WAPR: A Foundation Model for Wide-Angle Refinement in Unseen Object Pose Estimation

Yulin Wang*[1], Mengting Hu*[1], Hongli Li[2], Jianghao Zhou[1], and Chen LUO†[1]

[1] Southeast University, Nanjing, China
{yulinwang, mengtinghu, jianghaozhou, chenluo}@seu.edu.cn
[2] Purdue University, West Lafayette, IN, USA
li5125@purdue.edu
*Equal contribution. †Corresponding author.

Abstract. Real-world applications require 6D pose estimation to be accurate, fast, and scalable to unseen objects. This paper introduces WAPR, a zero-shot wide-angle pose refinement model that refines candidate poses with rotational deviations up to 90°. With as few as 12 candidate poses per detected object instance, WAPR supports fast inference within 1 s per frame and reaches a pose-estimation throughput of up to 25 detected object instances per second. To support wide-angle training for rotationally symmetric objects, WAPR uses rotational symmetry priors to canonicalize symmetry-equivalent pose targets before loss computation. We further construct SA6D, a large-scale 6D training dataset with such priors. SA6D obtains KASAL-assisted rotational symmetry priors for 944 GSO scans and expands them through geometry and texture augmentation into ∼50K augmented object instances and ∼2M rendered RGB-D images. In addition, an angle-balanced loss stabilizes learning across different angular ranges by reducing the influence of uninformative large-error cases. Experiments on seven BOP core datasets show that WAPR achieves state-of-the-art performance in unseen-object 6D pose localization and detection under both fast and unconstrained inference settings. Project page: https://github.com/WangYuLin-SEU/WAPR.

Keywords: Unseen Object Pose Estimation · 6D Pose Refinement · Wide-Angle Pose Refinement · Rotational Symmetry Priors · Synthetic RGB-D Dataset · BOP Challenge

1 Introduction

6D object pose estimation [18–20, 40] is a fundamental problem in computer vision, with wide-ranging applications in augmented reality [29, 43, 58], robotics [33, 49, 59, 61], autonomous driving [22, 55, 62], and industrial automation [23, 45]. Traditional pipelines typically combine 2D detection or segmentation [11, 14, 46, 50] with 6D pose estimation [48, 51, 52], followed by optional refinement [30, 60]. While effective for trained objects, their reliance on object-specific training limits scalability in real-world deployments.

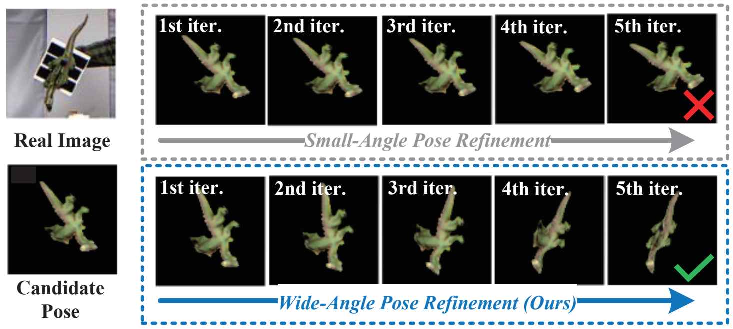

Fig. 1: Visualization of multi-iteration refinement for candidate poses with large angular deviations. The example is from TUD-L [18], where the handheld object moves continuously. With large deviations between the candidate pose and the ground-truth pose, conventional small-angle refiners often fail, whereas WAPR corrects the pose through multi-iteration refinement.

To improve scalability, unseen-object pose estimation [3, 37, 54] estimates poses for novel objects from object models with minimal or no retraining. Recent methods [3,36,37,39,63] commonly follow a candidate-based pipeline, where coarse pose hypotheses are generated from rendered views or feature correspondences and then refined or scored. This design makes candidate coverage and refinement range closely coupled. Using many candidates improves pose coverage but increases runtime, while sparse candidates are efficient but may leave larger angular deviations to the refiner. As shown in Fig. 1, such deviations can arise from coarse hypotheses, occlusion, and large viewpoint changes, motivating a refinement model that remains reliable across a wider angular range.

This paper introduces WAPR, a zero-shot foundation model for wide-angle pose refinement. WAPR follows the render-and-compare paradigm by comparing a rendered RGB-D observation and mask under a candidate pose with the real RGB-D input, and directly regresses rotation and translation updates. During training, ground-truth poses are perturbed by up to 90°, allowing WAPR to learn pose updates over a wide angular range. At inference time, this wide-angle refinement capability enables accurate pose estimation with a small number of candidates, including only 12 candidates in the fast setting.

Wide-angle training also makes pose ambiguity more prominent for rotationally symmetric objects. Different symmetry-equivalent poses may produce indistinguishable or nearly indistinguishable observations, and assigning arbitrary raw rotation targets can make visually similar candidate–target pairs receive inconsistent supervision. Following the common practice in 6D pose estimation of treating symmetry-equivalent poses consistently [26, 48, 52, 57], WAPR uses rotational symmetry priors to canonicalize symmetry-equivalent pose targets before loss computation. This makes wide-angle candidate–target pairs well-defined during training.

To support this training at scale, this paper constructs SA6D, a large-scale 6D dataset with rotational symmetry priors. Large-scale synthetic datasets for unseen-object pose estimation [27,54] provide diverse object models and rendered RGB-D supervision. SA6D complements this direction by adding dataset-level rotational symmetry priors. It uses KASAL [53] to obtain reliable symmetry priors for 944 GSO [8] scans, and expands these annotated assets through geometry and texture augmentation into $\sim$50K augmented object instances and 2M synthetic RGB-D images. The training objective also includes an angle-balanced loss, which reduces the dominance of low-correspondence large-error samples during wide-angle optimization.

Experiments on seven core BOP datasets [1,7,9,17,18,24,57] evaluate WAPR for zero-shot unseen-object 6D pose localization and detection. In the fast setting (<1 s per frame) with MUSE detections, WAPR reaches 80.6% AR / 78.6% AP, improving over Co-op [37] by +4.7% AR / +9.3% AP. With Co-op's F3DT2D detector under the same fast-inference protocol, WAPR obtains 78.1% AR / 76.5% AP, giving a matched-detector gain of +2.2% AR / +7.2% AP. In the unconstrained multi-detector setting, WAPR reaches 84.5% AR / 85.3% AP, improving over FreeZeV2 [3,4] by +1.2% AR / +2.0% AP.

The main contributions are as follows.

- This paper presents WAPR, a zero-shot foundation model for wide-angle pose refinement. With render-and-compare RGB-D inputs and iterative refinement, WAPR handles rotational deviations up to 90° and enables accurate pose estimation with sparse candidate poses.
- This paper constructs SA6D, a large-scale 6D dataset with rotational symmetry priors. SA6D uses a cost-efficient KASAL-assisted strategy to expand 944 annotated GSO scans into $\sim$50K augmented object instances and 2M synthetic RGB-D images.
- This paper develops wide-angle training and sparse candidate-based inference for unseen-object pose estimation. Rotational symmetry priors canonicalize ambiguous training pairs, angle-balanced loss stabilizes wide-angle learning, and score-based selection produces accurate final poses from a small candidate set.

2 Related Works

This section reviews 6D pose estimation for seen objects [12, 13, 21, 34, 48, 56], unseen-object 6D pose estimation [3, 27, 32, 36, 37, 54], and symmetry handling in 6D pose learning and evaluation.

2.1 6D Pose Estimation for Seen Objects

Seen-object 6D pose estimation assumes that training data are available for each target object. Early methods directly regress object rotation and translation from RGB or RGB-D observations [2, 57]. Later methods improve robustness by predicting keypoints or dense surface coordinates and recovering the pose with PnP [16, 28, 31, 42, 44, 52]. This correspondence-based formulation provides explicit 2D–3D constraints and is effective under partial occlusion.

Pose refinement further improves accuracy by aligning rendered observations with real images. Representative methods such as DeepIM [30] and DPOD [60] refine an initial pose by comparing rendered and observed images. Recent methods such as GDR-Net [51] and ZebraPose [48] combine detection, dense correspondence prediction, PnP, and optional refinement to achieve strong performance on trained objects. These object-specific pipelines provide the basis for many recent unseen-object methods, while scalable deployment to newly introduced objects motivates methods that reduce object-specific training.

2.2 6D Pose Estimation for Unseen Objects

Unseen-object 6D pose estimation estimates poses for novel objects from object models with minimal or no object-specific retraining. MegaPose [27] trains large-scale render-and-compare models on synthetic data and demonstrates strong generalization to unseen objects. FoundationPose [54] combines pose initialization, iterative refinement, and pose scoring in a unified framework. SAM-6D [32] and CNOS [38] use large visual models for zero-shot object localization and segmentation. Recent methods including FreeZe [3], GenFlow [36], FoundPose [63], GigaPose [39], and Co-op [37] improve different stages of the pipeline, including object detection, coarse pose retrieval, feature matching, refinement, and pose selection.

Most unseen-object pipelines generate candidate poses from rendered templates or correspondence matching, refine these candidates, and then select the final pose. This design makes candidate coverage and inference efficiency closely connected. Dense candidate sampling improves pose coverage but increases runtime, while sparse candidates are more efficient but place higher demand on the refinement stage. Many existing refinement modules are most effective when the initial hypothesis is already close to the target pose. WAPR targets the complementary wide-angle regime, where sparse candidates leave larger residual deviations for the refiner.

2.3 Symmetry Handling in 6D Pose Learning and Evaluation

Object symmetry is a long-standing issue in 6D pose estimation because multiple poses may correspond to the same or nearly the same visual observation. Benchmarks and evaluation protocols therefore use symmetry-aware pose metrics to avoid penalizing equivalent poses [19]. Learning-based methods also incorporate symmetry into supervision. PoseCNN [57] introduces the ShapeMatch loss to make pose learning invariant to indistinguishable symmetric poses, while CosyPose [26] uses symmetry-aware pose objectives for object-specific pose estimation. Coordinate-based methods such as ZebraPose [48] and HccePose [52] also rely on symmetry-consistent definitions for learning or evaluation.

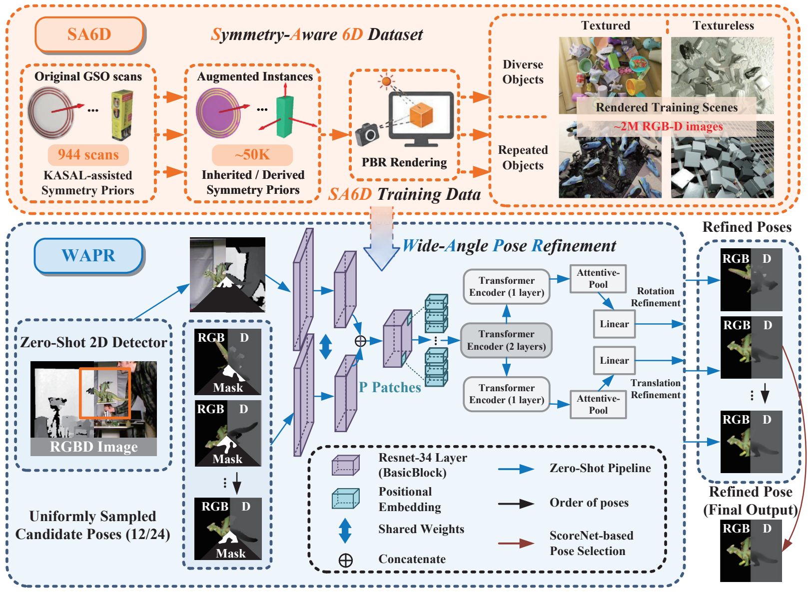

Fig. 2: Overview of the proposed framework. Top: SA6D constructs large-scale RGB-D training data with KASAL-assisted rotational symmetry priors. These priors are used to canonicalize symmetry-equivalent training pairs. Bottom: WAPR refines sparse candidate poses by comparing real and rendered RGB-D observations with masks. The refined candidates are selected by a score-based pose selection module, enabling accurate 6D pose estimation under large initial misalignments of up to 90°.

These works establish symmetry-consistent supervision as a standard principle in 6D pose estimation. Beyond loss design and evaluation, rotational symmetry priors are also useful as dataset-level annotations when synthetic pose pairs are generated for training. For wide-angle refinement, candidate and target poses may span symmetry-equivalent pose regions, and such pairs should receive consistent pose-update targets. Large-scale synthetic datasets have mainly focused on object diversity, rendering realism, and scalable pose supervision [27, 54]. SA6D complements this direction by adding KASAL-assisted rotational symmetry priors at scale and using them to canonicalize synthetic candidate–target pose pairs before training.

3 Methodology

Figure 2 summarizes the proposed framework. It consists of two connected parts. First, SA6D provides large-scale synthetic RGB-D training data with rotational symmetry priors. These priors are used to canonicalize symmetry-equivalent candidate–target pose pairs before training, so that wide-angle pose updates are well-defined for rotationally symmetric objects. Second, WAPR learns a render-and-compare refinement model that predicts rotation and translation updates

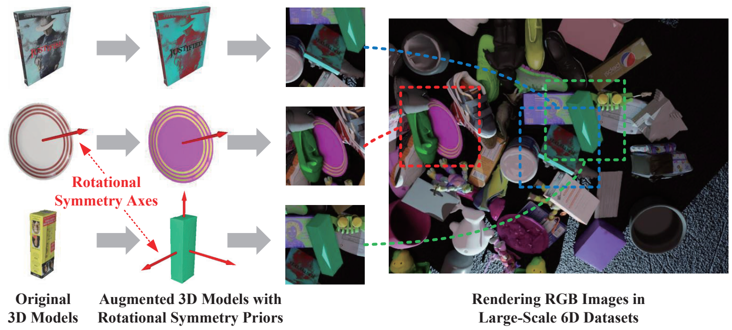

Fig. 3: Large-scale construction of SA6D with rotational symmetry priors. KASAL-assisted annotation provides texture-aware and geometry-only priors for the original GSO scans. Geometry scaling and texture randomization reuse these priors to generate ∼20K non-symmetric and ∼30K rotationally symmetric augmented instances, and BlenderProc [6] renders ∼2M synthetic RGB-D images for training.

from paired real/rendered RGB-D observations and masks. An angle-balanced loss further stabilizes training over a wide range of pose deviations, and sparse candidate generation enables efficient inference with only 12 or 24 candidates.

3.1 Symmetry-Aware 6D Dataset (SA6D)

SA6D is designed to train WAPR with synthetic RGB-D pairs whose pose targets remain consistent under rotational symmetry. It starts from 944 GSO scans [8], obtains KASAL-assisted rotational symmetry priors for these scans, and expands them through geometry and texture augmentation into ∼50K augmented object instances, including ∼30K rotationally symmetric instances. BlenderProc [6] is then used to render ∼2M synthetic RGB-D images.

Object Collection and Symmetry Augmentation
The construction of SA6D starts from the original GSO scans. For each scan, human annotators specify the rotational symmetry type and order, and KASAL [53] localizes the corresponding symmetry-axis direction and center. The annotated scans are then expanded into augmented instances by geometry scaling and texture randomization, while reusing the corresponding symmetry priors. This avoids repeatedly annotating each augmented instance.

SA6D uses two types of texture augmentation. The first modifies the texture while preserving the original texture-aware rotational symmetry prior, as illustrated by the disc in Fig. 3. The second replaces the texture with a uniform color, which removes texture-induced asymmetries and makes the augmented instance follow the geometry-only rotational symmetry prior, as shown by the carton in Fig. 3. Thus, each original scan is associated with both a texture-aware prior and a geometry-only prior. These priors are used in the offline canonicalization procedure described below to generate ambiguity-free candidate–target pose pairs for WAPR training.

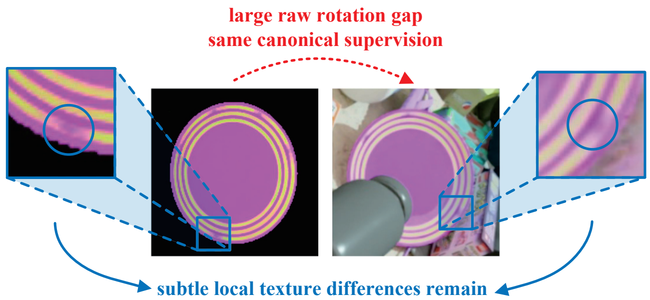

Fig. 4: Rotational symmetry ambiguity in wide-angle training. The candidate and target views can have a large raw rotation gap, while their visible difference may be limited to subtle local texture details, as shown by the zoomed regions. With rotational symmetry priors, such pairs are treated as symmetry-equivalent and canonicalized to the same supervision before loss computation.

Symmetry Disambiguation

To train WAPR, candidate poses are simulated by applying random rotations and translations to ground-truth poses in SA6D. For rotationally symmetric objects, multiple symmetry-equivalent poses may be visually valid for the same training pair. Directly using the raw perturbed rotation can therefore assign inconsistent pose-update targets to visually similar candidate–target pairs, as illustrated in Fig. 4. SA6D resolves this ambiguity by applying an offline min-over-symmetry canonicalization before loss computation. The procedure selects a canonical representative from the symmetry-equivalent poses, making the pose update used for training unique and prediction-independent.

Given a ground-truth pose, a set of augmented poses $\mathcal{P}_{\text{aug}}$ is generated by applying random perturbations to both rotation and translation:

$$\mathcal{P}_{\text{aug}} = \{(R_j, t_j) \mid j = 1, 2, \ldots, M\} \tag{1}$$

Each object is aligned so that its rotational symmetry center coincides with the origin, allowing symmetry transformations to be represented as pure rotations. For an object with n-fold discrete rotational symmetry, the symmetry group is represented by n rotation matrices. For continuous rotational symmetry, the symmetry orbit is discretized into 360 uniformly spaced rotations, which bounds the angular quantization error by $0.5°$. The symmetry-consistent rotation set is

$$\mathcal{R}_{\text{sym}} = \{\, R_k \in \mathrm{SO}(3) \mid k = 1, 2, \ldots, K \,\} \tag{2}$$

where each R_k represents a valid rotational symmetry of the object.

For each augmented pose (R_j, t_j), symmetry-extended variants are constructed as:

$$\mathcal{P}_{\text{sym}}^{(j)} = \{(R_j R_k, t_j) \mid R_k \in \mathcal{R}_{\text{sym}}\} \tag{3}$$

A canonical representative is selected by choosing the variant with the smallest rotation from the identity:

$$(R_{\mathrm{j}}^{*}, t_{\mathrm{j}}^{*}) = \arg\min_{(R,t)\in\mathcal{P}_{\mathrm{sym}}^{(j)}} \|R - I\|_2 \tag{4}$$

The selected pose (R_j^{*}, t_j^{*}) is used as the final augmented candidate pose for training. The target pose is canonicalized under the same symmetry prior, yielding a consistent candidate–target pair with a unique rotation update for training. This canonicalization is performed when constructing SA6D training pairs and does not depend on the network prediction.

Segmentation Mask Generation
WAPR also needs mask inputs during training. Instead of using perfect synthetic masks, which are cleaner than inference-time masks, SA6D generates training masks with a large-scale segmentation model such as SAM [25]. For each training sample, SAM produces candidate masks $\mathcal{S} = \{S_i\}_{i=1}^{N}$. Masks whose intersection-over-union (IoU) with the ground-truth mask G exceeds 0.5 are kept:

$$\mathcal{Q} = \{i \mid \mathrm{IoU}(S_i, G) > 0.5\} \tag{5}$$

If $\mathcal{Q}$ is non-empty, one mask is randomly sampled for training; otherwise, a blank mask is used. Together, symmetry-prior canonicalization and SAM-generated masks provide WAPR with well-defined pose targets and realistic mask inputs for wide-angle refinement training.

3.2 Wide-angle Pose Refinement

Neural Network

Given a candidate pose and its canonicalized target pose from SA6D, WAPR learns to predict the pose update by comparing two RGB-D observations. One branch takes the observed RGB-D crop and mask, which corresponds to a PBR-rendered sample during training and a real sensor observation during inference. The other branch takes the rendered RGB-D image and mask generated under the candidate pose. Two shared-weight ResNet layers based on BasicBlock-style residual units extract paired feature maps from the two branches.

The paired feature maps are concatenated and augmented with positional encodings. A two-layer Transformer encoder models global interactions in the fused representation, as illustrated in Fig. 2. Two independent one-layer Transformer encoders then produce separate representations for rotation and translation. Each branch applies attention-based pooling and a linear layer to regress the corresponding pose update.

Rotation updates are parameterized as 3D rotation vectors and converted to rotation matrices in SO(3) before being applied to candidate poses. For a candidate pose (R_i, t_i), the network predicts a rotation vector $\Delta\tilde{\mathbf{r}}_i \in \mathbb{R}^3$ and a translation offset $\Delta\tilde{t}_i \in \mathbb{R}^3$.

WAPR is trained and applied iteratively. During training, candidate poses are generated by perturbing the ground-truth poses with rotational deviations up to 90°. After each iteration, the predicted rotation and translation updates are applied to the current candidate pose, and the updated pose is used to re-render the input for the next iteration. The rotation and translation are updated by

$$\tilde{R}_i = \mathrm{Rot}(\Delta\tilde{\mathbf{r}}_i)R_i \tag{6}$$

$$\tilde{t}_i = \Delta\tilde{t}_i + t_i \tag{7}$$

where $\mathrm{Rot}(\cdot)$ maps a 3D rotation vector to a rotation matrix in SO(3).

The 90° perturbation range is used as a practical wide-angle training range for render-and-compare refinement. As analyzed in the supplementary material, the Visual Correspondence Ratio (VCR) between the candidate and target views decreases with angular deviation, and very low-VCR pairs provide limited shared visual evidence for learning reliable updates. Perturbations far beyond 90°, especially near 180°, also make training convergence less stable. Therefore, 90° provides a wide yet learnable range for WAPR training.

Angle-Balanced Loss

Wide-angle training contains both easy local refinements and hard large-error pairs. When the rendered and observed views have limited overlap, high-error samples can dominate the gradients while providing little useful correspondence. WAPR therefore uses an angle-balanced loss to emphasize learnable wide-angle cases and down-weight uninformative large-error samples. The prediction error of each sample is measured in rotation-vector space as the L1 residual magnitude between the predicted and target rotation-update vectors:

$$\theta_i = \|\Delta\tilde{\mathbf{r}}_i - \Delta\mathbf{r}_i\|_1 \tag{8}$$

Based on θ_i, a dynamic weight w_i is assigned to each training sample:

$$w_i = \exp\left(\sigma \cdot \min\left\{\theta_i, \tau - \theta_i\right\}\right) \tag{9}$$

where τ is empirically set to 0.25π. Samples with rotation-vector residuals below τ are treated as learnable wide-angle cases and receive larger weights, while samples with little overlap often yield $\theta_i > \tau$ and are down-weighted. The coefficient σ controls the suppression strength and is set to 3.

The final angle-balanced loss is then formulated as:

$$L = \alpha \cdot \frac{B \cdot \sum_{i=1}^{B} w_i \cdot |\Delta\tilde{\mathbf{r}}_i - \Delta\mathbf{r}_i|}{\sum_{i=1}^{B} w_i} + \beta \cdot \sum_{i=1}^{B} \left|\Delta\tilde{\mathbf{t}}_i - \Delta\mathbf{t}_i\right| \tag{10}$$

where B denotes the number of samples in the mini-batch, α and β are the balancing coefficients for the rotational and translational losses, respectively, and $|\cdot|$ denotes the L1 loss. We set $\alpha = 3$ and $\beta = 1$.

Inference Pipeline
WAPR's wide-angle refinement ability allows the inference pipeline to use a sparse set of candidate poses. Given a 2D bounding box and mask, candidate poses are generated by combining a small set of viewing directions with evenly spaced in-plane rotations. In the fast setting, 12 candidates are formed by pairing the 4 viewing directions defined by the vertices of a regular tetrahedron with 3 in-plane rotations. In the unconstrained setting, 24 candidates are formed by pairing the 8 viewing directions defined by the vertices of a regular cube with the same 3 in-plane rotations.

For each candidate, WAPR pairs the cropped RGB-D image and mask with rendered RGB-D and mask images generated under the candidate pose. In the unconstrained setting, all 24 candidates are refined for five rounds with re-rendered inputs at each round. In the fast setting, WAPR uses the two-stage schedule described in the supplementary material: all 12 candidates are first refined for two rounds and scored, after which only the top-4 candidates are refined for three additional rounds. Among the refined candidates, the pose with the highest confidence score predicted by the ScoreNet module of Foundation-Pose [54] is selected as the final result.

4 Experiments

This section evaluates WAPR through ablation studies and comparisons with state-of-the-art methods [3, 4, 37]. We first describe the experimental setup in Sec. 4.1, then analyze the main design choices in Sec. 4.2, and finally report BOP localization and detection results in Sec. 4.3.

4.1 Experimental setup

Network Settings
Both input branches of WAPR use a fixed resolution of 160×160. Each branch extracts features with a shared-weight encoder composed of a 7×7 convolution, a 3×3 convolution, and two ResNet basic blocks with 128 channels. The paired features are concatenated and processed by a fusion stage with 256- and 512-channel ResNet blocks, yielding a 512-dimensional fused feature representation. The Transformer encoder and attention-based pooling module both use four attention heads and a feed-forward dimension of 512. WAPR is trained for 800K iterations using Adam with a learning rate of 0.0002 and a batch size of 256. Each iteration samples RGB-D pairs from a single object category. Candidate poses are generated with translation offsets up to 50% of the object diameter and rotational deviations up to $90°$, followed by five-step iterative refinement.

Datasets & Metrics
The experiments follow the BOP 2024 single-view model-based unseen-object protocol, where comparable baselines and default 2D detections are available. We evaluate on seven BOP-Classic-Core datasets: LM-O [1], T-LESS [17], ITODD

Table 1: Analysis of candidate-pose coverage on IC-BIN. IC-BIN contains multiple identical objects in cluttered bin-like scenes, making it sensitive to the coverage of initial pose hypotheses. It is therefore used as a representative stress test for evaluating the number of candidate poses. "Num" denotes the number of candidate poses. Results are reported in AP for MSSD and MSPD, together with their mean.

Num	4	8	12	24	40	60
MSSD	59.6	67.3	71.4	71.9	72.8	**73.3**
MSPD	54.2	63.1	68.9	69.8	70.4	**70.9**
Mean	56.9	65.2	70.1	70.9	71.6	**72.1**

[9], HB [24], YCB-V [57], IC-BIN [7], and TUD-L [18]. Following the BOP protocol [19, 20], we use the average of VSD, MSSD, and MSPD as the BOP score. We report average recall (AR) for 6D localization and average precision (AP) for 6D detection.

Fast inference
For fast inference, WAPR uses the 12-candidate tetrahedron setting described in Sec. 3.2. For each MUSE detection, the associated SAM-based mask is used to estimate the initial translation by a depth-weighted average within the mask and the depth map. Detections are filtered using MUSE-predicted object labels and confidence scores with a threshold of 0.35. On an RTX 5090 system, this setting estimates poses for up to 25 detected object instances per second using the two-stage refinement and scoring schedule described in the supplementary material.

Unconstrained inference
When inference time is not constrained, WAPR uses the 24-candidate cube setting described in Sec. 3.2. When multiple 2D detectors are used, their detections are merged and NMS is applied to remove redundant boxes. For each retained detection, all category–detection–candidate combinations are evaluated without applying the confidence threshold used in the fast setting.

4.2 Ablation Study on Pose Refinement

This section analyzes the main factors affecting wide-angle pose refinement, including candidate-pose coverage, training inputs, symmetry priors, SA6D, angle-balanced optimization, network architecture, and plug-in refinement. The ablations in Table 2 are conducted on three representative datasets for controlled component analysis, while the final all-seven-dataset evaluation is reported in Table 4.

Number of Candidate Poses
Table 1 studies how candidate-pose coverage affects WAPR on IC-BIN. Increasing the number of candidates improves the chance that at least one hypothesis falls within the effective refinement range. The gain is large when increasing

Table 2: Ablations of WAPR components on BOP detection AP. Results are reported as AP for MSSD and MSPD on three representative datasets, together with their mean. IC-BIN, TUD-L, and LM-O are selected to cover crowded multi-instance scenes, large object motion and viewpoint variation, and heavy occlusion, respectively. "w/o" denotes "without." "WAPR (±90°)" is trained with candidate poses obtained by randomly perturbing the ground-truth pose within ±90°.

Row	Method	IC-BIN [7]		TUD-L [18]		LM-O [1]		Mean
		MSSD	MSPD	MSSD	MSPD	MSSD	MSPD	
A0	WAPR(±90°)	71.9	69.8	96.0	95.4	77.0	80.7	81.8
B0	WAPR(±20°)	61.5	58.0	77.4	76.5	68.0	71.0	68.7
B1	WAPR(±40°)	69.9	67.1	95.0	94.7	74.4	77.8	79.8
B2	WAPR(±60°)	71.7	68.9	95.6	95.1	74.9	78.3	80.8
B3	WAPR(±140°)	70.6	68.1	95.5	95.0	74.4	77.7	80.2
B4	WAPR(±180°)	63.3	60.2	89.6	89.2	72.9	76.2	75.2
C0	A0 + w/o mask input	71.0	68.5	95.2	94.9	75.6	78.9	80.7
C1	A0 + train with GT mask	61.8	59.4	94.3	93.9	73.9	77.2	76.8
D0	A0 + w/o sym prior	70.8	68.4	93.6	92.6	75.8	79.4	80.1
D1	A0 + w/o angle-balanced Loss	69.8	67.3	91.9	91.1	72.8	76.5	78.2
D2	D1 + w/o sym prior	61.8	58.0	91.0	91.0	69.3	73.0	74.0
D3	A0 on GSO PBR (w/o SA6D)	64.8	62.2	92.0	91.1	71.5	74.7	76.1
D4	D0 on GSO PBR (w/o SA6D)	63.9	61.1	90.6	89.7	69.3	72.6	74.5
E0	A0 + w avg pool	71.5	68.8	95.5	95.3	75.5	78.8	80.9
E1	E0 + w/o transformer	70.9	68.5	95.3	94.9	74.3	77.7	80.3

Row	Method	IC-BIN		TUD-L		LM-O		Mean	Time/s
		MSSD	MSPD	MSSD	MSPD	MSSD	MSPD		
F0	FreeZeV2 [3, 4]	73.4	70.6	98.6	99.1	78.5	81.9	83.7	27.8
F1	F0 + C0	74.9	72.4	98.7	98.9	78.6	82.1	84.3	27.8 (+0.54)
F2	GigaPose [39]+GenFlow [36]	42.0	38.4	71.4	71.7	56.1	59.5	56.5	4.5
F3	F2+C0	43.4	40.0	72.7	73.0	57.3	60.2	57.8	4.5 (+0.37)

from 4 to 12 candidates, where the mean AP improves from 56.9% to 70.1%. Further increasing the number to 24, 40, or 60 provides smaller gains, indicating that 12 candidates already provide a strong accuracy–efficiency trade-off for the fast setting, while 24 candidates are used in the unconstrained setting for higher coverage.

Angle Range for Training

Rows B0–B4 in Table 2 evaluate the rotation range used to generate candidate poses during training. Small perturbation ranges such as ±20° provide insufficient coverage for wide-angle errors, while overly large ranges such as ±180° introduce many low-correspondence samples and make optimization less stable. The ±90° setting achieves the best overall performance, consistent with the VCR and training-convergence analysis discussed in Sec. 3.2.

Mask Input Rows C0 and C1 evaluate the role of mask inputs. Removing the mask slightly degrades performance, showing that masks provide useful object-region guidance during render-and-compare refinement. Training with perfect ground-truth masks leads to a larger drop, since such masks differ from the predicted masks used at inference time. SAM-generated masks therefore provide a more realistic training signal for WAPR.

Rotational Symmetry Priors and SA6D Dataset Rows D0–D4 evaluate symmetry priors and SA6D through controlled comparisons. Rows D1 and D2 isolate the effect of rotational symmetry priors without angle-balanced loss, where adding the priors improves AP by 4.2%. Rows A0 and D0 show the effect when angle-balanced loss is used, where the priors improve AP by 1.7%. Rows D3 and D4 train the corresponding configurations on original GSO PBR data without SA6D augmentation. Compared with these GSO-PBR counterparts, A0 and D0 improve AP by 5.7% and 5.6%, respectively, showing the benefit of SA6D training data under matched model configurations.

Angle-Balanced Loss Function Rows D0–D2 analyze angle-balanced optimization. Compared with D2, which uses neither angle-balanced loss nor symmetry priors, D0 improves AP by 6.1%, showing that angle-balanced loss stabilizes wide-angle training. Combining angle-balanced loss with symmetry priors further improves performance, with A0 outperforming D2 by 7.8%.

Network Architecture Rows E0 and E1 ablate the feature aggregation modules. Replacing attention-based pooling with average pooling reduces AP by 0.9%, and further removing the Transformer encoder causes an additional 0.6% drop, resulting in a total 1.5% drop from A0. These results indicate that global feature interaction and attention-based aggregation are both useful for wide-angle render-and-compare refinement.

Plug-in Refinement Rows F0–F3 evaluate the mask-free WAPR variant C0 as a plug-in refiner for existing pose estimators. When applied to FreeZeV2 poses, C0 improves AP by 0.6% with a 0.54 s per-image overhead on top of FreeZeV2's 27.8 s runtime. When applied to GigaPose+GenFlow poses, C0 improves AP by 1.3% with 0.37 s additional time. These results show that WAPR-style wide-angle refinement can further improve strong existing estimators, although the remaining gain is naturally smaller when the initial estimates are already strong.

Comparison with Refinement Baselines Table 3 compares WAPR with representative RGB-D refinement baselines on LM-O and YCB-V. This comparison complements the full-pipeline BOP evaluation in Table 4 by focusing on refinement behavior under comparable settings. Compared with OSOP+ICP, WAPR is 4× faster and improves mean AR by 29.7%; compared with PPF/SIFT+Zephyr, it improves AR by 26.7%. Under the same hypothesis setting, replacing WAPR with ICP reduces AR by 54.2%, reflecting the sensitivity of ICP to large hypothesis errors. Compared with FoundationPose, which uses small-angle refinement with a maximum correction of 22.5°, WAPR improves AR by 8.3%.

Table 3: Comparison with representative RGB-D refinement baselines on LM-O and YCB-V, where comparable prior-refinement results are available. The last three rows use MUSE detections and 12 hypotheses per instance.

Method	LM-O	YCB-V	Mean	Time (s/img)
OSOP [47] + ICP	48.2	57.2	52.7	4
(PPF [10], SIFT) + Zephyr [41]	59.8	51.6	55.7	–
MegaPose-RGBD [27]	58.3	63.3	60.8	–
FoundationPose [54] (MUSE) + 12 hypotheses	67.4	80.8	74.1	–
ICP (MUSE) + 12 hypotheses	26.6	29.7	28.2	35
WAPR (MUSE) + 12 hypotheses	**75.5**	**89.3**	**82.4**	**1**

Table 4: Comparison with SOTA methods on seven BOP-Classic-Core datasets for 6D localization (AR, top) and detection (AP, bottom). "2D Det." reports the detector or hypothesis source. WAPR is trained on SA6D, while other methods use original training or released settings. "Multi-Det." combines NIDS [35], MUSE [5], SAM6D [32], and CNOS [38]. FP denotes FoundationPose [54]. "Time" reports per-image runtime, and "*" marks proposed methods.

Model-based 6D localization of unseen objects – BOP-Classic-Core

	Method	2D Det.	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V	Mean	Time (s)
fast	Co-op [37]	F3DT2D	73.0	68.0	92.9	62.4	60.0	86.3	88.6	75.9	0.765
	WAPR*	F3DT2D	**75.4**	72.0	92.5	66.3	65.0	86.7	**89.1**	78.1	0.904
	WAPR*	MUSE	**75.4**	**73.1**	**96.1**	**70.9**	**70.5**	**90.1**	88.3	**80.6**	0.72
unconstrained	MegaPose [27]	CNOS	62.6	48.7	85.1	46.7	46.8	73.0	76.4	62.8	141.965
	Genflow [36]	CNOS	67.8	55.6	81.1	56.3	57.5	79.1	82.5	68.6	11.14
	SAM6D [32]	SAM6D	69.9	51.5	90.4	58.8	60.2	77.6	84.5	70.4	4.367
	FP [54]	SAM6D	75.6	64.6	92.3	50.8	58.0	83.5	88.9	73.4	29.317
	Co-op [37]	F3DT2D	73.8	69.5	**92.9**	63.5	62.9	87.8	89.8	77.1	6.924
	WAPR*	CNOS	**76.8**	**82.6**	92.4	**70.7**	**74.3**	**89.1**	**91.1**	**82.4**	43.068
	FreeZeV2 [3,4]	Multi-Det.	77.7	79.9	**97.5**	71.1	75.4	89.5	**91.8**	83.3	21.709
	WAPR*	Multi-Det.	**78.1**	**81.7**	96.1	**73.5**	**79.8**	**91**	91.5	**84.5**	50.816

Model-based 6D detection of unseen objects – BOP-Classic-Core

	Method	2D Det.	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V	Mean	Time (s)
fast	Co-op [37]	F3DT2D	69.8	62.0	84.1	56.4	57.6	74.6	80.8	69.3	0.865
	WAPR*	F3DT2D	74.7	**64.5**	96.3	67.2	68.1	77.6	87.1	76.5	0.987
	WAPR*	MUSE	**75.6**	63.6	**98.6**	**71.8**	**71.5**	**81.6**	**87.6**	**78.6**	0.828
unconstrained	Genflow [36]	CNOS	59.7	56.5	68.8	43.9	42.5	70.7	60.7	57.5	15.486
	Co-op [37]	CNOS	68.3	59.6	80.8	46.9	56.0	74.3	78.2	66.3	2.202
	WAPR*	CNOS	**78.8**	**82.9**	**95.7**	**70.9**	**77.6**	**86.3**	**91.1**	**83.3**	43.068
	FreeZeV2 [3,4]	Multi-Det.	80.2	77.7	**98.8**	72.0	77.5	85.4	**91.1**	83.3	27.776
	WAPR*	Multi-Det.	**80.7**	**82.5**	98.7	**75.0**	**82.2**	**88.7**	89.6	**85.3**	50.816

4.3 BOP Results on 6D Localization and Detection

Table 4 compares WAPR with state-of-the-art methods on seven BOP-Classic-Core datasets for 6D localization and detection. In the unconstrained Multi-Det. setting, WAPR achieves 84.5% AR and 85.3% AP, improving over FreeZeV2 by 1.2% AR and 2.0% AP. This accuracy-oriented setting uses a larger candidate budget and obtains higher mean accuracy with a longer per-image runtime. The fast setting reports the efficiency-oriented operating point under the < 1 s per-image budget.

In the fast setting, WAPR with MUSE detections achieves 80.6% AR and 78.6% AP, improving over Co-op by 4.7% AR and 9.3% AP. For a matched-detector comparison, WAPR is also evaluated with the same F3DT2D detector used by Co-op under the same fast-inference protocol. This variant achieves 78.1% AR and 76.5% AP, improving over Co-op by 2.2% AR and 7.2% AP. These results show that the gain is preserved under the same detector, while the MUSE-based variant gives the highest mean performance among the evaluated fast-setting variants.

5 Conclusion

This paper presented WAPR, a zero-shot wide-angle pose refinement model for unseen-object 6D pose estimation. WAPR refines sparse candidate poses by comparing real and rendered RGB-D observations with masks, enabling accurate pose estimation under large candidate-pose deviations. To support wide-angle training, this paper further constructed SA6D, a large-scale synthetic RGB-D dataset with KASAL-assisted rotational symmetry priors. These priors are used to canonicalize symmetry-equivalent candidate–target pose pairs, while angle-balanced loss stabilizes learning across different angular ranges. Experiments on seven BOP core datasets show that WAPR achieves strong performance in both fast and unconstrained settings. These results demonstrate that wide-angle refinement, scalable training data with symmetry priors, and sparse candidate evaluation provide an effective solution for unseen-object 6D pose estimation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 52375487, the Special Fund of Jiangsu Province for Key Research and Development under Grant No. BE2023041, and the Doctoral Research Innovation Ability Improvement Program of Southeast University under Grant No. CXJH_SEU 26032. The authors sincerely thank the area chairs and anonymous reviewers for their careful reading, constructive feedback, and detailed suggestions.

References

1. Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J., Rother, C.: Learning 6D object pose estimation using 3D object coordinates. In: ECCV. pp. 536–551 (2014)
2. Capellen, C., Schwarz, M., Behnke, S.: ConvPoseCNN: Dense convolutional 6D object pose estimation. In: VISIGRAPP. pp. 162–172 (2020)
3. Caraffa, A., Boscaini, D., Hamza, A., Poiesi, F.: FreeZe: Training-free zero-shot 6D pose estimation with geometric and vision foundation models. In: ECCV. pp. 414–431 (2024)
4. Caraffa, A., Boscaini, D., Poiesi, F.: Accurate and efficient zero-shot 6D pose estimation with frozen foundation models (2025), arXiv preprint arXiv:2506.09784
5. Cho, S., Park, S., Oh, I.: Muse: Model-based uncertainty-aware similarity estimation for zero-shot 2d object detection and segmentation (2025), arXiv preprint arXiv:2510.17866
6. Denninger, M., Sundermeyer, M., Winkelbauer, D., Zidan, Y., Olefir, D., Elbadrawy, M., Lodhi, A., Katam, H.: Blenderproc (2019), arXiv preprint arXiv:1911.01911
7. Doumanoglou, A., Kouskouridas, R., Malassiotis, S., Kim, T.K.: Recovering 6D object pose and predicting next-best-view in the crowd. In: CVPR. pp. 3583–3592 (2016)
8. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: ICRA. pp. 2553 – 2560 (2022)
9. Drost, B., Ulrich, M., Bergmann, P., Hartinger, P., Steger, C.: Introducing MVTec ITODD — a dataset for 3D object recognition in industry. In: ICCVW. pp. 2200–2208 (2017)
10. Drost, B., Ulrich, M., Navab, N., Ilic, S.: Model globally, match locally: Efficient and robust 3d object recognition. In: CVPR. pp. 998–1005 (2010)
11. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: Exceeding YOLO series in 2021 (2021), arXiv preprint arXiv:2107.08430
12. Hai, Y., Song, R., Li, J., Salzmann, M., Hu, Y.: Rigidity-aware detection for 6D object pose estimation. In: CVPR. pp. 8927–8936 (2023)
13. Haugaard, R.L., Buch, A.G.: SurfEmb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings. In: CVPR. pp. 6739–6748 (2022)
14. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: ICCV. pp. 2980–2988 (2017)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
16. Hodan, T., Barath, D., Matas, J.: EPOS: Estimating 6D pose of objects with symmetries. In: CVPR. pp. 11700–11709 (2020)
17. Hodan, T., Haluza, P., Obdrzalek, S., Matas, J., Lourakis, M., Zabulis, X.: T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. In: WACV. pp. 880–888 (2017)
18. Hodan, T., Michel, F., Brachmann, E., Kehl, W., Buch, A.G., Kraft, D., Drost, B., Vidal, J., Ihrke, S., Zabulis, X., Sahin, C., Manhardt, F., Tombari, F., Kim, T.K., Matas, J., Rother, C.: BOP: Benchmark for 6D object pose estimation. In: ECCV. pp. 19–35 (2018)

19. Hodan, T., Sundermeyer, M., Drost, B., Labbe, Y., Brachmann, E., Michel, F., Rother, C., Matas, J.: BOP challenge 2020 on 6D object localization. In: ECCV. pp. 577–594 (2020)
20. Hodan, T., Sundermeyer, M., Labbe, Y., Nguyen, V.N., Wang, G., Brachmann, E., Drost, B., Lepetit, V., Rother, C., Matas, J.: BOP challenge 2023 on detection, segmentation and pose estimation of seen and unseen rigid objects. In: CVPRW. pp. 5610–5619 (2024)
21. Hong, Z.W., Hung, Y.Y., Chen, C.S.: RDPN6D: Residual-based dense point-wise network for 6DoF object pose estimation based on RGB-D images. In: CVPRW. pp. 5251–5260 (2024)
22. Hoque, S., Xu, S., Maiti, A., Wei, Y., Arafat, M.Y.: Deep learning for 6D pose estimation of objects — a case study for autonomous driving. Expert Syst. Appl. **223**(119838), 1–15 (2024)
23. Kalra, A., Stoppi, G., Marin, D., Taamazyan, V., Shandilya, A., Agarwal, R., Boykov, A., Chong, T.H., Stark, M.: Towards co-evaluation of cameras, HDR, and algorithms for industrial-grade 6DoF pose estimation. In: CVPR. pp. 22691–22701 (2024)
24. Kaskman, R., Zakharov, S., Shugurov, I., Ilic, S.: HomebrewedDB: RGB-D dataset for 6D pose estimation of 3D objects. In: ICCVW. pp. 2767–2776 (2019)
25. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: ICCV. pp. 3992–4003 (2023)
26. Labbe, Y., Carpentier, J., Aubry, M., Sivic, J.: CosyPose: Consistent multi-view multi-object 6D pose estimation. In: ECCV (2020)
27. Labbé, Y., Manuelli, L., Mousavian, A., Tyree, S., Birchfield, S., Tremblay, J., Carpentier, J., Aubry, M., Fox, D., Sivic, J.: MegaPose: 6D pose estimation of novel objects via render & compare. In: CoRL. pp. 1–10 (2022)
28. Lepetit, V., Moreno-Noguer, F., Fua, P.: EPnP: An accurate O(n) solution to the PnP problem. IJCV **81**, 155–166 (2009)
29. Li, S., Schieber, H., Corell, N., Egger, B., Kreimeier, J., Roth, D.: Gbot: Graph-based 3d object tracking for augmented reality-assisted assembly guidance. In: VR. pp. 513–523 (2024)
30. Li, Y., Wang, G., Ji, X., Xiang, Y., Fox, D.: DeepIM: Deep iterative matching for 6D pose estimation. In: ECCV. pp. 695–711 (2018)
31. Li, Z., Wang, G., Ji, X.: CDPN: Coordinates-based disentangled pose network for real-time RGB-based 6-DoF object pose estimation. In: ICCV. pp. 7678–7687 (2019)
32. Lin, J., Liu, L., Lu, D., Jia, K.: SAM-6D: Segment anything model meets zero-shot 6D object pose estimation. In: CVPR. pp. 27906–27916 (2024)
33. Liu, J., Sun, W., Liu, C., Zhang, X., Fu, Q.: Robotic continuous grasping system by shape transformer-guided multiobject category-level 6-d pose estimation. IEEE TII **19**(11), 11171–11181 (2023)
34. Liu, X., Zhang, R., Zhang, C., Wang, G., Tang, J., Li, Z., Ji, X.: GDRNPP: A geometry-guided and fully learning-based object pose estimator. PAMI **47**(7), 5742 – 5759 (2025)
35. Lu, Y., P, J.J., Guo, Y., Ruozzi, N., Xiang, Y.: Adapting pre-trained vision models for novel instance detection and segmentation (2024), arXiv preprint arXiv:2405.17859
36. Moon, S., Son, H., Hur, D., Kim, S.: GenFlow: Generalizable recurrent flow for 6D pose refinement of novel objects. In: CVPR. pp. 10039–10049 (2024)

37. Moon, S., Son, H., Hur, D., Kim, S.: Co-op: Correspondence-based novel object pose estimation. In: CVPR. pp. 11622 – 11632 (2025)
38. Nguyen, V.N., Groueix, T., Ponimatkin, G., Lepetit, V., Hodan, T.: CNOS: A strong baseline for CAD-based novel object segmentation. In: ICCVW. pp. 2126–2132 (2023)
39. Nguyen, V.N., Groueix, T., Salzmann, M., Lepetit, V.: GigaPose: Fast and robust novel object pose estimation via one correspondence. In: CVPR (2024)
40. Nguyen, V.N., Tyree, S., Guo, A., Fourmy, M., Gouda, A., Lee, T., Moon, S., Son, H., Ranftl, L., Tremblay, J., Brachmann, E., Drost, B., Lepetit, V., Rother, C., Birchfield, S., Matas, J., Labbe, Y., Sundermeyer, M., Hodan, T.: BOP challenge 2024 on model-based and model-free 6D object pose estimation. In: CVPRW (2025)
41. Okorn, B., Gu, Q., Hebert, M., Held, D.: Zephyr: Zero-shot pose hypothesis rating. In: ICRA. pp. 14141–14148 (2021)
42. Park, K., Patten, T., Vincze, M.: Pix2Pose: Pixel-wise coordinate regression of objects for 6D pose estimation. In: ICCV. pp. 7668–7677 (2019)
43. Park, K.B., Choi, S.H., Lee, J.Y.: Self-training based augmented reality for robust 3d object registration and task assistance. Expert Syst. Appl. **238**(122331), 1–16 (2024)
44. Peng, S., Zhou, X., Liu, Y., Lin, H., Huang, Q., Bao, H.: PVNet: Pixel-wise voting network for 6DoF object pose estimation. PAMI **44**(6), 3212–3223 (2022)
45. Qian, K., Erden, M.S., Kong, X.: Scalable network and adaptive refinement module for 6D pose estimation of diverse industrial components. In: IROS. pp. 11646–11653 (2024)
46. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. PAMI **39**(6), 1137–1149 (2017)
47. Shugurov, I., Li, F., Busam, B., Ilic, S.: Osop: A multi-stage one shot object pose estimation framework. In: CVPR. pp. 6825–6834 (2022)
48. Su, Y., Saleh, M., Fetzer, T., Rambach, J., Navab, N., Busam, B., Stricker, D., Tombari, F.: ZebraPose: Coarse to fine surface encoding for 6DoF object pose estimation. In: CVPR. pp. 6728–6738 (2022)
49. Thalhammer, S., Bauer, D., Hönig, P., Weibel, J.B., García-Rodríguez, J., Vincze, M.: Challenges for monocular 6-d object pose estimation in robotics. IEEE TRO **40**(1), 4065–4084 (2024)
50. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: ICCV. pp. 9626–9635 (2019)
51. Wang, G., Manhardt, F., Tombari, F., Ji, X.: GDR-Net: Geometry-guided direct regression network for monocular 6D object pose estimation. In: CVPR. pp. 16606–16616 (2021)
52. Wang, Y., Hu, M., Li, H., Luo, C.: HccePose(BF): Predicting front & back surfaces to construct ultra-dense 2D-3D correspondences for pose estimation. In: ICCV. pp. 7166–7175 (2025)
53. Wang, Y., Luo, C.: Key-axis-based localization of symmetry axes in 3d objects utilizing geometry and texture. TIP **33**, 6720–6733 (2024)
54. Wen, B., Yang, W., Kautz, J., Birchfield, S.: FoundationPose: Unified 6D pose estimation and tracking of novel objects. In: CVPR. pp. 17868–17879 (2024)
55. Wu, D., Zhuang, Z., Xiang, C., Zou, W., Li, X.: 6D-VNet: End-to-end 6DoF vehicle pose estimation from monocular RGB images. In: CVPRW. pp. 1238–1247 (2019)
56. Wu, Y., Javaheri, A., Zand, M., Greenspan, M.: Keypoint cascade voting for point cloud based 6DoF pose estimation. In: 3DV. pp. 176–186 (2022)
57. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. In: RSS. pp. 1–10 (2018)

58. Yang, X., Cai, J., Li, K., Fan, X., Cao, H.: A monocular-based tracking framework for industrial augmented reality applications. The International Journal of Advanced Manufacturing Technology **128**(5), 2571–2588 (2023)
59. Yu, S., Zhai, D.H., Guan, Y., Xia, Y.: Category-level 6-d object pose estimation with shape deformation for robotic grasp detection (2024), early access in *IEEE TNNLS*.
60. Zakharov, S., Shugurov, I., Ilic, S.: DPOD: 6D pose object detector and refiner. In: ICCV. pp. 1941–1950 (2019)
61. Zhang, H., Liang, Z., Li, C., Zhong, H., Liu, L., Zhao, C., Wang, Y., Wu, Q.M.J.: A practical robotic grasping method by using 6-d pose estimation with protective correction. IEEE TIE **69**(14), 3876–3886 (2022)
62. Zou, W., Wu, D., Tian, S., Xiang, C., Li, X., Zhang, L.: End-to-end 6DoF pose estimation from monocular RGB images. IEEE TCE **67**(1), 87–96 (2021)
63. Örnek, E.P., Labbé, Y., Tekin, B., Ma, L., Keskin, C., Forster, C., Hodan, T.: Foundpose: Unseen object pose estimation with foundation features. In: ECCV. pp. 163–182 (2024)