate a reference (Sec. 3.2). Finally, we describe how this reference information is leveraged to estimate the query’s geometric correspondence and iteratively refine the pose to reach the final result (Sec. 3.3).

3.1. Overview

RefPose follows a multi-stage pipeline, similar to other recent methods [15, 28, 30, 32], comprising object detection/segmentation, coarse pose estimation, and pose refinement. Following these methods, we use an off-the-shelf model [29] for object detection and segmentation to pre-process the input image.

Fig. 2 illustrates the overall pipeline of RefPose. Starting with an RGB image, the target object is detected, cropped, and resized to 256×256 to create the query image, I_q . In the coarse pose estimation stage, a classification network selects a set of templates, $\mathcal{S} = \{S_1, S_2, \dots, S_k\}$, from a pre-rendered template set, $\mathcal{T} = \{T_1, T_2, \dots, T_N\}$. This selected set, \mathcal{S} , is then used to estimate the geometric correspondence of the query, G_q , providing an initial pose esti-

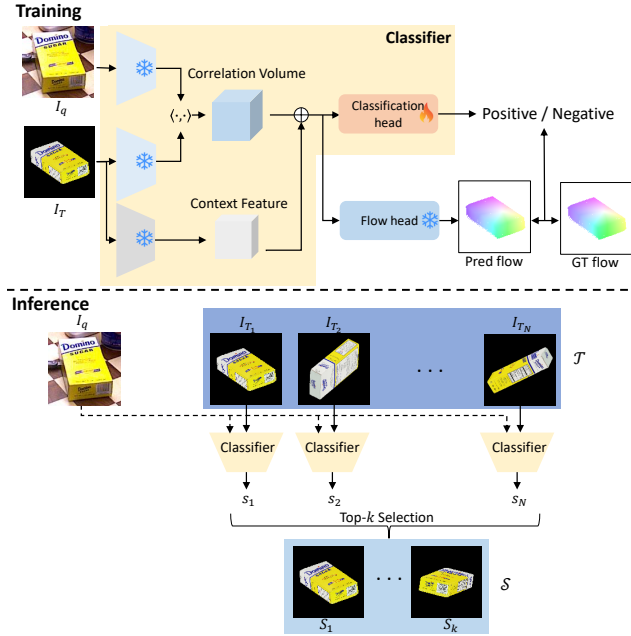


Figure 3. Templates selection using a classification network. The classification network scores pre-rendered templates based on how well optical flow can be estimated between each template and the query image. During inference, these scores are used to select the top- k templates. The classifier leverages a frozen feature encoder from a pre-trained optical flow network, with only the classification head trained.

mate, P_0 . Using P_0 , we render a reference image, I_r , and a geometric correspondence, G_r^{pos} , that are closely aligned with the query. Here, G_r^{pos} represents the geometric correspondence with positional encoding applied. In the refinement stage, this reference information is then used to estimate a more accurate geometric correspondence for the query, G_q^{pos} . Finally, the relative pose, ΔP , between the query and reference is iteratively refined through a render-and-compare approach to reach a final pose estimate.

3.2. Coarse pose estimation

Templates selection. We start by randomly sampling poses and rendering images along with geometric correspondences for each pose based on the given 3D model. The geometric correspondence $G \in \mathbb{R}^{h \times w \times 3}$ represents a dense map that indicates the corresponding 3D model point for each pixel. For simplicity, we refer to geometric correspondence as “geometry” throughout this paper. Inspired by MegaPose [15], we select a set of multiple templates, \mathcal{S} , using a classification network. However, unlike MegaPose, where the classifier is trained based on the refining capability of its refiner, we introduce a new training criterion. Since our subsequent pose estimation stage relies on optical flow, we train the classifier to assess the accuracy of optical

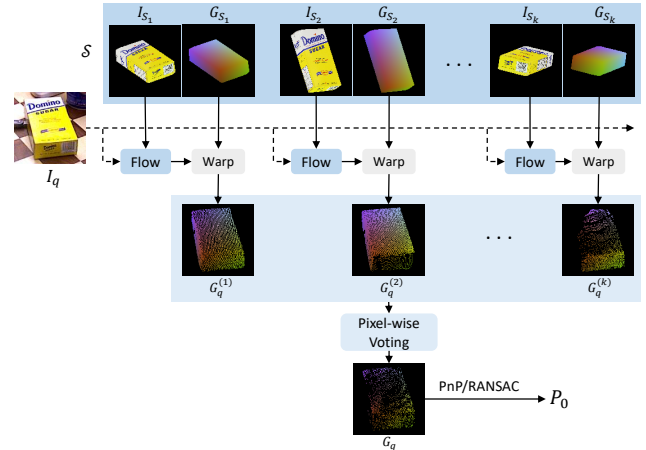


Figure 4. Warping-based geometry estimation process. The optical flow between each template image I_S in the selected set \mathcal{S} and the query image I_q is used to warp the corresponding template geometries G_S , generating candidate geometries for the query, G_q . A pixel-wise voting scheme refines these candidates, and the resulting 2D-3D correspondences in G_q are then applied with PnP/RANSAC to estimate the initial pose P_0 .

flow estimation between each template image, I_T , and the query image, I_q .

During training, we utilize a pre-trained optical flow network [43] to estimate the optical flow between I_q and each template image in \mathcal{T} . Positive and negative pairs are identified by comparing the predicted flow with the ground truth flow, which serves as labels for training the classifier. Additionally, rather than designing and training a new feature encoder for the classification network, we leverage the feature encoder from the optical flow network to enhance both the classifier’s performance and coherence with subsequent stages. Specifically, as in RAFT, we extract the correlation volume and context features, then attach a simple CNN as the classification head to complete the classifier. The classifier is trained using Binary Cross-Entropy (BCE) Loss [40]. During inference, the classifier ranks template images from the set \mathcal{T} , and we select the top- k templates based on the predicted scores. The classifier’s structure and training and inference processes are illustrated in Fig. 3.

Warping-based geometry estimation. To predict an initial pose P_0 , we estimate the query geometry G_q based on a set of selected templates \mathcal{S} . First, we calculate the optical flow between each template image I_S and the query image I_q , then warp the template geometries G_S accordingly. These warped geometries serve as candidates for G_q . However, due to potential inconsistencies and inaccuracies in the optical flow estimated from each template, these candidates may indicate different 3D points for the same query pixel. To address this, we employ a voting scheme to select a single 3D point per pixel. PFA [13] follows a similar approach

by aggregating multiple optical flows from different templates to estimate the 2D-3D correspondences. However, it performs this aggregation without explicitly addressing the inconsistencies and inaccuracies, which can lead to unreliable correspondences. In contrast, our method employs a more robust medoid-based voting scheme, where the most representative 3D point per pixel is selected rather than aggregating all candidates indiscriminately. This approach ensures that errors in optical flow estimation do not adversely affect the final result, leading to a more accurate and stable coarse pose estimation. Finally, based on the established 2D-3D correspondences in G_q , we apply the PnP/RANSAC algorithm to estimate the pose P_0 . The overall process is illustrated in Fig. 4.

3.3. Pose refinement

Using the initial pose P_0 obtained from the coarse pose estimation stage, we render a single reference image I_r and geometry G_r^{pos} that are more closely aligned with the query image I_q than the previously selected templates in \mathcal{S} . Rather than using raw 3D coordinates, we apply positional encoding to the 3D coordinates of G to enrich the geometry representation, thereby improving estimation accuracy [34]. This encoding results in a representation $G^{pos} \in \mathbb{R}^{h \times w \times 6N_{freq}}$, where N_{freq} denotes the frequency bands used for encoding [27].

Correlation volume-guided attention mechanism. Accurately retrieving relevant geometric information from the reference geometry G_r^{pos} requires precise pixel-level correspondence between the query image I_q and the reference image I_r . To achieve this, we introduce a correlation volume-guided attention mechanism. In this setup, the query image I_q serves as the query, the reference image I_r acts as the key, and the reference geometry G_r^{pos} is treated as the value. The mechanism computes attention weights based on the similarity between the query (I_q) and the key (I_r), enabling the extraction of relevant features from the value (G_r^{pos}).

In contrast to vanilla attention mechanisms that implicitly learn relevance, our approach uses explicit pixel-level correspondence information, as both the query image I_q and reference image I_r represent the same object from different viewpoints. We obtain the attention weights by applying a softmax operation to the correlation volume from the optical flow network, which encodes pixel-level similarities between I_q and I_r . By applying these weights to the reference geometry features G_r^{pos} , we retrieve highly relevant geometric information, facilitating accurate correspondence estimation for improved pose refinement.

Geometry estimation network. Our geometry estimation network processes the query image I_q , reference image I_r , and reference geometry G_r^{pos} as inputs and outputs the query geometry G_q^{pos} . To capture multi-scale features ef-

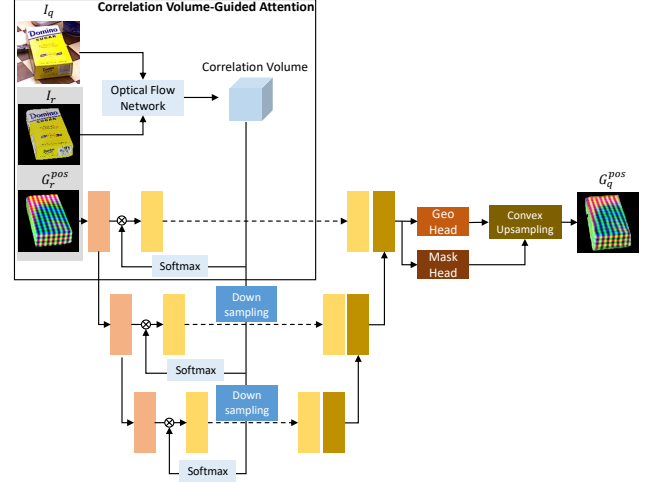


Figure 5. Architecture of the geometry estimation network. The network takes the query image I_q , reference image I_r , and reference geometry G_r^{pos} as inputs to estimate the query geometry G_q^{pos} . A correlation volume-guided attention mechanism is applied at each level of the U-Net to effectively integrate these inputs. The geo head and mask head output a low-resolution geometry map and mask, which are then refined through convex upsampling to produce the final high-resolution geometry G_q^{pos} .

fectively, we employ a U-Net structure [39], with correlation volume-guided attention mechanisms applied at each level to incorporate reference information. As features are downsampled within the U-Net, the correlation volume is correspondingly downsampled. The U-Net concludes with two heads: a geo head, which estimates geometry at an 8x downsampled resolution, and a mask head, which predicts an up-mask for convex upsampling. Following RAFT [43], convex upsampling is applied to reconstruct G_q^{pos} at the original resolution, providing higher fidelity than bilinear upsampling. The complete network architecture is illustrated in Fig. 5.

The network is optimized by minimizing the \mathcal{L}_1 loss, which measures the difference between the predicted and ground truth geometry:

$$\mathcal{L}_{geo} = \|\bar{G}_q^{pos} - G_q^{pos}\|_1, \quad (1)$$

where \bar{G}_q^{pos} is the ground truth geometry and G_q^{pos} is the predicted geometry.

Iterative pose refinement. The estimated query geometry G_q^{pos} enables iterative pose refinement through a render-and-compare approach [14, 18, 33]. Initially, a relative pose ΔP is estimated between G_q^{pos} and the reference geometry G_r^{pos} , updating the initial pose P_0 . With each update, a new reference geometry is rendered and compared to the fixed query geometry G_q^{pos} , progressively refining the pose. This iterative process continues until the final pose is accurately determined.

The relative pose estimator, based on a CNN, is trained using sequence loss [43], which is inspired by SCFlow [10], to improve learning efficiency and ensure consistent prediction quality across iterations. The sequence loss, \mathcal{L}_{seq} , is defined as follows:

$$\mathcal{L}_{seq} = \sum_{m=1}^M \gamma^{M-m} \mathcal{L}_{pose}^{(m)}, \quad (2)$$

where M denotes the total number of refinement iterations, γ is an exponential weighting factor, and $\mathcal{L}_{pose}^{(m)}$ represents the pose loss at each iteration m .

After updating the reference pose, we compare it with the ground truth pose to assess the refinement strategy. The pose loss \mathcal{L}_{pose} , which measures this alignment, combines grid-matching and grid-distance loss functions [33], formulated as follows:

$$\mathcal{L}_{pose} = \|\bar{\mathcal{G}} - \mathcal{G}\|_2 + \|\|\bar{t}\|_2 - \|t\|_2\|_1, \quad (3)$$

where $\bar{\mathcal{G}}$ and \bar{t} are the ground truth grid and translation vectors, respectively, while \mathcal{G} and t are derived from the estimated pose.

4. Experiment

4.1. Experimental Setup

Datasets. We train our model on the synthetic dataset, Google Scanned Objects (GSO) [9], as provided by [15]. This dataset contains nearly 1 million images representing 1,000 different object types. Although [15] also includes data from approximately 50,000 ShapeNet [4] objects, we opted to use only the GSO dataset due to computational and memory constraints. This choice is further supported by the higher-quality mesh models in the GSO dataset, which, as noted in [15], contribute more critically to model performance than ShapeNet’s. For evaluation, we test our approach on the seven primary datasets in the BOP benchmark [12], including YCB-V, LM-O, T-LESS, TUD-L, ICBIN, ITODD, and HB. Our method relies solely on RGB images and 3D object models, without leveraging any depth information.

Evaluation metrics. We follow the standard BOP evaluation protocol [12], which employs three core metrics: Visible Surface Discrepancy (VSD), Maximum Symmetry-Aware Surface Distance (MSSD), and Maximum Symmetry-Aware Projection Distance (MSPD). Overall performance, referred to as Average Recall (AR), is calculated by averaging the individual recall scores for each metric across a range of error thresholds.

Implementation details. Our model is trained with the AdamW optimizer [21], using a batch size of 8 and a learning rate of 0.0001 for a total of 400k training steps. A cosine annealing scheduler [22] with a 10k step period is employed

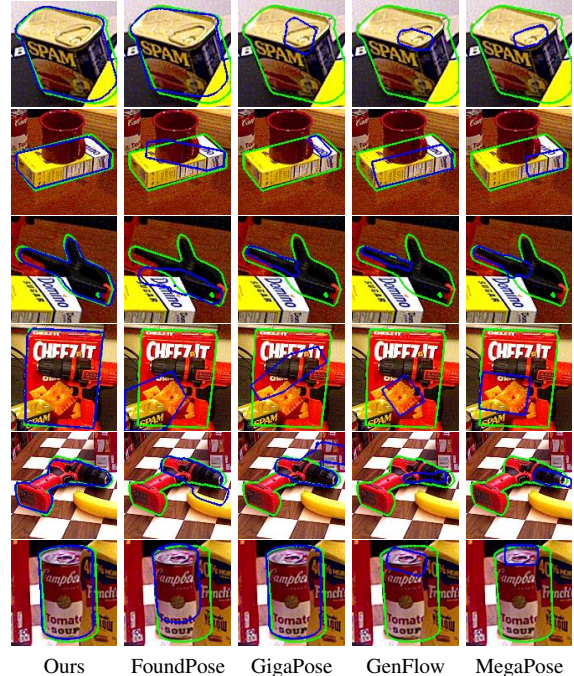


Figure 6. Qualitative comparison of pose estimation results. We present a qualitative comparison of our method against other approaches, with the projected contours from the ground-truth pose shown in green and those from the predicted pose in blue.

to adjust the learning rate. Both training and evaluation are conducted on an RTX-3090 GPU. For the optical flow network in RefPose, we use the large model of RAFT [43], which is pre-trained on the FlyingChairs [8] and FlyingThings3D [26] datasets and then fine-tuned on the GSO dataset for our specific application. In the pose refinement stage, we generate the geometry G^{pos} using sine and cosine positional encoding with five frequency bands, following the approach in [34]. The number of pose refinement iterations, M , is set to 5, balancing accuracy and runtime efficiency.

4.2. Comparison with state-of-the-art methods

Tab. 1 presents the results of our method on the BOP benchmark datasets. Except for OSOP [41], which uses its own detection model, all methods employ CNOS [29] as the detection/segmentation model to ensure a fair comparison. Our method achieves the best performance in both the coarse pose estimation and the refined results following the refinement stage. While our approach slightly underperforms the current state-of-the-art methods on LM-O, T-LESS, and ITODD, it achieves state-of-the-art results on the other datasets, with particularly notable improvements on YCB-V and HB. Overall, our method demonstrates the best performance across all datasets. Fig. 6 provides qualitative comparisons with existing methods, further illustrating the robustness of our approach.

Table 1. Evaluation results on BOP benchmark datasets. The table reports Average Recall (AR) scores across the seven datasets in the BOP challenge, where higher AR scores indicate better performance. The best-performing method is highlighted in bold, and the second-best is underlined. The top section presents results from coarse pose estimation alone, while the bottom section displays results after applying the refinement stage. ‘‘MH’’ denotes MegaPose and GenFlow versions that incorporate a multi-hypotheses strategy in the refinement stage, and ‘‘featuremetric’’ refers to the refinement method introduced in FoundPose.

Method	Refinement	YCB-V	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	Mean	Run-time
OSOP [41]	-	29.6	27.4	40.3	-	-	-	-	-	-
ZS6D [1]	-	32.4	29.8	21.0	-	-	-	-	-	-
MegaPose [15]	-	28.1	22.9	17.7	25.8	15.2	10.8	25.1	20.8	15.5s
GenFlow [28]	-	27.7	25.0	21.5	30.0	16.8	15.4	28.3	23.	3.8s
GigaPose [30]	-	27.8	29.6	26.4	30.0	22.3	17.5	34.1	26.8	0.4s
FoundPose [32]	-	<u>45.2</u>	39.6	33.8	<u>46.7</u>	23.9	<u>20.4</u>	50.8	<u>37.2</u>	<u>1.7s</u>
RefPose (Ours)	-	50.0	<u>35.8</u>	<u>38.1</u>	48.5	<u>23.1</u>	21.5	<u>49.8</u>	38.1	3.1s
MegaPose [15]	MegaPose	60.1	49.9	47.7	65.3	36.7	31.5	65.4	50.9	17.0s
MegaPose [15]	MegaPose, MH	62.1	56.0	50.7	68.4	41.4	33.8	70.4	54.7	21.9s
MegaPose [15]	RefPose (Ours)	65.3	56.0	52.8	66.4	45.3	41.1	73.6	57.2	16.4s
GenFlow [28]	GenFlow, MH	63.3	56.3	52.3	68.4	45.3	39.5	73.9	57.0	20.8s
GigaPose [30]	MegaPose	63.2	55.7	54.1	58.0	45.0	37.6	69.3	54.7	2.3s
GigaPose [30]	GenFlow, MH	65.2	63.1	58.2	66.4	<u>49.8</u>	45.3	<u>75.6</u>	<u>60.5</u>	10.6s
FoundPose [32]	MegaPose, MH + Featuremetric	<u>69.0</u>	<u>61.0</u>	57.0	<u>69.4</u>	47.9	40.7	72.3	59.6	20.5s
RefPose (Ours)	MegaPose	63.7	56.3	51.1	65.8	43.7	41.4	71.8	56.3	4.6s
RefPose (Ours)	RefPose (Ours)	72.7	59.6	<u>57.8</u>	69.7	51.2	<u>43.8</u>	76.2	61.4	<u>3.9s</u>

The 10th and 15th rows of Tab. 1 report results where our proposed coarse pose estimation and refinement methods are independently combined with MegaPose to assess their standalone effectiveness. A comparison among the 8th, 12th, and 15th rows shows that our coarse pose estimation achieves better results even when using the same refinement method as MegaPose. Additionally, comparing the 8th and 10th rows demonstrates that our refinement method is more effective even when applied to the coarse poses estimated by MegaPose.

The reported runtime represents the average speed for processing all objects within a single image. Though our method takes slightly longer in the coarse pose estimation stage, it significantly reduces refinement time by minimizing the number of renderings and using a lightweight model. As a result, our method achieves superior performance with runtime comparable to other state-of-the-art approaches.

4.3. Ablation study

Number of pre-rendered templates. The results of the ablation study on the number of pre-rendered templates in the set \mathcal{T} , represented by N , are shown in Tab. 2. With only 64 templates, the sampled poses are too sparse across the object’s orientation space, leading to reduced performance in pose estimation due to insufficient coverage of various possible object poses. Increasing to 128 templates provides a denser sampling, significantly enhancing query-template matching and improving performance. Using more than 128 templates yields only marginal improvements while increasing both memory and computational overhead. Additionally, a larger template set results in longer inference times, as more templates must be evaluated. Therefore, using 128 templates achieves an optimal balance between ef-

Table 2. Ablation study on the number of pre-rendered templates. The table presents AR scores, illustrating the impact of varying numbers of pre-rendered templates, \mathcal{T} , on coarse pose estimates.

N	YCB-V	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	Mean
64	45.2	29.8	36.5	43.7	20.1	18.8	43.5	33.9
128	50.0	35.8	38.1	48.5	23.1	21.5	49.8	38.1
256	48.8	37.2	39.5	45.8	24.3	23.1	49.8	38.3

Table 3. Ablation study on the number of selected templates. The table presents AR scores, showing the impact of varying numbers of selected templates, \mathcal{S} , on coarse pose estimates.

k	YCB-V	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	Mean
1	42.6	31.0	28.8	40.1	18.8	18.1	40.8	31.5
2	46.8	37.2	34.0	43.1	19.8	18.8	46.5	35.2
4	50.0	35.8	38.1	48.5	23.1	21.5	49.8	38.1
8	50.0	29.8	36.4	44.5	21.7	19.9	51.0	36.2

iciency and performance.

Number of selected templates. The results of our ablation study on the number of selected templates \mathcal{S} , denoted by k , are shown in Tab. 3. Selecting four templates from the pre-rendered set \mathcal{T} yields the best results, achieving a balance between diversity and alignment accuracy. Using fewer templates limits diversity, increasing the risk of alignment errors during warping-based geometry estimation if a template is poorly selected or if the optical flow is inaccurately estimated potentially, leading to inaccurate pose estimation. Conversely, selecting more than four templates raises the likelihood of including misaligned templates, which may reduce the effectiveness of the medoid-based voting stage in handling outliers. With too many templates, the medoid’s robustness can be compromised, as it becomes more challenging to filter out misaligned correspondence effectively.

Table 4. Ablation study on components in coarse pose estimation stage. This table presents AR scores for the coarse pose estimates.

Setting	YCB-V	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	Mean
w/o Flow features	46.8	34.2	33.2	47.1	22.3	20.9	45.8	35.8
w/o Medoid	49.6	29.6	36.8	45.9	21.3	18.8	50.2	36.0
Ours	50.0	35.8	38.1	48.5	23.1	21.5	49.8	38.1

Table 5. Ablation study on components of the geometry estimation network in the pose refinement stage. This table presents AR scores for the pose refinement results.

Setting	YCB-V	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	Mean
w/o P.E.	69.3	57.1	57.0	68.4	48.3	39.5	73.9	59.1
w/o Convex.	70.0	56.8	58.2	68.4	47.9	43.1	75.6	60.0
w/o C.G. attn.	68.8	61.0	52.8	66.0	47.3	41.9	74.8	58.9
Ours	72.7	59.6	57.8	69.7	51.2	43.8	76.2	61.4

Thus, selecting four templates provides an effective balance, ensuring sufficient diversity while minimizing alignment errors, which is crucial for accurate geometry and pose estimation.

Components in the coarse pose estimation stage. Tab. 4 presents the results of our ablation study on key components used in the coarse pose estimation stage. The first row evaluates the impact of using the feature encoder from the optical flow network as the classifier’s feature encoder. This feature encoder outputs flow features, including the correlation volume and context feature, which provide rich cues related to optical flow, enhancing the classifier’s accuracy in selecting templates. By directly leveraging the same encoder applied in the warping-based geometry estimation stage, consistency across stages is maintained, contributing to superior results.

The “w/o Medoid” variant represents a model in which, during pixel-wise voting in the warping-based geometry estimation, the medoid is replaced with a simple average for selecting correspondences. Using the medoid rather than averaging mitigates the impact of outliers that may arise from imperfect optical flow estimations, leading to a more robust correspondence selection. This robustness directly translates to greater accuracy of the coarse pose estimation. **Components of the geometry estimation network in the pose refinement stage.** Tab. 5 presents the ablation study results on key components of the geometry estimation network within the pose refinement stage. The first row shows that applying positional encoding to the 3D coordinates, inspired by [34], improves performance over using raw 3D coordinates as geometric correspondence. Specifically, the “w/o P.E.” variant, which omits positional encoding and directly uses 3D coordinates, shows lower accuracy.

Additionally, estimating an up-mask and utilizing convex upsampling yield superior performance compared to bilinear upsampling, as indicated by the “w/o Convex.” variant. Convex upsampling more effectively preserves spatial detail at the original resolution, leading to a closer alignment between the estimated geometry and the true geometry

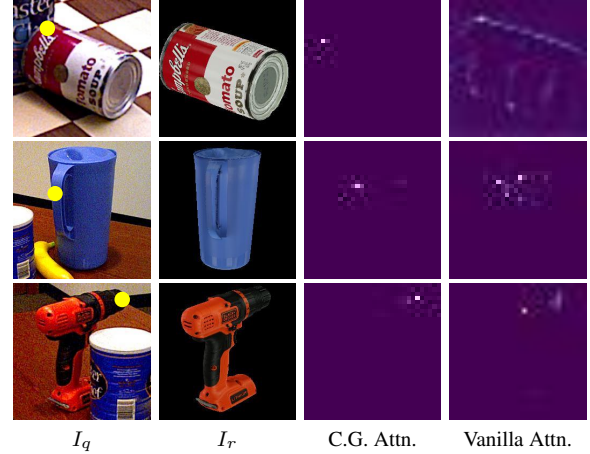


Figure 7. Visual comparison of attention mechanisms in the geometry estimation network. Each row shows a query image I_q with a yellow dot marking the point of interest, the corresponding reference image I_r , and the attention maps produced by the correlation volume-guided (C.G.) attention and vanilla attention. The C.G. attention accurately focuses on relevant regions in I_q and I_r , while vanilla attention lacks this precision.

try of the object.

Finally, the proposed correlation volume-guided (C.G.) attention mechanism outperforms vanilla attention, as demonstrated by the “w/o C.G. attn.” variant, which replaces C.G.attention with vanilla attention. This result underscores the effectiveness of C.G. attention in accurately capturing relevant correspondences between the query and reference images. As illustrated in Fig. 7, C.G. attention maps exhibit a sharper focus on relevant regions in the reference image that correspond to points of interest in the query, while vanilla attention maps appear less precise. This comparison highlights the effectiveness of C.G. attention in establishing accurate correlations between the query and reference images.

5. Conclusion

In this paper, we have proposed RefPose, a two-stage method designed for enhanced accuracy and generalization in unseen object pose estimation. Starting with a coarse pose estimation using template selection and medoid-based voting, RefPose builds an initial pose, which is then used to assist the geometry estimation through a correlation volume-guided attention mechanism. This refined geometry for the query supports an iterative render-and-compare process, producing a precise final pose. Extensive experiments on the BOP benchmark demonstrate RefPose’s strong performance and generalization ability, while ablation studies confirm the effectiveness of each component. RefPose advances adaptable and efficient solutions for 6D pose estimation in complex, real-world scenarios.

Acknowledgements This work was supported by Samsung Electronics Co., Ltd.(No. 0423-20240056), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], and in part by the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2025.

References

- [1] Philipp Ausserlechner, David Haberger, Stefan Thalhammer, Jean-Baptiste Weibel, and Markus Vincze. Zs6d: Zero-shot 6d object pose estimation using vision transformers. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 463–469. IEEE, 2024.
- [2] Benjamin Busam, Marco Esposito, Simon Che’Rose, Nassir Navab, and Benjamin Frisch. A stereo vision approach for cooperative robotic movement therapy. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 127–135, 2015.
- [3] Ming Cai and Ian Reid. Reconstruct locally, localize globally: A model free method for object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3153–3163, 2020.
- [4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.
- [5] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017.
- [6] Xu Chen, Zijian Dong, Jie Song, Andreas Geiger, and Otmar Hilliges. Category level object pose estimation via neural analysis-by-synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 139–156. Springer, 2020.
- [7] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, and Federico Tombari. So-pose: Exploiting self-occlusion for direct 6d pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12396–12405, 2021.
- [8] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.
- [9] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022.
- [10] Yang Hai, Rui Song, Jiaojiao Li, and Yinlin Hu. Shape-constraint recurrent flow for 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4831–4840, 2023.
- [11] Tomas Hodan, Daniel Barath, and Jiri Matas. Epos: Estimating 6d pose of objects with symmetries. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11703–11712, 2020.
- [12] Tomas Hodan, Martin Sundermeyer, Yann Labbe, Van Nguyen Nguyen, Gu Wang, Eric Brachmann, Bertram Drost, Vincent Lepetit, Carsten Rother, and Jiri Matas. Bop challenge 2023 on detection segmentation and pose estimation of seen and unseen rigid objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5610–5619, 2024.
- [13] Yinlin Hu, Pascal Fua, and Mathieu Salzmann. Perspective flow aggregation for data-limited 6d object pose estimation. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022.
- [14] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 574–591. Springer, 2020.
- [15] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022.
- [16] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep n p: An accurate o (n) solution to the p n p problem. *International journal of computer vision*, 81:155–166, 2009.
- [17] Fu Li, Ivan Shugurov, Benjamin Busam, Minglong Li, Shaowu Yang, and Slobodan Ilic. Polarmesh: A star-convex 3d shape approximation for object pose estimation. *IEEE Robotics and Automation Letters*, 7(2):4416–4423, 2022.
- [18] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 683–698, 2018.
- [19] Zhigang Li, Gu Wang, and Xiangyang Ji. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7678–7687, 2019.
- [20] Ruyi Lian and Haibin Ling. Checkerpose: Progressive dense keypoint localization for object pose estimation with graph neural network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14022–14033, 2023.
- [21] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [22] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- [23] Fabian Manhardt, Wadim Kehl, and Adrien Gaidon. Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In *Proceedings of the IEEE/CVF Conference*

- on *Computer Vision and Pattern Recognition*, pages 2069–2078, 2019.
- [24] Fabian Manhardt, Gu Wang, Benjamin Busam, Manuel Nickel, Sven Meier, Luca Minciullo, Xiangyang Ji, and Nassir Navab. Cps++: Improving class-level 6d pose and shape estimation from monocular images with self-supervised learning. *arXiv preprint arXiv:2003.05848*, 2020.
- [25] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 22(12):2633–2651, 2015.
- [26] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016.
- [27] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- [28] Sungphill Moon, Hyeontae Son, Dongcheol Hur, and Sangwook Kim. Genflow: Generalizable recurrent flow for 6d pose refinement of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2024.
- [29] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Vincent Lepetit, and Tomas Hodan. Cnos: A strong baseline for cad-based novel object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2134–2140, 2023.
- [30] Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. Gigapose: Fast and robust novel object pose estimation via one correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9903–9913, 2024.
- [31] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [32] Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomas Hodan. Foundpose: Unseen object pose estimation with foundation features. In *European Conference on Computer Vision*, pages 163–182. Springer, 2025.
- [33] Jaewoo Park and Nam Ik Cho. Dprost: Dynamic projective spatial transformer network for 6d pose estimation. In *European Conference on Computer Vision*, pages 363–379. Springer, 2022.
- [34] Jaewoo Park, Jaeguk Kim, and Nam Ik Cho. Leveraging positional encoding for robust multi-reference-based object 6d pose estimation. *arXiv preprint arXiv:2401.16284*, 2024.
- [35] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7668–7677, 2019.
- [36] Luis Pérez, Íñigo Rodríguez, Nuria Rodríguez, Rubén Usamentiaga, and Daniel F García. Robot guidance using machine vision techniques in industrial environments: A comparative review. *Sensors*, 16(3):335, 2016.
- [37] Mahdi Rad and Vincent Lepetit. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In *Proceedings of the IEEE international conference on computer vision*, pages 3828–3836, 2017.
- [38] Jason Rambach, Alain Pagani, Michael Schneider, Oleksandr Artemenko, and Didier Stricker. 6dof object tracking based on 3d scans for augmented reality remote live support. *Computers*, 7(1):6, 2018.
- [39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.
- [40] Usha Ruby and Vamsidhar Yendapalli. Binary cross entropy with deep learning technique for image classification. *Int. J. Adv. Trends Comput. Sci. Eng.*, 9(10), 2020.
- [41] Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic. Osop: A multi-stage one shot object pose estimation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6835–6844, 2022.
- [42] Yongzhi Su, Mahdi Saleh, Torben Fetzner, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari. Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6738–6748, 2022.
- [43] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.
- [44] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 292–301, 2018.
- [45] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. 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*, pages 16611–16621, 2021.
- [46] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019.
- [47] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1941–1950, 2019.