2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

GFreeDet2: Exploiting Gaussian Splatting and Foundation Models for RGB-based Model-free 2D and 6D Detection of Unseen Objects

Gu Wang*,†, Xingyu Liu*, Jingyi Tang, Chengxi Li, Yingyue Li, Ziqin Huang, and Xiangyang Ji†

***Abstract*— We introduce GFreeDet2, which leverages Gaussian Splatting and foundation models to address RGB-based model-free 2D detection and 6D detection of unseen objects. GFreeDet2 reconstructs 3D Gaussian object models from multi-view RGB references, enabling efficient model-free detection without relying on CAD models. To accelerate reconstruction and consistently handle both pinhole and fisheye cameras, we propose projection-aware perspective cropping (PAPC) with visual hull initialization. PAPC further improves coarse 6D detection by accurately extracting pinhole crops from fisheye query images. The Gaussian objects enable rendering in place of CAD models within foundation model-driven pipelines, allowing existing state-of-the-art RGB-based methods for unseen 2D and 6D detection to be extended to the model-free setting with minimal modifications. Extensive experiments on all three BOP-H3 datasets demonstrate that GFreeDet2 achieves state-of-the-art performance and establishes a strong baseline for RGB-based, model-free 2D and 6D unseen object detection. The code is publicly available at github.com/wangg12/GFreeDet2.git.**

I. INTRODUCTION

Object pose estimation, which seeks to recover the position and orientation of an object relative to the camera from images, has long been regarded as a fundamental problem in robotics [1] and 3D computer vision [2]. Despite recent remarkable advances, classic instance-level and category-level formulations remain restricted to fixed sets of objects [3], [4], [5], [6], [7] or categories [8], [9], [10], [11], and require retraining when faced with previously unseen ones. These limitations hinder scalability and flexibility, thereby restricting their applicability in open-world environments. As a result, unseen object pose estimation [12], [13], [14], [15], [16], [17], [18] has recently emerged as a growing research focus. Nevertheless, most existing approaches still rely on CAD models as references, whose acquisition is costly. Moreover, most 6D pose estimation methods [16] rely on 2D detection to provide object candidates and coarse localization. This dependency is particularly strong for unseen objects, yet prior works study them separately, leaving unified solutions largely unexplored.

Recent efforts tackle unseen object pose estimation from a single RGB reference [19], a single RGBD reference [20], [21], [22], or multi-view RGB references [23], [24], as illustrated in Fig. 1. However, these methods remain limited: they either estimate only rotation, require depth for full 6D pose recovery, or still depend on detectors trained on known objects. In contrast, model-free 2D detection and 6D pose estimation of unseen objects from RGB images remain largely unexplored. While existing pipelines predominantly focus on pinhole cameras, fisheye imagery is increasingly critical in robotics and mixed reality (MR) applications [25], [26]. To reflect this trend, the recently introduced BOP-H3 benchmark [16] contains datasets captured by both pinhole (HANDAL [27], HOPEv2 [28]) and fisheye (HOT3D [26]) cameras, motivating the need for a unified treatment of both image types. Notably, on the BOP-H3 benchmark for model-free unseen tasks [16], only GFreeDet [29] has reported 2D detection results across three datasets, while no method has yet achieved complete results for 6D detection. The joint requirements of being model-free, unseen, RGB-based, and applicable to both pinhole and fisheye cameras render this problem especially challenging.

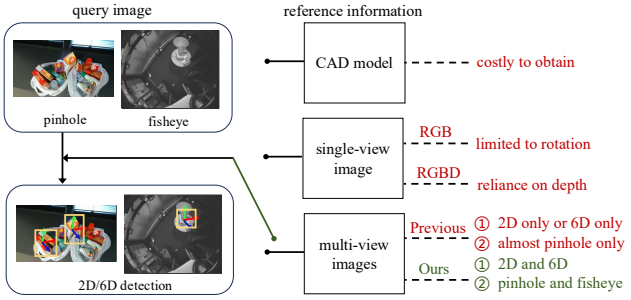

Fig. 1: **Comparison of paradigms for unseen object pose estimation.** Existing methods are limited by cost (CAD), sensor requirements (RGBD), or specialization to a limited task or camera type. In contrast, GFreeDet2 delivers full 2D and 6D unseen detection from RGB in a model-free manner, working seamlessly across pinhole and fisheye imagery.

To overcome these challenges, we propose GFreeDet2, which leverages Gaussian splatting and foundation models to address RGB-based, model-free 2D detection and 6D detection of unseen objects. Benefiting from recent advances in 3D Gaussian Splatting (3DGS) [30], we aim to reconstruct Gaussian objects (GS objects) from multi-view reference RGB images to support model-free 2D and 6D detection of unseen objects. Direct reconstruction from high-resolution reference images, however, is computationally prohibitive. To mitigate this, we crop and resize the images to a fixed,

[1]Gu Wang is with the Lab for High Technology, Tsinghua University, Beijing, 100084, China. E-mail: wangg12@tsinghua.org.cn.
[2]Xingyu Liu, Jingyi Tang, Chengxi Li, Yingyue Li, Ziqin Huang, and Xiangyang Ji are with the Department of Automation, Tsinghua University, Beijing, 100084, China, and also with BNRist, Beijing, 100084, China. E-mail: {liuxy21, lichengx21, yingyue-21, huang-zq24}@mails.tsinghua.edu.cn, {tangjy,xyji}@tsinghua.edu.cn.
*: Gu Wang and Xingyu Liu have equally contributed.
†: Corresponding authors.

smaller resolution and further employ the visual hull [31] technique for fast geometric initialization, which allows rapid 3DGS training. To further unify pinhole and fisheye inputs, we introduce projection-aware perspective cropping (PAPC), which eliminates the border effect of naive perspective cropping [14] and ensures geometrically consistent pinhole crops for efficient 3DGS reconstruction.

Once the Gaussian models are reconstructed, we replace CAD-based renderings with GS-based renderings in existing pipelines for 2D and 6D unseen object detection. This approach extends detection pipelines to a model-free setting with minimal modifications, fully capitalizing on recent advancements in the field. Notably, our PAPC ensures that fisheye images are correctly cropped for coarse 6D detection, improving accuracy and maintaining compatibility with existing pose estimation methods [14]. This design enables GFreeDet2 to build upon top-performing foundation model–driven methods for RGB-based unseen 2D and 6D detection. Meanwhile, it provides a crucial alternative that effectively bridges the gap between CAD-model-based and fully model-free approaches. In a nutshell, our main contributions are threefold:

- We propose projection-aware perspective cropping (PAPC), which enables efficient and unified GS object reconstruction from multi-view RGB references, handles both pinhole and fisheye cameras, and improves coarse 6D detection.
- We showcase that existing CAD-based, foundation-model-driven methods for unseen 2D and 6D detection can be plugged into the model-free setting with minimal modifications.
- To the best of our knowledge, GFreeDet2 is the first to achieve complete model-free RGB-based 2D and 6D unseen detection on the BOP-H3 benchmark, establishing a strong baseline with state-of-the-art performance.

II. RELATED WORK

A. Gaussian Splatting for 3D Object Reconstruction

3D object models form the foundation for many 3D tasks. Most methods for 2D detection and 6D pose estimation of unseen objects assume access to a mesh model for rendering [13], [14], [12], which is usually acquired through CAD, scanning, or classical 3D reconstruction processes that are costly and labor-intensive.

Neural rendering has recently enabled high-quality 3D reconstruction at much lower cost [32]. Among these advances, 3D Gaussian Splatting (3DGS) [30], which integrates explicit point-based representations with optimization techniques from neural rendering, has proven particularly effective. Subsequent variants have further enhanced the practicality of 3DGS. For instance, Mip-Splatting [33] enables alias-free rendering across scales, Fisheye-GS [34] extends it to distortion-free fisheye imagery, and GaussianObject [35] achieves object-level reconstruction from sparse views using visual hull initialization and generative priors. Moreover, gsplat [36] offers an optimized, user-friendly framework that incorporates various enhancements over vanilla 3DGS.

Despite these advances, existing methods either primarily focus on scene-level or are limited to specific camera models. To address this, we propose an efficient Gaussian object reconstruction method that unifies pinhole and fisheye camera models to enable RGB-based, model-free 2D and 6D detection of unseen objects.

B. Vision Foundation Models

Vision foundation models, particularly the DINO [37], [38] and SAM [39], [40], [41] families, have recently reshaped core vision tasks such as detection, segmentation, and pose estimation. DINOv2 [37], pre-trained on massive data, yields highly transferable visual representations that can be applied to downstream tasks in a training-free manner or with minimal fine-tuning. Its successor, DINOv3 [38], further scales the model to 7B parameters with improved self-distillation. Segment Anything (SAM) [39] revolutionizes the segmentation paradigm by enabling zero-shot segmentation via flexible prompts (e.g., points, bounding boxes, or masks). Moreover, FastSAM [41] integrates YOLOv8 [42] for improved efficiency, while SAM2 [40] extends the framework to temporally consistent video segmentation.

C. Unseen Object Detection and Pose Estimation

With advances in vision foundation models, research on object detection and pose estimation has shifted toward open-world scenarios, emphasizing generalization to unseen objects. For *2D unseen detection*, CNOS [43] sets a strong baseline by matching SAM/FastSAM [39], [41] mask proposals with CAD-rendered templates using DINOv2 [37]. SAM-6D [17] proposes an improved matching scoring scheme that significantly surpasses CNOS when depth is available. It also incorporates a transformer to estimate the 6D pose for unseen objects. GFreeDet [29] instead uses reconstructed Gaussian objects for model-free unseen 2D detection, but requires separate reconstruction and rendering strategies for pinhole and fisheye cameras. For *unseen pose estimation*, CAD-based approaches [12], [13] typically follow a render-and-compare paradigm [3] trained on large-scale synthetic data, including variants that predict optical flow for correspondence-based pose recovery [44], [45], [46]. These methods are often used for pose refinement, whereas coarse pose estimation leverages foundation models either in a training-free manner [14], [15] or with minimal finetuning [47]. While most methods [16], [18] still rely on CAD models, model-free alternatives are emerging, such as reconstructing sparse 3D points from multi-view RGB references [23], [24] or using a single RGBD reference image [20], [22], [21].

Nevertheless, RGB-based model-free approaches that jointly tackle 2D and 6D unseen detection, and work across both pinhole and fisheye cameras, remain underexplored.

III. METHOD

A. Overview

Problem Statement. Given a set of previously unseen objects $\mathcal{O}$, the goal of model-free 2D and 6D detection is to estimate the 2D bounding boxes, instance masks, and

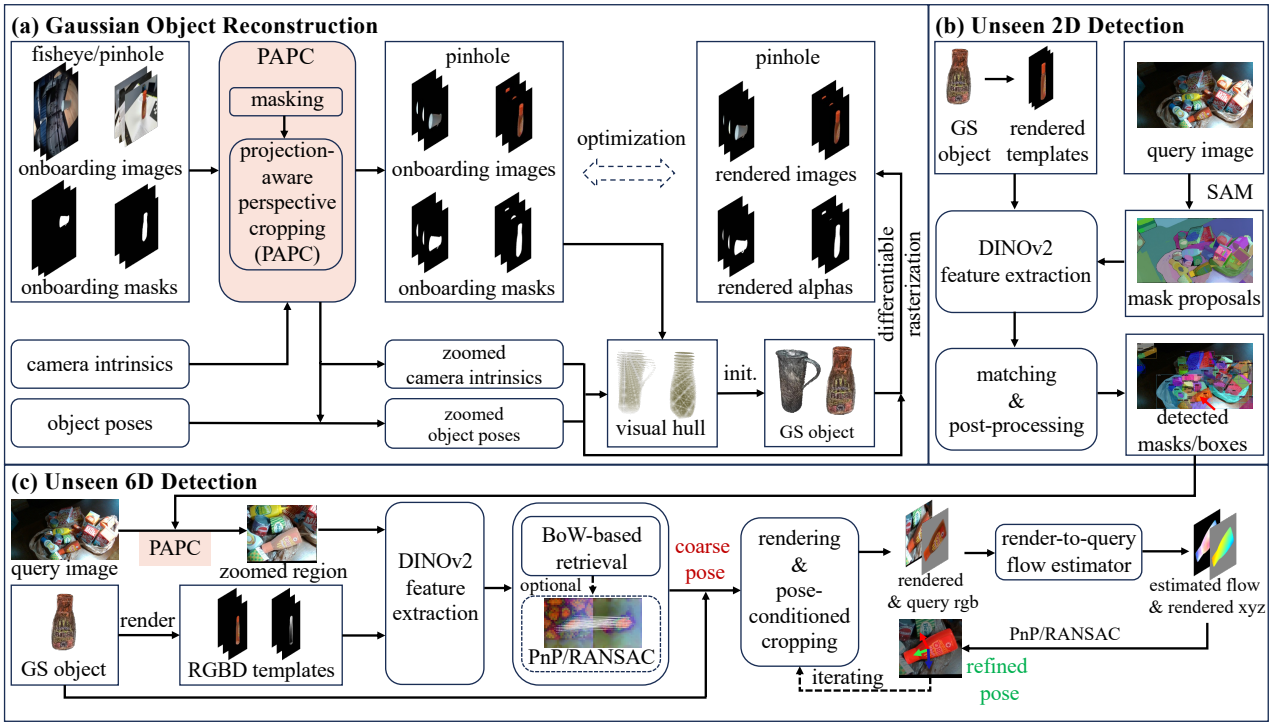

Fig. 2: **The framework of GFreeDet2.** For an unseen object, we reconstruct its Gaussian representation (GS object) from zoomed onboarding data obtained via projection-aware perspective cropping (PAPC). GS-rendered templates enable RGB-based, model-free unseen 2D and 6D detection with vision foundation models. For coarse pose estimation, PAPC is applied to crop query images, followed by Bow-based retrieval (with optional PnP/RANSAC). The pose is then iteratively refined through a render-to-query flow-based refiner using the GS object for rendering.

6D object poses of all possible objects from a query RGB image $\mathbf{I}_q$ without knowing the CAD models of the objects. To compensate for the absence of CAD models, each object is provided with corresponding reference images $\mathcal{I}_r$ that can be used for onboarding. The object poses and masks are assumed to be available for each reference image, as they can be obtained by using toolkits such as COLMAP [48] and SAM2 [40]. Additionally, we assume the camera intrinsics are available for both query and reference images, as they are typically provided in commonly used datasets and can be readily obtained through calibration. Note that, unlike localization, where the possible object categories and instance counts are known in advance for each query image, detection must be performed without such prior knowledge, making it a more challenging task [16].

Framework of GFreeDet2. Fig. 2 illustrates the framework of GFreeDet2, which leverages Gaussian Splatting [30] and vision foundation models [37], [39] to achieve RGB-based, model-free unseen 2D detection and 6D detection. For an unseen object without the CAD model, we first reconstruct its 3D Gaussian representation (GS object) from zoomed onboarding data, processed via projection-aware perspective cropping (PAPC). The GS object is then used to render multiview templates that enable 2D unseen detection and coarse 6D unseen detection, assisted by vision foundation models like SAM [39] and DINOv2 [37]. During coarse pose estimation, PAPC is also employed to crop query images, thereby supporting consistent handling of both pinhole and fisheye images. The coarse pose is estimated through BoW-based retrieval, optionally followed by PnP/RANSAC [14]. Finally, the pose is further iteratively refined through a render-to-query flow-based refiner [46], with the GS object serving as the rendering source.

B. Gaussian Object Reconstruction

To better leverage prior advances in CAD-based unseen 2D and 6D detection, we reconstruct the Gaussian representation (GS object O_{GS}) of an unseen object from its onboarding data, which consists of RGB images, estimated masks, camera intrinsics, and object poses. However, training O_{GS} directly on full-resolution images is computationally costly. We instead train O_{GS} using cropped and resized onboarding data. Unlike GFreeDet [29], which relies on separate pipelines for fisheye and pinhole images, we propose a unified projection-aware perspective cropping (PAPC) strategy to crop and resize the onboarding data.

Projection-Aware Perspective Cropping (PAPC). PAPC converts both pinhole and fisheye images into a standardized, cropped pinhole view. It first unprojects the 2D bounding box corners into 3D rays, then estimates the centroid and radius of the enclosing sphere, and constructs a virtual pinhole (perspective) camera pointing to the centroid, with its focal length adjusted to ensure that the object fits the desired viewport.

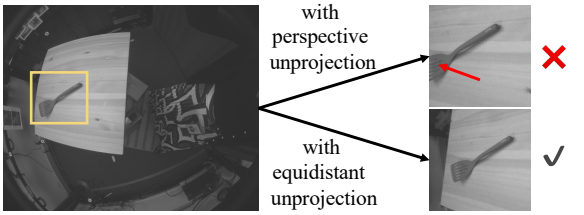

Fig. 3: In our projection-aware perspective cropping (PAPC), equidistant unprojection enables correct pinhole crops from fisheye images.

For a 2D pixel coordinate (u, v) and camera intrinsics with focal length (f_x, f_y) and principal point (c_x, c_y), the corresponding unprojected ray for a pinhole input image is

$$\mathbf{d}_p = \phi([u - c_x, v - c_y, f]^\top), \tag{1}$$

where $f = 0.5(f_x + f_y)$ is the approximate focal length for simplicity, and $\phi(\cdot)$ denotes vector normalization.

For a fisheye image, PAPC applies the equidistant model:

$$\mathbf{d}_f = \phi([(u - c_x)\sin\theta, (v - c_y)\sin\theta, \cos\theta]^\top), \tag{2}$$

with $r = \sqrt{(u - c_x)^2 + (v - c_y)^2}$ and $\theta = \frac{r}{f}$ denoting the radial distance to the principal point and the incidence angle.

Using the unprojected rays $\{\mathbf{d}_i\}_{i=1}^4$ from the 2D bounding box corners, PAPC computes the centroid $\mathbf{c} = \frac{1}{4}\sum_{i=1}^4 \mathbf{d}_i$ and radius $\rho = \max_i \|\mathbf{d}_i - \mathbf{c}\|$ of the enclosing sphere. Transforming $\mathbf{c}$ into the world frame gives $\tilde{\mathbf{c}}_w = (\mathbf{T}_w^c)^{-1}\tilde{\mathbf{c}}$, where $\tilde{\cdot}$ denotes homogeneous coordinates. A virtual pinhole camera is then constructed such that its optical axis passes through $\mathbf{c}_w$, using a look-at transformation $\mathbf{T}_w^{vc} = \texttt{LootAt}(\mathbf{T}_w^c, \mathbf{c}_w)$. The centroid in this virtual camera is $\tilde{\mathbf{c}}_{vc} = \mathbf{T}_w^{vc}\tilde{\mathbf{c}}_w$. The projected 2D radius, enlarged by a relative padding factor s_{pad}, is $(\rho_x^{\text{2D}}, \rho_y^{\text{2D}}) = \frac{\rho(1+s_{pad})}{z_{vc}}(f_x, f_y)$, where z_{vc} is the third component of $\mathbf{c}_{vc}$.

Given the desired crop size (s_x, s_y), we set the principal point and focal lengths of the virtual perspective camera as

$$(c_x^{(vc)}, c_y^{(vc)}) = \frac{1}{2}(s_x, s_y), \tag{3}$$

$$(f_x^{(vc)}, f_y^{(vc)}) = \left(\frac{f_x s_x}{2\rho_x^{\text{2D}}}, \frac{f_y s_y}{2\rho_y^{\text{2D}}}\right). \tag{4}$$

With these parameters, PAPC tightly fits the padded object region within the virtual viewport. Finally, a warp function φ reprojects pixels via the original and virtual camera parameters, transforming the object region from the original pinhole or fisheye image to the zoomed pinhole view [26].

PAPC extends the perspective cropping in [14] to fisheye images by applying equidistant unprojection, enabling consistent cropped pinhole views from both pinhole and fisheye inputs. As shown in Fig. 3, this ensures that objects from fisheye images are properly centered and scaled.

GS object training. As depicted in Fig. 2 (a), given the zoomed pinhole onboarding data, we initialize the GS object using a 3D visual hull motivated by [35]. The visual hull is constructed by keeping only the valid 3D points whose 2D projections lie within the object masks. For efficiency, up to 8 posed masks are selected via farthest point sampling (FPS) for hull generation. The GS object is then trained for 10K iterations using a composite loss that combines L1, SSIM [49], and mask terms.

C. 2D Detection of Unseen Objects

We adopt a template-based, foundation-model-driven approach for unseen 2D detection, extending [29] by adding real templates and morphological post-processing.

Using FPS, we sample 64 templates from onboarding data to complement the GS-rendered ones. Then, for each SAM-generated query mask proposal [39], we assign a class and a confidence score by matching against templates using DINOv2 [37] global and local features. The global feature $\mathbf{g}_p$ of a proposal retrieves the top 5 global template features $\{\mathbf{g}_{o,i}\}_{i=1}^5$ per object o, and the global score is computed as the average cosine similarity

$$\xi_{g,o} = \frac{1}{5}\sum_{i=1}^{5} \text{sim}_{\cos}(\mathbf{g}_p, \mathbf{g}_{o,i}). \tag{5}$$

The object class o is selected by the highest global score among all candidate objects. Following [29], we define a local appearance score to adjust the matching quality as

$$\xi_{l,o} = \frac{1}{N_l}\sum_{k=1}^{N_l} \max_{i=1,\ldots,N_l} \text{sim}_{\cos}(\mathbf{l}_p^k, \mathbf{l}_o^i), \tag{6}$$

where $\{\mathbf{l}_{p|o}^i\}_{i=1}^{N_l}$ denotes the local features. The overall matching score is

$$\xi_o = \frac{1}{2}(\xi_{g,o} + \xi_{l,o}). \tag{7}$$

Afterwards, we discard low-score predictions and remove duplicates via non-maximum suppression (NMS). We further apply morphological closing and opening as a post-processing step to fill small holes and remove small isolated regions. The final 2D detections are derived from the tight bounding boxes of the masks (Fig. 2 (b)).

D. 6D Detection of Unseen Objects

As illustrated in Fig. 2 (c), our unseen 6D detection pipeline comprises coarse pose estimation followed by pose refinement. Leveraging O_{GS}, it extends model-based unseen pose estimation approaches to the model-free scenario with minimal modifications while retaining their core strengths.

The coarse stage builds on FoundPose [14], a model-based approach that operates on RGB images and combines bag-of-words (BoW) template retrieval with an optional PnP/RANSAC step. The BoW descriptors are computed from DINOv2 local features through PCA and clustering. Distinct from [14], we employ PAPC to generate zoomed query regions, handling fisheye query images accurately. For efficiency, feature-metric refinement is omitted, and PnP/RANSAC can also be skipped following [46]. The retrieved pose in the virtual camera frame is converted to the original camera frame through inverse warping φ^{-1}. While

this slightly reduces accuracy, it substantially accelerates the pipeline, as subsequent refinement primarily relies on the 2D projection initialization already sufficiently provided by retrieval.

For pose refinement, we adapt the RGB-based iterative refiner GoTrack [46] to the model-free setting by replacing CAD-based rendering with Gaussian splatting, forming GoTrack-GS. GoTrack-GS leverages the frozen GoTrack network to predict render-to-query flow and confidences using zoomed rendered and query images cropped according to the coarse pose. Pose-conditioned cropping uses the 3D GS object points to construct a virtual pinhole camera, making PAPC's unprojection unnecessary. The refined pose is then obtained via PnP/RANSAC from flow-established 2D-3D correspondences, completing a unified, efficient pipeline for unseen 6D object detection.

IV. EXPERIMENTS

A. Experimental Setup

Implementation Details. Our method is implemented in PyTorch [50] and evaluated on an NVIDIA L20 GPU. GS object reconstruction uses 280×280 crops (slightly larger than [29]) with a padding factor of 0.1. GS objects are trained on all static onboarding images with a maximum spherical harmonic (SH) degree of 2. For coarse pose estimation, we follow [14] with a default crop scale of 420 and padding of 0.2, and a lightweight variant with 280 and 0.05. Unlike [14], [46], we adopt EPnP [51] instead of iterative PnP for RANSAC-based pose recovery to improve robustness.

Datasets. We evaluate GFreeDet2 on the BOP-H3 benchmark [16], which consists of three datasets. HOT3D [26] contains 33 objects across 1,028 egocentric scenes with 5,140 test frames, captured by two fisheye cameras (Quest3 and Aria). HOPEv2 [28] has 28 toy grocery objects in 47 household and office scenes with 457 test images, and HANDAL [27] includes 40 manipulable objects from 7 categories with 1,684 test images. Both HOPEv2 and HANDAL use pinhole cameras. For the model-free setting, we use the static onboarding images for each object provided by BOP Challenge [16], and for HOT3D we specifically use the Aria static onboarding sequences.

Evaluation Metrics. We report average precision (AP) for 2D and 6D detection using the online server from [16]. For 2D detection, AP_{2D} follows the COCO protocol [52], averaging per-object AP over IoU thresholds $\{0.50, 0.55, \ldots, 0.95\}$. The dataset score is the mean over objects, and we also report the mean across datasets. For 6D detection, IoU is replaced by symmetry-aware pose errors: **MSSD** (maximum symmetry-aware surface distance) and **MSPD** (maximum symmetry-aware projection distance). For each metric in $\{\text{MSSD}, \text{MSPD}\}$, AP is averaged over standard correctness thresholds and then over objects. We report AP_{MSSD} and AP_{MSPD}, with the overall 6D detection score as their mean: $AP_{6D} = \frac{1}{2}\left(AP_{MSSD} + AP_{MSPD}\right)$. We also provide the average time per image in seconds, as in [16].

TABLE I: **Results of unseen 2D detection.** Best results for each onboarding type are marked as 1st, 2nd, and 3rd.

Method	Onboard.	HOT3D	HOPEv2	HANDAL	AP_{2D}	Time
MUSE (SAM2)	CAD	43.8	46.0	35.7	41.8	0.9
CNOS-FastSAM [43]	CAD	35.0	31.3	24.6	30.3	0.3
CNOS-SAM [43]	CAD	31.7	36.5	19.7	29.3	1.8
GFreeDet-FastSAM [29]	Static	33.8	36.4	25.5	31.9	0.3
GFreeDet-SAM [29]	Static	30.9	38.4	26.4	31.9	2.1
CNOS-FastSAM [43]	Static	37.3	34.3	30.4	34.0	0.4
GFreeDet2-2D-FastSAM	Static	40.3	36.9	34.4	37.2	0.2
GFreeDet2-2D-SAM2.1p	Static	42.1	45.2	37.3	41.5	1.8
GFreeDet2-2D-SAMp	Static	42.7	44.0	39.7	42.1	2.3

B. Results

We evaluate GFreeDet2 on BOP-H3 for model-free unseen object detection in both 2D and 6D, with comparisons from the official BOP leaderboard [1] and [16]. Our method is denoted as GFreeDet2-2D for 2D detection and GFreeDet2-6D for 6D detection.

Results of Unseen 2D Detection. Tab. I reports the results of unseen 2D detection. We evaluate three GFreeDet2-2D variants, each adopting a different mask proposal model: GFreeDet2-2D-SAMp (using SAM [39]), GFreeDet2-2D-SAM2.1p (using SAM2.1 [40]), and GFreeDet2-2D-FastSAM (using FastSAM [41]), where 'p' indicates morphological post-processing. For comparison, we also include the available CAD-based methods at the BOP leaderboard, given the limited number of existing model-free approaches.

Our GFreeDet2-2D-SAMp achieves the highest overall AP_{2D} of 42.1%, surpassing the best CAD-based method, MUSE (SAM2). GFreeDet2-2D-SAM2.1p attains a performance comparable to MUSE (SAM2). In the model-free setting, all of our variants outperform prior methods [43], [29], with GFreeDet2-2D improving by about 10% over GFreeDet [29], highlighting the benefit of our unified GS object reconstruction with PAPC over GFreeDet's separate pipelines. Additionally, GFreeDet2-2D-FastSAM outperforms CNOS-FastSAM (Static) [43] by 3%. Across different mask proposal models, SAM delivers the highest accuracy, followed by SAM2.1 and then FastSAM. In terms of efficiency, SAM2.1 is slightly faster than SAM, while FastSAM runs nearly 10× faster with only a 4-5% drop in AP.

Results of Unseen 6D Detection. Tab. II presents unseen 6D detection results, including available CAD-based methods and several GFreeDet2-6D variants that differ mainly in the underlying 2D detection method. Notably, GFreeDet2-6D is the only method with complete results on BOP-H3. Except for OPFormer (Static) at C4, all methods have both coarse and refined results. GFreeDet2-6D-lite is a lightweight variant that uses DINOv2S instead of DINOv2L for feature extraction with smaller crops. For reference, we also include GFDet-6D, a CAD-based counterpart of our method that relies on CNOS-FastSAM (CAD) for 2D detection and CAD rendering for pose estimation.

As shown in Tab. II, stronger 2D detectors generally yield better pose estimates, though the gap in final AP_{6D}

[1] bop.felk.cvut.cz/leaderboards

TABLE II: **Results of unseen 6D detection.** Best coarse and refined results are highlighted as 1st, 2nd, and 3rd.

Row	Method	Refinement	Onboard. type	2D Method	AP_{2D}	HOT3D	HOPEv2	HANDAL	AP_{6D}	Time
Coarse Pose Estimation										
C1	GigaPose [47]	-	CAD	-	-	7.2	16.7	4.1	9.4	0.9
C2	OPFormer	-	CAD	CNOS [43]	-	-	35.1	19.2	-	-
C3	GFDet-6D-c (Ours)	-	CAD	CNOS-FastSAM [43]	30.3	26.0	25.4	18.7	23.4	15.5
C4	OPFormer	-	Static	CNOS [43]	-	-	33.5	20.4	-	-
C5	GFreeDet2-6D-lite-c (Ours)	-	Static	GFreeDet2-2D-FastSAM	37.2	7.4	9.2	1.7	6.1	1.4
C6	GFreeDet2-6D-c (Ours)	-	Static	GFreeDet2-2D-FastSAM	37.2	29.4	25.9	25.1	26.8	12.4
C7	GFreeDet2-6D-c (Ours)	-	Static	GFreeDet2-2D-SAM2.1p	41.5	29.4	31.6	25.9	29.0	12.2
C8	GFreeDet2-6D-c (Ours)	-	Static	GFreeDet2-2D-SAMp	42.1	30.9	31.6	28.2	30.2	23.5
With Pose Refinement										
F1	GigaPose [47]	GenFlow [44]	CAD	-	-	26.8	41.1	25.6	31.2	5.3
F2	OPFormer	MegaPose [12]	CAD	CNOS [43]	-	-	39.2	26.2	-	-
F3	GFDet-6D (Ours)	GoTrack [46]	CAD	CNOS-FastSAM [43]	30.3	46.1	40.5	37.7	41.4	28.4
F5	GFreeDet2-6D-lite (Ours)	GoTrack-GS	Static	GFreeDet2-2D-FastSAM	37.2	42.3	39.6	40.0	40.6	10.7
F6	GFreeDet2-6D (Ours)	GoTrack-GS	Static	GFreeDet2-2D-FastSAM	37.2	48.7	42.5	44.2	45.1	22.6
F7	GFreeDet2-6D (Ours)	GoTrack-GS	Static	GFreeDet2-2D-SAM2.1p	41.5	47.1	47.1	44.7	46.3	21.4
F8	GFreeDet2-6D (Ours)	GoTrack-GS	Static	GFreeDet2-2D-SAMp	42.1	48.9	45.2	45.8	46.6	39.6

is smaller than that in AP_{2D}. All GFreeDet2-6D variants outperform previous methods such as GigaPose [47] and OPFormer. The best AP_{6D} of 46.6% is achieved by GFreeDet2-6D with GFreeDet2-2D-SAMp. Surprisingly, GFreeDet2-6D-lite-c, while running 2-4× faster, reaches 40.6% after refinement with the flow-based refiner GoTrack-GS, despite a coarse AP of only 6.1%. This performance is just 5–6% below the heavier variants and comparable to GFDet-6D.

Overall, GFreeDet2-6D establishes a strong model-free baseline with state-of-the-art performance by exploiting the strengths of CAD-based methods with minimal changes. Its performance and efficiency will continue to advance with ongoing progress in foundation models and pose estimation.

C. Ablation Studies

We present several ablation studies in Tab. III and Tab. IV. For 2D detection, all results are reported on BOP-H3, while for 6D detection, key ablations are performed on HOT3D using GFreeDet2-6D-lite-c for efficiency.

Effectiveness of PAPC. Tab. III (A0 *v.s.* B0) shows that, compared to GFreeDet's separate GS object workflows [29] for fisheye and pinhole images, our projection-aware perspective cropping (PAPC) unifies GFreeDet2-2D and improves AP_{2D} via higher-quality GS objects. Tab. IV (Row 1 *v.s.* Row 3) further confirms that applying PAPC to fisheye queries boosts coarse pose estimation accuracy in both 2D and 3D.

Onboarding Camera Type for HOT3D. Tab. III (B0 *v.s.* B1) and Tab. IV (Row 1 *v.s.* Row 2) show that using Aria for onboarding outperforms Quest3 on HOT3D, as Aria provides RGB images while Quest3 captures only grayscale.

Templates for 2D and 6D detection. Inspired by CNOS-FastSAM-Static (Tab. III A1) [43], sampled real onboarding images can already provides reasonable AP_{2D}. Our PAPC and appearance score further boost accuracy (C0), though adding more real images gives only marginal gains (C1). Combining GS-rendered templates and real images yields the best 2D detection results (D0), and reducing the real templates leads to slightly lower AP_{2D} (D1). For 6D detection, however, the trend is reversed: including real onboarding images degrades pose AP (Tab. IV Row 4 *v.s.* Row 1), likely because cross-domain features interfere with PCA dimension reduction.

Foundation Feature Extractor. Tab. III (E0 *v.s.* E2) shows that replacing DINOv2L with the updated DINOv3L [38] for feature extraction and matching reduces 2D performance. Therefore, we retain DINOv2 in all other experiments.

Mask Proposal Model and Morphological Post-Processing. Tab. III (E0-1) shows that SAM and SAM2.1 are slightly more accurate than FastSAM but considerably slower. While morphological post-processing (F1-2) degrades the performance of the FastSAM variant, it significantly enhances GFreeDet2-2D with SAM and SAM2.1 due to the superior quality of their initial masks.

V. CONCLUSIONS

This work has presented GFreeDet2, a strong baseline for RGB-based, model-free unseen 2D and 6D detection that leverages Gaussian splatting and vision foundation models. At its core, projection-aware perspective cropping (PAPC) unifies the GS workflow and consistently handles both pinhole and fisheye images during both GS training and coarse 6D detection. The full GFreeDet2 framework draws on many strengths of existing CAD-based approaches for 2D and 6D detection, allowing continuous improvement as advances are made at each stage. Currently, 6D detection efficiency remains far from real-time, partly due to numerous false positives in 2D detection. Moreover, unseen 2D detection performance continues to be a bottleneck for 6D detection, as similarly noted in [16].

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant No. 62406169, and in part by the China Postdoctoral Science Foundation under Grant No. 2024M761673. This paper used generative AI tools like ChatGPT for language editing. All intellectual content, analyses, and results are the original work of the authors.

REFERENCES

[1] S. Thalhammer, D. Bauer, P. Hönig, J.-B. Weibel, J. García-Rodríguez, and M. Vincze, "Challenges for monocular 6-d object pose estimation in robotics," *IEEE Transactions on Robotics (T-RO)*, vol. 40, pp. 4065–4084, 2024.

TABLE III: **Ablations of unseen 2D object detection on BOP-H3.** Best results are highlighted as 1st, 2nd, and 3rd.

Row	Method	HOT3D	HOPEv2	HANDAL	AP_{2D}	Time
A0	GFreeDet [29]	33.8	36.4	25.5	31.9	0.3
A1	CNOS-FastSAM-Static [43] (100 static onboard. images as templates)	37.3	34.3	30.4	34.0	0.4
B0	A0: separate GS pipeline → unified GS with PAPC	36.1	34.9	28.1	33.0	0.2
B1	B0: Quest3 onboard. seq. → Aria onboard. seq.	38.2	34.9	28.1	33.7	0.2
C0	A1 → templates cropped with PAPC; CNOS with appearance score	40.3	35.5	32.6	36.1	0.2
C1	C0: 100 static onboard. images → 162 static onboard. images	40.4	35.6	32.8	36.3	0.2
D0	B1: 162 GS templates → adding 64 static onboard. images (**GFreeDet2-2D-FastSAM**)	40.3	36.9	34.4	37.2	0.2
D1	D0 → 162 GS templates + 32 static onboard. images	40.1	36.5	33.9	36.8	0.2
E0	D0: FastSAM → SAM	38.1	39.0	36.2	37.8	1.6
E1	E0: SAM → SAM2.1	39.6	39.8	34.1	37.8	1.4
E2	E0: matching with DINOv2 → matching with DINOv3	35.4	34.9	34.2	34.8	2.1
F0	D0 → with mask post-processing	36.3	26.6	31.3	31.4	0.6
F1	E1 → with mask post-processing (**GFreeDet2-2D-SAM2.1p**)	42.1	45.2	37.3	41.5	1.8
F2	E0 → with mask post-processing (**GFreeDet2-2D-SAMp**)	42.7	44.0	39.7	42.1	2.3

TABLE IV: **Ablations of unseen 6D object detection on HOT3D.** Row 1 represents GFreeDet2-6D-lite-c (Tab. II C5). Best results are highlighted as 1st, 2nd, and 3rd.

Row	Onboard. seq.	Num. static onboard. im	Query PAPC	AP_{6D}	AP_{MSSD}	AP_{MSPD}
1	Aria	0	✓	7.4	0.9	13.9
2	Quest3	0	✓	6.8	0.8	12.8
3	Aria	0	✗	6.9	0.8	13.0
4	Aria	64	✓	5.2	0.7	9.7

[2] A. Ioannidou, E. Chatzilari, S. Nikolopoulos, and I. Kompatsiaris, "Deep learning advances in computer vision with 3d data: A survey," *ACM computing surveys (CSUR)*, vol. 50, no. 2, pp. 1–38, 2017.

[3] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "DeepIM: Deep iterative matching for 6D pose estimation," *International Journal of Computer Vision (IJCV)*, vol. 128, no. 3, pp. 657–678, 2020.

[4] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," *Robotics: Science and Systems Conference (RSS)*, 2018.

[5] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "Pvnet: Pixel-wise voting network for 6dof pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4561–4570.

[6] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "GDR-Net: Geometry-guided direct regression network for monocular 6D object pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 611–16 621.

[7] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, "CosyPose: Consistent multi-view multi-object 6D pose estimation," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 574–591.

[8] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6D object pose and size estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2642–2651.

[9] M. Tian, M. H. Ang Jr, and G. H. Lee, "Shape prior deformation for categorical 6D object pose and size estimation," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 530–546.

[10] X. Liu, G. Wang, Y. Li, and X. Ji, "Catre: Iterative point clouds alignment for category-level object pose refinement," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 499–516.

[11] R. Zhang, Z. Huang, G. Wang, C. Zhang, Y. Di, X. Zuo, J. Tang, and X. Ji, "Lapose: Laplacian mixture shape modeling for rgb-based category-level object pose estimation," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 467–484.

[12] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, and J. Sivic, "Megapose: 6d pose estimation of novel objects via render & compare," in *Conference on Robot Learning (CoRL)*. PMLR, 2023, pp. 715–725.

[13] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "FoundationPose: Unified 6d pose estimation and tracking of novel objects," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 17 868–17 879.

[14] E. P. Örnek, Y. Labbé, B. Tekin, L. Ma, C. Keskin, C. Forster, and T. Hodan, "Foundpose: Unseen object pose estimation with foundation features," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 163–182.

[15] A. Caraffa, D. Boscaini, A. Hamza, and F. Poiesi, "Freeze: Training-free zero-shot 6d pose estimation with geometric and vision foundation models," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 414–431.

[16] V. N. Nguyen, S. Tyree, A. Guo, M. Fourmy, A. Gouda, T. Lee, S. Moon, H. Son, L. Ranftl, J. Tremblay, E. Brachmann, B. Drost, V. Lepetit, C. Rother, S. Birchfield, J. Matas, Y. Labbe, M. Sundermeyer, and T. Hodan, "Bop challenge 2024 on model-based and model-free 6d object pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025.

[17] J. Lin, L. Liu, D. Lu, and K. Jia, "Sam-6d: Segment anything model meets zero-shot 6d object pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 27 906–27 916.

[18] J. Huang, H. Yu, K.-T. Yu, N. Navab, S. Ilic, and B. Busam, "Matchu: Matching unseen objects for 6d pose estimation from rgb-d images," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 10 095–10 105.

[19] V. N. Nguyen, T. Groueix, G. Ponimatkin, Y. Hu, R. Marlet, M. Salzmann, and V. Lepetit, "Nope: Novel object pose estimation from a single image," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 17 923–17 932.

[20] X. Liu, G. Wang, R. Zhang, C. Zhang, F. Tombari, and X. Ji, "UNOPose: Unseen object pose estimation with an unposed rgb-d reference image," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 22 023–22 034.

[21] M. Liu, S. Li, A. Chhatkuli, P. Truong, L. Van Gool, and F. Tombari, "One2any: One-reference 6d pose estimation for any object," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 6457–6467.

[22] T. Lee, B. Wen, M. Kang, G. Kang, I. S. Kweon, and K.-J. Yoon, "Any6d: Model-free 6d pose estimation of novel objects," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 11 633–11 643.

[23] J. Sun, Z. Wang, S. Zhang, X. He, H. Zhao, G. Zhang, and X. Zhou, "Onepose: One-shot object pose estimation without cad models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 6825–6834.

[24] X. He, J. Sun, Y. Wang, D. Huang, H. Bao, and X. Zhou, "Onepose++: Keypoint-free one-shot object pose estimation without cad models," *Conference on Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 35 103–35 115, 2022.

[25] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal Manipulation Interface: In-The-Wild Robot Teaching Without In-The-Wild Robots," in *Robotics: Science and Systems Conference (RSS)*, 2024.

[26] P. Banerjee, S. Shkodrani, P. Moulon, S. Hampali, S. Han, F. Zhang, L. Zhang, J. Fountain, E. Miller, S. Basol, *et al.*, "Hot3d: Hand and object tracking in 3d from egocentric multi-view videos," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 7061–7071.

[27] A. Guo, B. Wen, J. Yuan, J. Tremblay, S. Tyree, J. Smith, and S. Birchfield, "HANDAL: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions," in *IEEE/RJS International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 11 428–11 435.

[28] S. Tyree, J. Tremblay, T. To, J. Cheng, T. Mosier, J. Smith, and S. Birchfield, "6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark," in *IEEE/RJS International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 13 081–13 088.

[29] X. Liu, G. Wang, C. Li, Y. Li, C. Zhang, Z. Huang, and X. Ji, "Gfreedet: Exploiting gaussian splatting and foundation models for model-free unseen object detection in the bop challenge 2024," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025.

[30] B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, July 2023.

[31] A. Laurentini, "The visual hull concept for silhouette-based image understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 16, no. 2, pp. 150–162, 1994.

[32] M. Z. Irshad, M. Comi, Y.-C. Lin, N. Heppert, A. Valada, R. Ambrus, Z. Kira, and J. Tremblay, "Neural fields in robotics: A survey," *arXiv preprint arXiv:2410.20220*, 2024.

[33] Z. Yu, A. Chen, B. Huang, T. Sattler, and A. Geiger, "Mip-splatting: Alias-free 3d gaussian splatting," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 19 447–19 456.

[34] Z. Liao, S. Chen, R. Fu, Y. Wang, Z. Su, H. Luo, L. Ma, L. Xu, B. Dai, H. Li, *et al.*, "Fisheye-gs: Lightweight and extensible gaussian splatting module for fisheye cameras," in *European Conference on Computer Vision Workshops (ECCVW)*, 2024.

[35] C. Yang, S. Li, J. Fang, R. Liang, L. Xie, X. Zhang, W. Shen, and Q. Tian, "Gaussianobject: High-quality 3d object reconstruction from four views with gaussian splatting," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 6, pp. 1–13, 2024.

[36] V. Ye, R. Li, J. Kerr, M. Turkulainen, B. Yi, Z. Pan, O. Seiskari, J. Ye, J. Hu, M. Tancik, *et al.*, "gsplat: An open-source library for gaussian splatting," *JMLR*, vol. 26, no. 34, pp. 1–17, 2025.

[37] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research (TMLR)*, 2024.

[38] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, *et al.*, "DINOv3," *arXiv preprint arXiv:2508.10104*, 2025.

[39] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.

[40] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollar, and C. Feichtenhofer, "SAM 2: Segment anything in images and videos," in *International Conference on Learning Representations (ICLR)*, 2025.

[41] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang, "Fast segment anything," *arXiv preprint arXiv:2306.12156*, 2023.

[42] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[43] V. N. Nguyen, T. Groueix, G. Ponimatkin, V. Lepetit, and T. Hodan, "Cnos: A strong baseline for cad-based novel object segmentation," in *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2023, pp. 2134–2140.

[44] S. Moon, H. Son, D. Hur, and S. Kim, "Genflow: Generalizable recurrent flow for 6d pose refinement of novel objects," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 10 039–10 049.

[45] Q. Wang, R. Song, J. Li, K. Cheng, D. Ferstl, and Y. Hu, "Scflow2: Plug-and-play object pose refiner with shape-constraint scene flow," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 22 045–22 054.

[46] V. N. Nguyen, C. Forster, B. Tekin, S. Shkodrani, V. Lepetit, C. Keskin, and T. Hodaň, "Gotrack: Generic 6dof object pose refinement and tracking," *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025.

[47] V. N. Nguyen, T. Groueix, M. Salzmann, and V. Lepetit, "Gigapose: Fast and robust novel object pose estimation via one correspondence," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 9903–9913.

[48] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113.

[49] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing (TIP)*, vol. 13, no. 4, pp. 600–612, 2004.

[50] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2019, pp. 8026–8037.

[51] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An Accurate O(n) Solution to the PnP Problem," *International Journal of Computer Vision (IJCV)*, vol. 81, no. 2, p. 155, 2009.

[52] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.