3D-Object Perception Transformer (3PT)

Agastya Kalra[1,2*†] Tim Salzmann[1*] Guy Stoppi[1] Dmitrii Marin[1] Rishav Agarwal[1]
Vage Taamazyan[1] Martin Bokeloh[1] Stefan Hinterstoisser[1] Anton Boykov[1]
Alberto Dall'Olio[1] Pravin Dangol[1] Kartik Venkataraman[1] Huaijin Chen[1,2†]
[1]Intrinsic Innovation LLC [2]University of Hawaii at Manoa

{agastyak, guystoppi, dmitriim, vage, kartikvp}@google.com

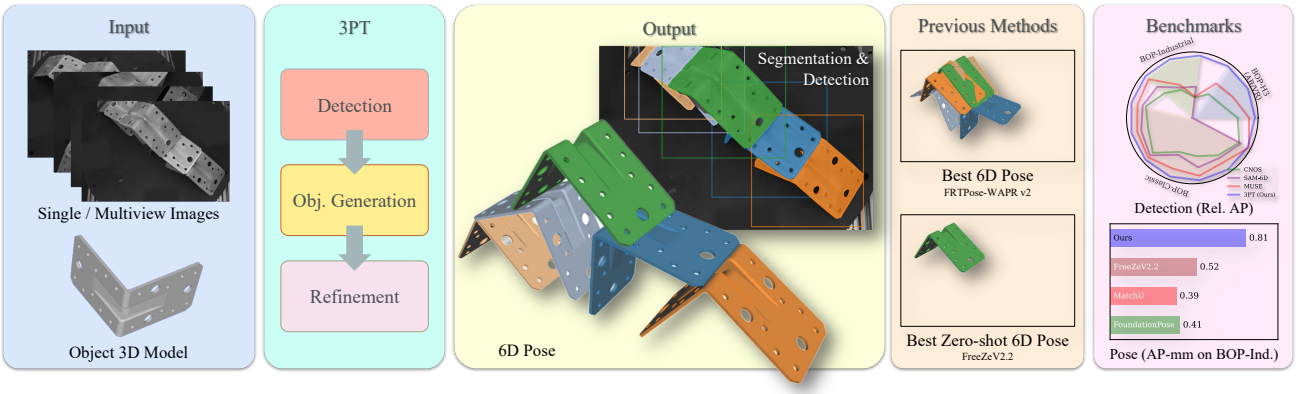


Figure 1. **3PT achieves state-of-the-art zero-shot 3D object perception**, including detection, segmentation, and 6DoF pose estimation. Unlike previous methods, 3PT successfully reconstructs poses in complex industrial scenes, such as the bracket *house-of-cards* structure shown above.

Abstract

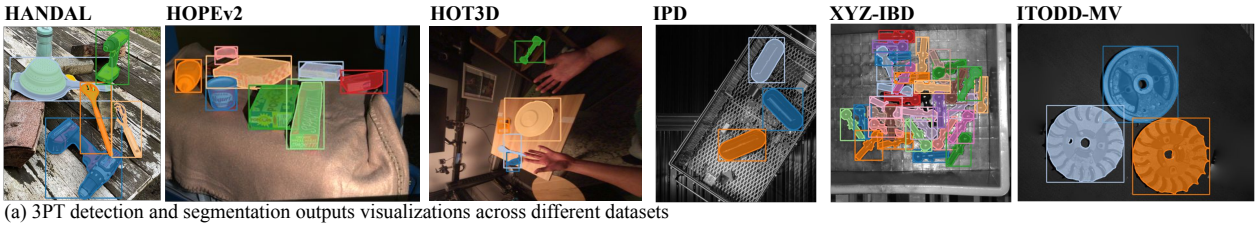
Current approaches to zero-shot 3D-object perception typically rely on ensembles of frozen foundation models. This limits deep object understanding and cross-domain generalization, making performance inadequate for real-world deployment. The 3D-Object Perception Transformer (3PT) addresses this limitation by unifying detection, segmentation, and 6DoF pose estimation in a single framework, directly trained for 3D-object perception. Based on two large-scale trained transformers that specialize in 2D and 3D object-centric scene understanding respectively, 3PT continuously refines its object representations without depth input, enhancing 3D understanding by incorporating multi-view information. 3PT surpasses task-specialized models for detection and pose estimation, often achieving double-digit percentage improvements on the diverse BOP-benchmarks, and in some cases outperforming non zero-shot methods. It also ranked first in 7 of 11 tracks at the BOP Challenge 2025. 3PT's high-accuracy and reliability is well-suited for practical industrial robotics applications such as bin picking and precise insertion.

Project Page can be found at https://www.intrinsic.ai/publications/3pt-cvpr2026

1. Introduction

Spatial understanding of 3D object instances in a scene is a fundamental challenge in computer vision. This understanding is typically demonstrated through tasks like object detection, segmentation, and pose estimation. These tasks become particularly challenging when geometric and visual descriptions of objects (3D-models) are only available at inference time or when per-object training is impractical due to continuously changing target objects and high-cost. This "zero-shot" object perception problem, where the model must reason about previously unseen objects from a reference representation, is critical for applications across multiple domains such as augmented reality, logistics, and industrial automation. 6DoF Pose estimation represents the most comprehensive task in object understanding, generating rich geometric and spatial understanding about objects in a scene. The conventional approach to pose estimation employs a two-stage pipeline: a *detection* stage

*Equal contribution †Work performed at Intrinsic with Intrinsic resources



(a) 3PT detection and segmentation outputs visualizations across different datasets

(b) 3PT pose outputs visualized as rendered edges. Zoom in to see alignment quality of our method.

Figure 2. **3PT generalizes zero-shot 3D-object perception** across camera types (fisheye, industrial, consumer), object types (household, industrial, dark), and environments (clutter, hand-held, and outdoor).

which localizes objects in the 2D image, followed by a *refinement* stage that estimates their 3D properties. This architecture reflects a hierarchical structure, where progressively more detailed, higher-dimensional representations are built upon the fundamental object understanding established at earlier stages aligned within the measurement evidence from the sensor(s). Existing detectors bypass genuine 3D geometric understanding by relying on frozen foundation models tailored for other vision tasks (SAM [34], GroundingDINO [39] etc.) to perform detection via visual similarity matching [10, 37, 45]. By operating on appearance similarity without fundamental understanding of the 3D object, these methods produce unreliable predictions when test objects diverge from their training distributions, for instance, when models pretrained on everyday web images encounter industrial objects. To mitigate this noise, state-of-the-art (SotA) approaches resort to multi-model ensembles within multi-stage pipelines [5, 59]. However, these heuristic-based, decoupled architectures fundamentally limit generalization due to their rigid inter-model interfaces and lack of task-specific finetuning. This limitation is evident in the continued superiority of *per-object trained* models over pipelined zero-shot approaches that use multiple frozen models [59].

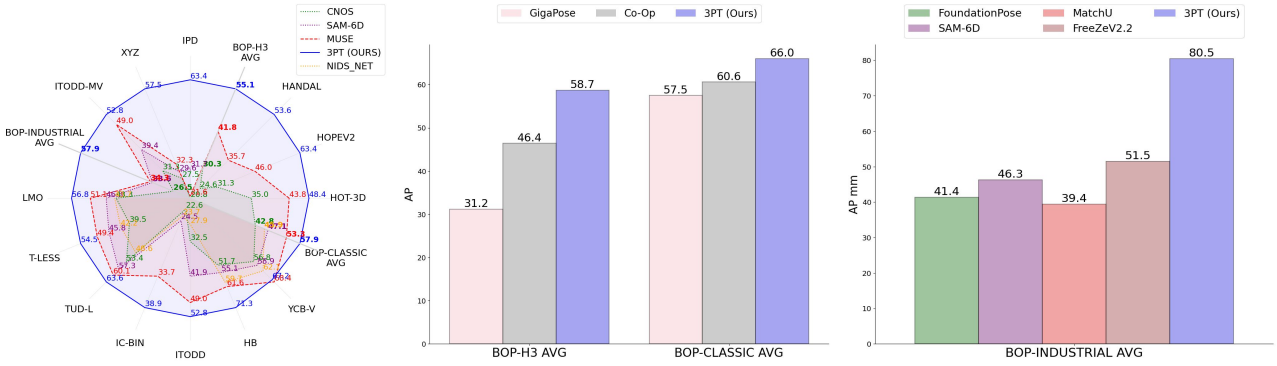
The second significant limitation of existing zero-shot pipelines is their reliance on depth data for accurate pose refinement. This limitation requires depth to be estimated from multi-view correspondence matching, either between multiple cameras or for industrial datasets [18, 30, 64], between a camera and a projector. Errors in these correspondences are easily introduced by occlusions, harsh lighting, or reflective surfaces and further propagate directly to the quality of the estimated pose. Eliminating this dependency on processed evidence by relying solely on raw multi-view RGB data is essential to prevent avoidable degradation.

Addressing these limitations, we introduce the 3D-Object Perception Transformer (3PT), a multi-stage framework for deep visual object understanding. Unlike existing approaches that rely on frozen foundation models for detection and similarity matching, 3PT's *detection* stage (3PT-D) is trained at scale, directly conditioned on 3D object models. This "early-fusion" design allows the model to jointly learn object representations and leverage this understanding to localize objects within images through a unified process, rather than separately embedding objects and image before "late-fusing" them through similarity matching. Consequently, 3PT exceeds the performance of prior methods on $12/13$ detection datasets, with 68.8% and 31.8% relative improvements on BOP-Industrial and BOP-H3 respectively (Fig. 3). 3PT's *refinememt* stage (3PT-R) requires only a single RGB image and does not rely on depth data; but, when available, natively incorporates multiview images to further improve 6DoF pose estimation. On the T-LESS [25] dataset, our zero-shot multi-view RGB method outperforms all prior supervised RGB methods, without training on the test objects. 3PT also achieves substantial improvements over prior methods across datasets spanning diverse object types: It achieves a 56.5% relative improvement on BOP-Industrial and a 26.5% relative improvement on the everyday-objects of BOP-H3. Beyond benchmarks, we demonstrate 3PT's industrial-grade accuracy and robustness in practical applications, including DIMM-insertion and sheet metal bin picking, with over 100 successful and repeatable real-world attempts in dense clutter Fig. 4b. We present comprehensive ablation studies and analysis to understand which design choices impact 3PT's performance.

2. Background and Related Work

Zero-Shot 2D-Detection SotA approaches for zero-shot 3D-object conditioned detection have converged on a two-stage **propose-and-match** workflow leveraging generic vision transformers.

In the *proposal* stage, methods generate class-agnostic



a) Combined BOP Benchmarks (2D Detection) b) BOP-H3 & BOP-Classic (Zero-Shot RGB 6DoF Pose) c) BOP-Industrial (Zero-Shot 6DoF Pose)

Figure 3. **3PT (blue) outperforms prior zero-shot methods** across tasks, objects, and environments achieving SotA in all three BOP-benchmarks for detection and single/multi-view RGB 6DoF pose estimation.

proposals using segmenters like FastSAM [65] (used by CNOS [45], SAM-6D [37]) or open-vocabulary detectors like GroundingDINO [39] (used by MUSE [10], F2D3D [9], NIDS-Net [40]). In the subsequent *match* stage, proposals are classified by comparing their features against template renders with complex heuristics. Most approaches (CNOS, MUSE, NIDS-Net) rely on frozen DINOv2 [48] features with heuristic matching, while SAM-6D [37] trains a lightweight custom matcher. Unlike these ensemble-based approaches, which depend on the understanding of generic foundation models trained for different tasks, we introduce a native vision transformer directly trained on 3D-object-conditioned detection.

Zero-Shot 6D-Pose Estimation The standard zero-shot pipeline follows a sequential *Initialize-Refine-Score* paradigm. Current SotA methods typically treat these stages as disjointed tasks, often requiring a combination of custom heuristics and specialized models.

Initialize: Most methods rely on depth [61] or bounding box priors [44, 46] for translation initialization. For rotation initialization, some methods perform dense rotation sampling [61] and postpone the decision of which sample to take to the scoring stage. Others utilize template matching [20, 46, 54, 66], or correspondence-based methods [29, 37, 44]. In contrast, our method receives the orientation information directly from the detection stage 3PT-D.

Refine: FoundationPose [61] employs a direct-update regression network. MegaPose [36] established the iterative render-and-compare paradigm for zero-shot objects, which GigaPose [46] and PicoPose [38] further accelerated using dense pixel-correspondence, and Co-op [44] using optical flow. FreeZe [5] and MatchU [29] perform point cloud based 3D matching augmented by learned visual and geometric features. Some RGB-D methods apply some variant of ICP [2] as a final refinement step [5, 29].

Score: Scoring is often utilized to filter hypotheses after refinement. Some methods use handcrafted scoring functions (geometric distance [5] or feature similarities [37]) while others require a separate network [61]. In contrast, we introduce a unified approach *3PT-R* that jointly predicts (1) edge-based reciprocal correspondences for pose updates, (2) segmentation masks, and (3) hypothesis confidence scores in a single forward pass.

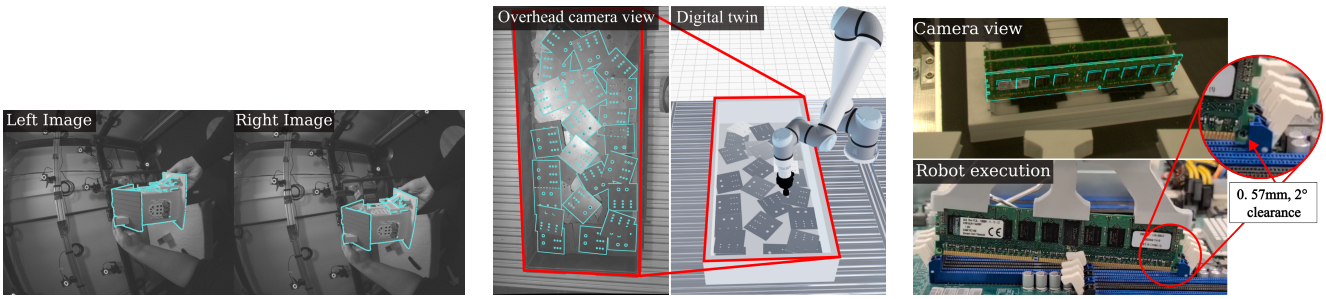
High-Precision Pose Ahead of all published methods, the current SotA submission on the online BOP-Industrial leaderboard [47] for high-precision pose is FRTPose-WAPRv2 [59]. It ensembles 4 detection approaches and 3 refinement approaches on high resolution RGB with >2 minute runtime to achieve maximum possible accuracy. For pose initialization, it trains a new model per object for 5 minutes, making it non zero-shot. We include it for completeness due to its high performance on the BOP-Industrial leaderboard.

AP vs. AR While many classical pose estimators [36, 66], and recent Diffusion [20, 31, 58], Gaussian-Splatting [3], and model-free [23, 56] methods report Average Recall (AR), this metric implicitly assumes known object counts and ignores false positives. In contrast, we evaluate on Average Precision (AP). This is a stricter metric that penalizes hallucinations and measures the reliability of the full pipeline (detection and pose estimation) in the wild, rather than just the pose accuracy of a pre-selected object crop.

Multi-View Pose While multi-view RGB methods exist [35, 49], they require per-object training rather than zero-shot. Model-free approaches [23, 56] use multi-view for novel object onboarding, but single-view scene-level inference. 3PT is focused on zero-shot multi-view RGB 6DoF pose in real scenes.

3. 3D Object Perception

3D object perception is decomposed into three sequential stages: *Detection* (instance identification from RGB images), *Object Generation* (proposal creation), and *Refine-*



(a) Multi-View Pose Tracking with AR Glasses (b) Long-Range Cluttered Bin-Picking (c) Sub-mm precise DIMM insertion

Figure 4. **3PT generalizes to challenging real-world applications in industrial robotics and augmented reality.** Demonstration videos are provided as supplementary materials.

ment (iterative optimization, selection of the best proposal based on image evidence, and generation of final segmentation maps and 6D poses).

3.1. Detection Stage

We introduce *3PT-D*, a two-tower Vision Transformer (ViT [15]) architecture adapted from recent encoder-only dense object detection methods [41, 42, 52]. Unlike prior works which process text embeddings, we introduce a 3D-object conditioned architecture that ingests rendered CAD views to perform detection and coarse orientation estimation. This two-tower architecture consists of an Object Encoder (Tower 1) and an Image Encoder (Tower 2) (see Fig. 5).

Object Encoder (Tower 1) The object encoder processes the 3D-model to generate two distinct sets of embeddings, which can be pre-computed and cached:

1. *Prior Embeddings (Condition):* We render the object in 12 pre-specified rotations. These are forwarded through the backbone and the resulting embeddings are pooled and concatenated to form the context embeddings used to condition the image encoder.
2. *Query Embeddings (Templates):* We generate N_t renders ($N_t = 5140$) densely covering the rotation space. These are encoded into a cache of template vectors used for orientation matching.

Image Encoder (Tower 2) The image encoder receives the input image tokens concatenated with the object *Prior Embeddings* (Early Fusion). This conditions the ViT to transform image patches into object-specific hypothesis embeddings. Each hypothesis is decoded by regression heads to predict a 2D bounding box and a confidence score indicating the presence of the object. We estimate the orientation of an object hypothesis by computing the cosine distance between the predicted hypothesis embedding and the cached *Query Embeddings* as a single matrix-multiply. The detection score is defined as the maximum similarity score over all orientation templates. This leads to three outputs: *Bounding Boxes, Detection Scores, and Orientation Distribution*.

In the case of multiple query objects, we run the model once per object, and then perform non-maximum suppression across boxes.

Scale-Range Prior To address the trade-off between resolution and computational cost, we employ a dynamic pan-and-scan strategy, where we tile the image into overlapping 512×512 patches. During training, we randomly crop and resize regions to ensure target objects fall within a specific pixel range (128–350px). This embeds a scale prior in the model. At inference time, we assume a defined depth range, enabling us to resize images such that the target objects fall within this specified size range.

While this prior is practical for real industrial applications, BOP benchmarks do not support the assumption of a fixed depth range. In that case, we run the model S times where $S = 3$ to cover a sufficiently large depth range. We ablate this in Tab. 3e.

Training The model is trained contrastively on $\sim$1B image-render pairs to align hypothesis embeddings with their corresponding rendered orientation templates using the *sigmoid focal cross-entropy* loss. Bounding box grounding is optimized by reducing both *L1* and *GiOU* [51] losses.

Soft-Scaled Matching Loss Standard matching-based losses [6, 41, 42] use a winner-takes-all approach, treating non-best matches as hard negatives regardless of their spatial overlap, leading to noisy gradients. 3PT-D instead down-weights the penalty for near-miss hypotheses inversely to their bounding box matching score (IoU). This stabilizes training by focusing the model on distinguishing hard negatives from true positives.

3.2. Object Generation Stage

This stage inputs bounding boxes and rotation distributions, and groups them by object identity. Each group is called *object*. Object i contains boxes belonging to the same instance (at most one per view) and K pose hypotheses $P_{i1}..P_{ij}..P_{iK}$.

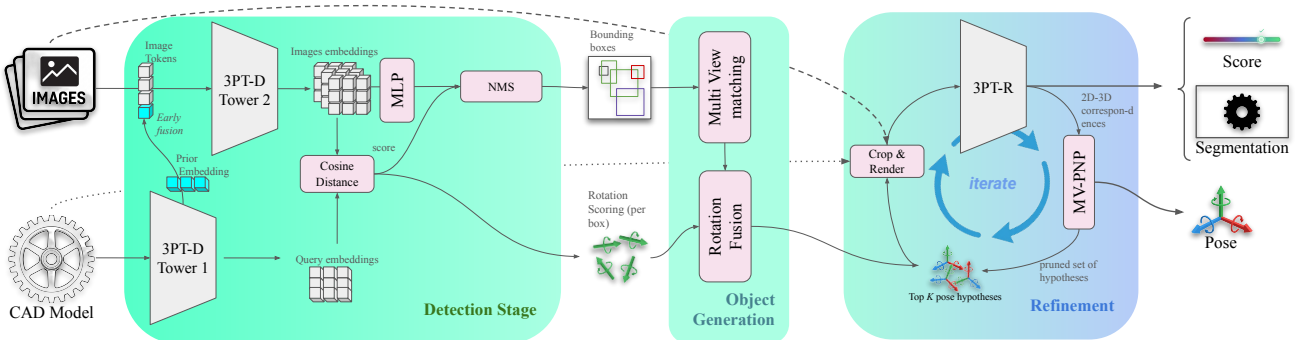


Figure 5. **Architecture overview of 3PT**. Details are given in Sec. 3.

When only a single RGB image is available, we associate each bounding box to an object. The z coordinate is estimated from the ratio of the bounding box diagonal and the CAD diagonal. The x, y coordinates are estimated by projecting the centroid to the estimated object z. The rotations are estimated by sampling the top K rotations that are at least $45°$ apart.

In the multi-view case, we use *epipolar matching* to convert 2D object detections from multiple images into 3D hypotheses and filter out outliers. We match along rays cast through the centers of the detected bounding boxes by triangulating the object's position in 3D space. An object is considered valid and is kept if it is successfully matched in $v \geq 3$ views.

For a single object, each estimated 3D position has vN_t associated orientation proposals. These can be noisy because a single camera view may have visual ambiguities. We fuse and filter the proposals from all available views as follows. We model the continuous distribution of orientations using a *Kernel Density Estimation* (KDE) over the rotation group $\mathrm{SO}(3)$. The kernel function used for this estimation is an isotropic *von Mises-Fisher* distribution [19]. To speed it up, we only use the top 500 rotations from each view. We re-weight the scores of all per-object orientation proposals using the KDE distribution. From these, we select the top K highest-scoring proposals, ensuring they maintain a minimum distance from each other. This yields K fused $\mathrm{SE}(3)$ pose hypotheses for each object.

3.3. Refinement Stage

The Refinement stage (3PT-R) generates outputs towards object-centric tasks, specifically segmentation, pose estimation, and scoring.

The input to this stage consists of V calibrated images and M object instances, where each instance represents a detected object of interest with K initial pose hypotheses. For each object instance, 3PT-R processes these inputs to produce V segmentation maps (one per image), a single refined 6DoF pose, and a single pose confidence score. Although the primary goal is pose estimation, segmentation outputs are retained for downstream tasks like bin-picking (*e.g.*, determining top-layer parts).

Refinement Methodology and Inference We formulate pose refinement as an iterative *render-and-compare* loop. During inference, this loop typically runs for 3 iterations to converge on a single best pose hypothesis. Each iteration consists of the following steps for every object and active hypothesis:

1. *Forward Pass:* We generate image-render pairs for current pose hypotheses and pass them through the 3PT-R network described below. This yields hypotheses confidence scores, segmentation masks, and dense real image feature map F and the rendered image feature map Z.
2. *Reciprocal Matching:* To estimate correspondences between the real and rendered images, we employ reciprocal matching. This technique finds pairs of pixels (i, j) in F and (x, y) in Z that are mutual nearest neighbors in feature space (based on cosine similarity). To reduce the computational load of high-resolution dense matching, we restrict this matching to the edges of the rendered image.
3. *Pose Update:* Sparse 2D-3D correspondences are obtained from the matched features (as every rendered pixel has a known 3D location). We then apply multi-view Perspective-n-Point (MV-PnP) [11] to minimize reprojection error and calculate the pose update for the next iteration.
4. *Hypothesis Selection:* Unlike methods that only output a final pose, our model estimates a confidence score for every hypothesis at each iteration. We use this score to filter out the bottom 50% of hypotheses after each iteration, progressively narrowing down to the best candidate.

Architecture

Encoding: The original images are cropped around the object using detection bounding boxes and resized to fixed dimensions. The object CAD model is rendered at each of the K pose hypotheses onto each of the V cropped views.

Both the real image crops and the rendered views are independently embedded using the same DINOv2 [48] encoder, resulting in two sets of embeddings.

Cross-Attention: To register object hypotheses with visual image evidence, we cross-attend the embeddings from each real view (query) with all render embeddings (key/value) similar to Co-Op [44].

Decoding: The resulting token sequences are processed by a DPT [50] decoder to predict dense outputs, specifically per-pixel descriptors (feature maps) and segmentation masks. The DINOv2 [48] register token is used to predict the scalar classification score for each hypothesis.

Training The network is trained on synthetic data where ground truth poses are known. Training hypotheses are generated by perturbing these poses. The training objective comprises three components: a segmentation loss, a classification loss, and a matching loss. The matching loss enforces high cosine similarity between feature maps for ground truth correspondence pairs while minimizing similarity for non-corresponding pairs.

Matching Loss: Let $P \in \mathrm{SE}(3)$ be the current hypothesis. Mapping $f : (x, y|P) \mapsto (i, j)$ projects a 3D point from the rendered feature map Z at pixel (x, y) to its corresponding pixel (i, j) in the real image feature map F using the ground truth pose. The matching loss

$$\mathcal{L}_m(\theta|P) \quad = \sum_{(x,y)\in \mathrm{Seg}[Z]} - \log p_{f(x,y|P)} \qquad (1)$$

where θ are trainable parameters, p_{ij} is the softmax of the dot product $Z_{xy} \cdot F_{ij}$.

Classification Loss: 3PT-R outputs confidence scores per render-image pair. During training, we sample 5 randomly jittered initial poses and generate render-image pairs for each. We then run one step of refinement, and train the scorer to estimate the render-image pair with the lower error using softmax and crossentropy loss.

Segmentation Loss: For segmentation, we use a standard binary cross entropy per-pixel.

3.4. Data Curation

We train both networks on 900,000 unique synthetic images, created using a Blender-based data generation pipeline. In the dataset, we used over 100,000 mesh files [12, 14, 17, 62], arranging them to simulate real challenges the networks would have to solve. On average, there are 110 annotated instances in each image, *i.e.* the dataset has approximately 100 million unique object instances before data augmentation. The synthetic scenes in the dataset are object-based and most are similar to NVIDIA's Synthetica [55], with meshes dropped onto a surface with a random background. In addition, some scenes showed either objects floating in the air or tightly packed together. More information about the dataset is in the supplement.

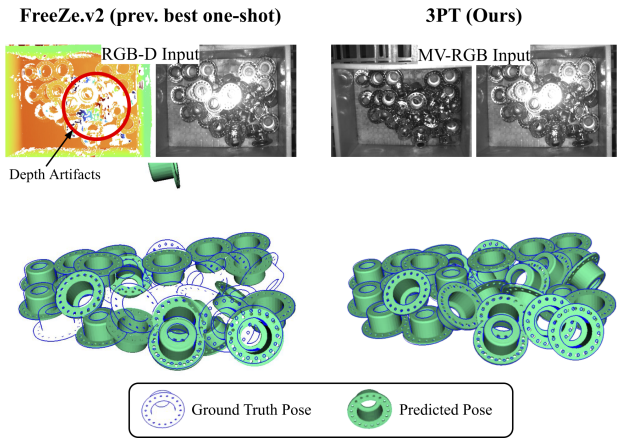


Figure 6. **3PT achieves accurate 6D pose estimation in densely cluttered scenes** under challenging lighting conditions where existing RGB-D methods fail.

4. Experiments

We evaluate 3PT extensively across 13 datasets from the *Benchmark on 6DoF Object Pose* (BOP) [26]. While prior work focuses primarily on the saturated **BOP-Classic** suite (LM [24], YCB-V [63], etc.), we jointly present results on the recently introduced **BOP-H3** [1, 21, 57] and **BOP-Industrial** [18, 30, 32] benchmarks. These datasets represent the frontier of 3D object perception, where existing zero-shot approaches typically struggle with performance drops of $10 - 30\%$.

- **BOP-H3** (e.g., HOPE, HOT3D) targets Augmented Reality challenges: lack of depth sensors, fisheye distortion, and hand occlusions.
- **BOP-Industrial** (e.g., ITODD, CNOS-Fast) targets robotic bin-picking: featuring severe clutter, metallic reflections, and harsh lighting.

Crucially, these modern suites utilize fully *hidden* test sets and high-fidelity ground truth, preventing overfitting and enabling precise evaluation. We evaluate exclusively on Average Precision (AP) and AP-mm (AP at $2 - 20$mm thresholds), excluding methods only evaluated on Average Recall (AR) [31, 38, 43, 66] as they assume known object counts, which conceals false positives.

We focus on BOP-benchmarks due to wide adoption and hidden ground truth compared to other datasets [7, 8, 60].

4.1. Main Results

Our method establishes new **state-of-the-art** detection results on 12/13 benchmarks, RGB pose on 9/10 benchmarks, and industrial mm-level pose on 3/3 benchmarks, with significant improvements on BOP-H3 ($+13.3$ detection AP, $+12.3$ pose AP) and BOP-Industrial ($+23.6$ detection AP and $+29.1$ pose AP-mm).

Table 1. **Main results:** Comprehensive detection and pose results across BOP benchmarks. Baseline numbers were taken from BOP Online Leaderboard [27]. (Δ) represents per-column improvement over the best competitor.

(a) **Detection AP:** Our detection demonstrates superior performance generalization, especially on industrial datasets, with an average 23.6 AP improvement over the previous SotA in BOP-Industrial and 13.3 AP improvement in BOP-H3.

Method	BOP-H3				BOP-Industrial				BOP-Classic							
	HOT-3D[1]	HOPEv2[57]	HANDAL[21]	Avg	IPD[32]	XYZ[30]	ITODD-MV[18]	Avg	LMO[4]	T-LESS[25]	TUD-L[28]	IC-BIN[16]	ITODD[18]	HB[33]	YCB-V[63]	Avg
CNOS [45]	35.0	31.3	24.6	30.3	20.8	27.5	31.3	26.5	43.3	39.5	53.4	22.6	32.5	51.7	56.8	42.8
SAM-6D [37]	–	–	–	–	31.7	29.6	39.4	33.6	46.3	45.8	57.3	24.5	41.9	55.1	58.9	47.1
MUSE [10]	43.8	46.0	35.7	41.8	21.6	32.3	49.0	34.3	51.1	49.4	60.1	33.7	49.0	61.6	**68.4**	53.3
3PT-D	**48.4**	**63.4**	**53.6**	**55.1**	**63.4**	**57.5**	**52.8**	**57.9**	**56.8**	**54.5**	**63.6**	**38.9**	**52.8**	**71.3**	67.2	**57.9**
Δ	**+4.6**	**+17.4**	**+17.9**	**+13.3**	**+31.7**	**+25.2**	**+3.8**	**+23.6**	**+5.7**	**+5.1**	**+3.5**	**+5.2**	**+3.8**	**+9.7**	-1.2	**+4.6**

(b) **Single-View RGB 6DoF Pose on BOP-H3 and BOP-Classic:** 3PT is SotA on 9/10 datasets and shows substantial gains on BOP-H3 (+12.3 average AP).

Method	BOP-H3 (AP)				BOP-Classic (AP)							
	H3D[1]	Hv2[57]	HDL[21]	Avg.	LMO[4]	T-L[25]	TUDL[28]	ICB[16]	IT[18]	HB[33]	YCBV[63]	Avg.
GigaPose [46]	26.8	41.1	25.6	31.2	59.7	56.5	68.8	43.9	42.5	70.7	60.7	57.5
Co-Op [44]	40.1	51.4	47.7	46.4	63.7	61.6	65.8	**46.7**	50.4	73.2	62.6	60.6
3PT (D+R)	**55.3**	**55.1**	**65.7**	**58.7**	**64.2**	**77.5**	**71.8**	37.1	**64.4**	**79.6**	**67.5**	**66.0**
Δ	+15.2	+3.7	+18.0	+12.3	+0.5	+15.9	+3.0	-9.6	+14.0	+6.4	+4.9	+5.4

(c) **6DoF Pose on BOP-Industrial (AP-mm):** Our depth-free pose estimation achieves a 29.1 AP-mm improvement over the previous best.

Method	0-Shot	Modality	IPD[32]	XYZ[30]	IT.-MV[18]	Av.
FRTPose [59]	×	MV-RGB-D	74.4	64.5	65.9	68.3
FoundationPose [61]	✓	RGB-D	30.4	48.3	45.5	41.4
SAM-6D [37]	✓	RGB-D	31.6	52.8	54.4	46.3
MatchU [29]	✓	RGB-D	34.0	40.9	43.4	39.5
FreeZeV2.2 [5]	✓	RGB-D	32.1	51.9	70.5	51.5
3PT (D+R)	✓	MV-RGB	**92.3**	**72.0**	**77.5**	**80.6**
Δ			+58.3	+19.2	+7.0	+29.1

Table 2. **Our *zero-shot* MV-RGB method outperforms prior *supervised/seen* MV-RGB(D) methods on AR for T-Less.**

Method	Modality	# Views	Zero-Shot	T-LESS (AR)
CenDerNet [13]	MV-RGB	5	×	71.3
DPODv2 [53]	MV-RGB	4	×	74.2
Haugaard et al. [22]	MV-RGB-D	4	×	85.8
3PT (D+R)	MV-RGB	4	✓	**90.0**
Δ				+4.2

2D Detection: We compare 3PT-D against the leading zero-shot detection approaches CNOS[45], SAM-6D[37], and MUSE[10] in Tab. 1a. 3PT-D significantly outperforms all priors, achieving AP gains of +13.3 (BOP-H3), +23.6 (BOP-Industrial), and +4.6 (BOP-Classic [4, 16, 18, 25, 28, 33, 63]). These results substantiate the enhanced generalization capability of 3PT's 3D-object conditioned training relative to ensemble-based approaches. 3PT-D currently ranks first for 2D detection of unseen objects on BOP-H3, BOP-Classic, and BOP-Industrial leaderboards.

RGB-Only 6DoF Pose: In Tab. 1b, we compare -D + -R against SotA methods Co-op [44] and GigaPose [46]. yields substantial gains in the generalized setting (unknown object counts), improving relative AP by +12.3 on BOP-H3 and +5.4 on BOP-Classic. 3PT-D currently ranks first for 6D detection of unseen objects on BOP-H3. It also ranks first on BOP-Classic against all RGB-only methods.

mm-level 6DoF Pose: Tab. 1c benchmarks our multi-view pose estimation against strong zero-shot RGB-D baselines (FoundationPose [61], SAM-6D [37], FreeZeV2 [5]) and the prior best non-zero-shot method FRTPose-WAPRv2 [59]. Remarkably, using *only RGB images*, 3PT outperforms all zero-shot RGB-D approaches by an average of 29.1 AP-mm, even per-object-fine-tuned RGB-D methods by 12.3 AP-mm. 3PT-D currently ranks first for 6D detection of unseen objects and seen objects on BOP-Industrial, even though it is zero-shot.

Supervised MV-RGB: We include a comparison on T-LESS [23] to supervised MV-RGB methods in Tab. 2. Even though they are **supervised/seen** methods that train directly on the test objects, our **zero-shot** method **outperforms** them by +4.2 AR.

Additional Challenges: We achieve SotA on zero-shot segmentation (53.6 AP) for BOP-Classic and model-free detection (48.2 AP) on BOP-H3 and currently rank first on both leaderboards. Details are provided in the supplement.

4.2. Analysis and Discussion

We analyze our methods on BOP-H3 [1, 21, 57] and BOP-Industrial [18, 30, 32], as these datasets offer the most reliable ground truth annotations for evaluation.

2D Detection: *Native 3D-conditioning at scale > generic ensembles of ViTs.* Previous zero-shot approaches leverage ensembles of generic ViTs (e.g., MUSE ensembles DinoV2, GroundingDINO, and SAM). Tab. 3b demonstrates that while powerful, these models are suboptimal at every component of 3D-object conditioned detection. Adding our classification, scoring, and regression separately to MUSE's predictions each improve performance. However, just as training a robust text-conditioned detector (e.g., OwL-ViT) requires massive vocabulary exposure, 3D-conditioned detection requires similar scale. We train on **76.8M images** and **110k unique 3D object models**, creating the first native 3D-model conditioned detector. Tab. 3e shows that increasing the 3D model "vocabulary" from 35k to 110k results in a substantial 5.1 AP improvement. This scale yields the improvements of **13.3 AP** (BOP-H3) and **23.6 AP** (BOP-Industrial) over MUSE in Tab. 1a.

Single-View RGB-Pose: In Tab. 3c, we analyze the source of our improvements upon prior methods on single-view pose estimation. Applying our refinement to Co-Op's [44] predictions yields a 1.4 AP gain, demonstrating the added

Table 3. **Ablations:** Top row (a–c): ablations show which parts of our approach, when added to prior SotA, explain our performance gains. Bottom row (d–e): hyperparameter ablations reveal that scaling 3D-model vocabulary and using multiple hypotheses drive performance.

(a) **Multi-view refinement accounts for most mm-level gains,** added on top of FRTPose. 6DoF Pose AP-mm reported.

Method / Ablation	ITODD[18]	XYZ[30]	IPD[32]	Average	Δ
FRTPose [59]	74.4	64.5	65.9	68.3	—
+ Our Refinement	**77.7**	67.6	89.5	78.3	+10.0
Ours Full	77.5	**72.0**	**92.3**	**80.6**	+12.3

(b) **Each detection component helps;** end-to-end training surpasses MUSE on HANDAL[21].

Method	AP	AP_{50}	AP_{75}	AR_1	ΔAP
MUSE [10]	35.7	45.1	38.4	45.0	—
+ class id	40.1	49.9	43.6	52.5	+4.4
+ scoring	42.3	52.9	46.2	52.3	+6.6
+ regression	43.7	53.1	50.0	55.2	+8.0
Ours Full	**53.6**	**66.2**	**61.2**	**64.9**	+17.9

(c) **Refinement adds single-view improvements on RGB 6D Pose;** AP reported.

Method	K	BOP-H3	Δ
Co-Op [44]	1	46.4	—
+ Our refinement	1	47.8	+1.4
Ours	1	55.1	+8.7
Ours - Full	5	**58.7**	+12.3

(d) **Refinement Ablations (AP-mm):** Our final solution uses $S = 3$, $K = 8$, and 3 Iters. It doesn't use known object classes O.

HPPs				BOP-Industrial				
O	K	Iters	S	ITODD[18]	XYZ[30]	IPD[32]	Avg.	Δ
✓	1	3	3	66.8	63.9	82.7	71.1	—
✓	4	3	3	76.3	70.9	90.9	79.4	+8.3
✓	8	1	3	75.4	67.5	88.7	77.2	+6.1
✓	8	2	3	77.4	72.6	92.2	80.7	+9.6
✓	8	3	1	75.2	70.7	92.0	79.3	+8.2
✓	8	3	3	**77.6**	**72.7**	**92.3**	**80.9**	+9.8
×	8	3	3	77.5	72.0	**92.3**	80.6	+9.5

(e) **Scaling 3D-model vocabulary and training iterations is key:** 110k CADs and 300k iters ($S = 3$) yield best mean AP across BOP-H3[1, 21, 57] and BOP-Industrial[18, 30, 32].

HPPs			Average Precision			
S	Train CADs	Train Iters	BOP-H3	BOP-Ind.	Mean	Δ
3	35k	100k	44.2	55.1	49.7	—
3	110k	100k	52.6	56.9	54.8	+5.1
1	110k	300k	52.0	**58.6**	55.3	+5.6
3	110k	300k	**54.8**	57.9	**56.4**	+6.7

value of our refinement stage, even in single-view. However, using 3PT-D initial poses with $K = 1$ yields an 8.7 AP gain, confirming that superior detection & pose initialization is a key driver for single-view performance. Finally, $K = 5$ orientation hypotheses give additional 3.6 AP improvement, showing that scoring with the refined 3-D object understanding 3PT-R is superior to 3PT-D's coarse scoring.

Precise Industrial Pose: *Multi-View RGB Refinement > High-Res RGB-D Refinement* In the main result, 3PT achieves an 80.6 AP-mm on Industrial datasets, a **56.5% relative improvement** over the RGB-D baseline FreeZeV2 [5] and a **29.1** AP-mm. Tab. 3a, shows that 10.0 of the 12.3 AP-mm improvement is achieved by applying our RGB-only multi-view refinement + scoring to FRTPose [59] predictions. Our full pipeline (3PT-D + 3PT-R) adds another 2.3 AP. Remarkably, our multi-view RGB refinement achieves mm-level pose accuracy that exceeds FRTPose, an ensemble of the best RGB-D refinement methods with high-end depth sensors on industrial data.

Dataset Size: We retrained with 3PT trained on a MegaPose [36] sized subset of our dataset. This only leads to a slight (−1.9 AP) drop across the BOP benchmarks and still maintains SoTA performance. Full details in supplement.

While most hyperparameter changes (Tab. 3d) yield modest shifts (∼ 2% AP-mm), generating multiple 3PT-D rotation hypotheses ($K = 8$ vs $K = 1$) and filtering them via 3PT-R's fine-grained object understanding contributes a substantial gain of 9.8 AP, similar to single-view RGB. Knowing object classes *a-priori* ($O = ✓$ vs $O = ×$) only leads to a 0.3 AP-mm gain.

Runtime Limitations: 3PT averages 30.5s on BOP-Industrial (H100), compared to 16.8s for FreeZeV2 [5] (L40) and 162s for FRTPose [59] (RTX 5090). The primary computational bottleneck for scenes containing numerous diverse objects is the per-object forward pass through the 3PT-D image tower, which we hope to address in future work. In our bin-picking application where the object class is known *a priori* ($O = ✓$), our method averages 11.5s for 50 object instances and 1.6s for 5. Further discussion on runtime and other limitations in supplement.

4.3. Applications

3PT enables multiple challenging real-world robotics tasks. Videos are provided in supplementary materials.

Bin-Picking: Precise 3D-poses enable bin picking with 3 cameras at 2m distances, with the robot fully clearing a bin with 100 densely cluttered parts (see Fig. 4b).

Precise-Insertion: We demonstrate 96/96 successful attempts of vision-only precision insertion of a DIMM into a PCB slot using 6D-pose (see Fig. 4c). This insertion has a 0.5mm and 2 degree pose tolerance.

Tracking: For AR glasses, where depth may not available [1], our approach performs multi-view tracking by repeatedly running 3PT-R Sec. 3.3 (see Fig. 4a).

5. Conclusion

We introduce the 3D-Object Perception Transformer (3PT), a unified, depth-free framework featuring a novel 3D-conditioned detection stage and a multi-view RGB refinement loop for robust 3D-object perception. This approach achieves state-of-the-art performance on BOP benchmarks, notably securing gains of +23.6 detection AP and +29.1 pose AP-mm on BOP-Industrial. Finally, 3PT achieves industrial-grade accuracy and reliability, demonstrated in real-world tasks like bin-picking and DIMM-insertion.

References

[1] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. *CVPR*, 2025. 6, 7, 8

[2] P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. 3

[3] Matteo Bortolon, Theodore Tsesmelis, Stuart James, Fabio Poiesi, and Alessio Del Bue. 6dgs: 6d pose estimation from a single image and a 3d gaussian splatting model. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[4] Eric Brachmann. 6D Object Pose Estimation using 3D Object Coordinates [Data], 2020. 7

[5] Andrea Caraffa, Davide Boscaini, Amir Hamza, and Fabio Poiesi. Freeze: Training-free zero-shot 6d pose estimation with geometric and vision foundation models. In *ECCV*, pages 414–431. Springer, 2024. 2, 3, 7, 8

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 4

[7] Long Chen, Han Yang, Chenrui Wu, and Shiqing Wu. Mp6d: An rgb-d dataset for metal parts' 6d pose estimation. *IEEE Robotics and Automation Letters*, 7(3):5912–5919, 2022. 6

[8] Xiaotong Chen, Huijie Zhang, Zeren Yu, Anthony Opipari, and Odest Chadwicke Jenkins. Clearpose: Large-scale transparent object dataset and benchmark. In *European conference on computer vision*, pages 381–396. Springer, 2022. 6

[9] Sungmin Cho, Sungbum Park, and Insoo Oh. F3dt2d: Method submission to the bop challenge 2024. https://bop.felk.cvut.cz/method_info/761/, 2024. BOP: Benchmark for 6D Object Pose Estimation. 3

[10] Sungmin Cho, Sungbum Park, and Insoo Oh. Muse: Model-based uncertainty-aware similarity estimation for zero-shot 2d object detection and segmentation. *arXiv preprint arXiv:2510.17866*, 2025. 2, 3, 7, 8

[11] Alvaro Collet and Siddhartha S Srinivasa. Efficient multi-view object recognition and full pose estimation. In *2010 IEEE International Conference on Robotics and Automation*, pages 2050–2055. IEEE, 2010. 5

[12] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F. Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21126–21136, 2022. 6

[13] Peter De Roovere, Rembert Daems, Jonathan Croenen, Taoufik Bourgana, Joris de Hoog, and Francis Wyffels. Cendernet: center and curvature representations for render-and-compare 6d pose estimation. In *European Conference on Computer Vision*, pages 97–111. Springer, 2022. 7

[14] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13142–13153, 2023. 6

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 4

[16] Andreas Doumanoglou, Rigas Kouskouridas, Sotiris Malassiotis, and Tae-Kyun Kim. Recovering 6d object pose and predicting next-best-view in the crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[17] 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, 2022. 6

[18] Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Hartinger, and Carsten Steger. Introducing mvtec itodd - a dataset for 3d object recognition in industry. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017. 2, 6, 7, 8

[19] Ronald Aylmer Fisher. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 217(1130):295–305, 1953. 5

[20] Bernd Von Gimborn, Philipp Ausserlechner, Markus Vincze, and Stefan Thalhammer. Diffusion features for zero-shot 6dof object pose estimation, 2024. 3

[21] Andrew Guo, Bowen Wen, Jianhe Yuan, Jonathan Tremblay, Stephen Tyree, Jeffrey Smith, and Stan Birchfield. HANDAL: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions. In *IROS*, 2023. 6, 7, 8

[22] Rasmus Laurvig Haugaard and Thorbjorn Mosekjaer Iversen. Multi-view object pose estimation from correspondence distributions and epipolar geometry. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1786–1792, 2023. 7

[23] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without CAD models. In *Advances in Neural Information Processing Systems*, 2022. 3

[24] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Computer Vision – ACCV 2012*, pages 548–562, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. 6

[25] Tomas Hodan, Pavel Haluza, Stepan Obdrzalek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-LESS: An

RGB-D Dataset for 6D Pose Estimation of Texture-Less Objects . In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 880–888, Los Alamitos, CA, USA, 2017. IEEE Computer Society. 2, 7

[26] Tomáš Hodaň, Frank Michel, Eric Brachmann, Wadim Kehl, Anders Glent Buch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiří Matas, and Carsten Rother. BOP: Benchmark for 6D object pose estimation. *European Conference on Computer Vision (ECCV)*, 2018. 6

[27] Tomáš Hodaň et al. BOP: Benchmark for 6d object pose estimation leaderboards. https://bop.felk.cvut.cz/leaderboards/, 2025. Accessed: 2025-11-13. 7

[28] Tomáš Hodaň, Frank Michel, Eric Brachmann, Wadim Kehl, Anders Glent Buch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiří Matas, and Carsten Rother. Bop: Benchmark for 6d object pose estimation. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, page 19–35, Berlin, Heidelberg, 2018. Springer-Verlag. 7

[29] Junwen Huang, Hao Yu, Kuan-Ting Yu, Nassir Navab, Slobodan Ilic, and Benjamin Busam. Matchu: Matching unseen objects for 6d pose estimation from rgb-d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10095–10105, 2024. 3, 7

[30] Junwen Huang, Jizhong Liang, Jiaqi Hu, Martin Sundermeyer, Peter KT Yu, Nassir Navab, and Benjamin Busam. Xyz-ibd: High-precision bin-picking dataset for object 6d pose estimation capturing real-world industrial complexity, 2025. 2, 6, 7, 8

[31] Junwen Huang, Shishir Reddy Vutukur, Peter KT Yu, Nassir Navab, Slobodan Ilic, and Benjamin Busam. Raypose: Ray bundling diffusion for template views in unseen 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9102–9112, 2025. 3, 6

[32] Agastya Kalra, Guy Stoppi, Dmitrii Marin, Vage Taamazyan, Aarrushi Shandilya, Rishav Agarwal, Anton Boykov, Tze Hao Chong, and Michael Stark. Towards co-evaluation of cameras hdr and algorithms for industrial-grade 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22691–22701, 2024. 6, 7, 8

[33] Roman Kaskman, Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. HomebrewedDB: RGB-D Dataset for 6D Pose Estimation of 3D Objects . In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2767–2776, Los Alamitos, CA, USA, 2019. IEEE Computer Society. 7

[34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, and et al. Segment anything. In *ICCV*, 2023. 2

[35] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *European Conference on Computer Vision*, pages 574–591. Springer, 2020. 3

[36] 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. In *Conference on Robot Learning (CoRL)*, 2022. 3, 8

[37] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27906–27916, 2024. 2, 3, 7

[38] Lihua Liu, Jiehong Lin, Zhenxin Liu, and Kui Jia. Picopose: Progressive pixel-to-pixel correspondence learning for novel object pose estimation. *arXiv preprint arXiv:2504.02617*, 2025. 3, 6

[39] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 3

[40] Yangxiao Lu, Jishnu Jaykumar P, Yunhui Guo, Nicholas Ruozzi, and Yu Xiang. Adapting pre-trained vision models for novel instance detection and segmentation. *arXiv preprint arXiv:2405.17859*, 2024. 3

[41] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *ECCV*, pages 728–755. Springer, 2022. 4

[42] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. 36:72983–73007, 2023. 4

[43] 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. 6

[44] Sungphill Moon, Hyeontae Son, Dongcheol Hur, and Sangwook Kim. Co-op: Correspondence-based novel object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3, 6, 7, 8

[45] 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. 2, 3, 7

[46] 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*, 2024. 3, 7

[47] Van Nguyen Nguyen, Stephen Tyree, Andrew Guo, Mederic Fourmy, Anas Gouda, Taeyeop Lee, Sungphill Moon, Hyeontae Son, Lukas Ranftl, Jonathan Tremblay, Eric Brachmann, Bertram Drost, Vincent Lepetit, Carsten Rother, Stan Birchfield, Jiri Matas, Yann Labbe, Martin Sundermeyer, and Tomas Hodan. BOP challenge 2024 on model-

based and model-free 6D object pose estimation. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW, CV4MR Workshop)*, 2025. 3

[48] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 3, 6

[49] Lukas Ranftl, Felix Brendel, Bertram Drost, and Carsten Steger. Mvtop: Multi-view transformer-based object pose-estimation. *arXiv preprint arXiv:2508.03243*, 2025. 3

[50] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021. 6

[51] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 658–666, 2019. 4

[52] Tim Salzmann, Markus Ryll, Alex Bewley, and Matthias Minderer. Scene-graph vit: End-to-end open-vocabulary visual relationship detection. In *ECCV*, pages 195–213. Springer, 2024. 4

[53] Ivan Shugurov, Sergey Zakharov, and Slobodan Ilic. Dpodv2: Dense correspondence-based 6 dof pose estimation. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7417–7435, 2021. 7

[54] Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic. Osop: A multi-stage one shot object pose estimation framework, 2022. 3

[55] Ritvik Singh, Jingzhou Liu, Karl Van Wyk, Yu-Wei Chao, Jean-Francois Lafleche, Florian Shkurti, Nathan Ratliff, and Ankur Handa. Synthetica: Large scale synthetic data for robot perception, 2024. 6

[56] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6825–6834, 2022. 3

[57] Stephen Tyree, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Jeffrey Smith, and Stan Birchfield. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In *International Conference on Intelligent Robots and Systems (IROS)*, 2022. 6, 7, 8

[58] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[59] Yulin Wang, Mengting Hu, Jianghao Zhuo, and Chen Luo. Frt-pose waprv2 (multi-cam). https://bop.felk.cvut.cz/method_info/1157/, 2025. Accessed: 2025-10-27. 2, 3, 7, 8

[60] Bowen Wen, Chaitanya Mitash, and Kostas Bekris. Data-driven 6d pose tracking by calibrating image residuals in synthetic domains. *arXiv preprint arXiv:2105.14391*, 2021. 6

[61] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *CVPR*, pages 17868–17879, 2024. 3, 7

[62] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 803–814, 2023. 6

[63] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. 2018. 6, 7

[64] Jun Yang, Yizhou Gao, Dong Li, and Steven L Waslander. Robi: A multi-view dataset for reflective objects in robotic bin-picking. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9788–9795. IEEE, 2021. 2

[65] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023. 3

[66] Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomas Hodan. Foundpose: Unseen object pose estimation with foundation features, 2024.