IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

InstantPose: Zero-Shot Instance-Level 6D Pose Estimation From a Single View

Francesco Di Felice, Alberto Remus, Stefano Gasperini, Benjamin Busam, Lionel Ott, Stefan Thalhammer, *Member, IEEE*, Federico Tombari, and Carlo Alberto Avizzano, *Senior Member, IEEE*

***Abstract*—Object pose estimation using visual data is crucial for robotic interaction with the environment. Many existing instance-level methods are restricted by their requirements for 3D CAD models or multiple object views, which limits their flexibility and generalizability. Overcoming this limitation is critical to enhance the adaptability of pose estimation systems. In this work, a novel pipeline that leverages recent advances in reconstruction techniques is presented to address these challenges. To this end, Large Reconstruction Models (LRM) represent an advanced neural architecture capable of generating 3D object models from a limited set of views. Nevertheless, the resulting 3D models often lack relevant geometric and texture details due to insufficient input information. This research presents InstantPose, an innovative zero-shot instance-level pose estimation method that, building upon LRM, can determine the pose of unseen objects using as little as a single unposed RGB reference and RGB-D query images. Extensive experiments demonstrate that InstantPose achieves remarkable performance in object pose estimation on the YCB-V dataset, compared to methods conceived to rely on a geometrically perfect object's model. Furthermore, the 6D pose provided through the presented approach facilitates successful object grasping, highlighting its practical utility in robotic manipulation tasks.**

***Index Terms*—RGB-D perception, deep learning for visual perception, deep learning methods.**

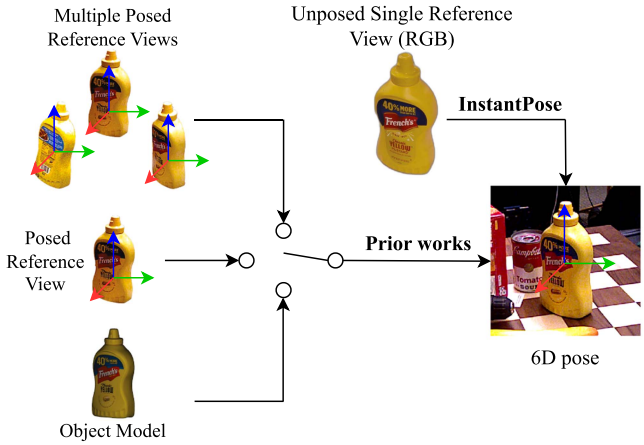

Fig. 1. **InstantPose**. Previous instance-level works have addressed 6D pose estimation for novel objects by using inputs such as geometrically perfect object models [4], full re-scans of objects with known poses [5], or support posed images [6]. In contrast, InstantPose introduces a novel approach that takes as input only a single unposed RGB view of the object and directly outputs the 6D pose estimation. This is achieved through a combination of pre-trained Vision Foundation Models, which provide general priors on the object's geometry, and a refinement process to enhance accuracy.

Received 4 December 2024; accepted 30 March 2025. Date of publication 21 April 2025; date of current version 5 May 2025. This article was recommended for publication by Associate Editor Lin Shao and Editor Abhinav Valada upon evaluation of the reviewers' comments. This work was supported by the Robotics Laboratory, Innovation Hub & Intellectual Property, Leonardo S.p.A., Genoa, Italy, under Grant LDO/CTI/P/0026995/21, July 2nd, 2021. *(Francesco Di Felice and Alberto Remus contributed equally to this work.) (Corresponding author: Francesco Di Felice.)*

Francesco Di Felice and Carlo Alberto Avizzano are with Department of Excellence in Robotics & AI, Mechanical Intelligence Institute, Scuola Superiore Sant'Anna, Pisa 56124, Italy (e-mail: francescodifelice96@gmail.com).

Alberto Remus is with the Department of Excellence in Robotics & AI, Mechanical Intelligence Institute, Scuola Superiore Sant'Anna, Pisa 56124, Italy, and also with Robotics Laboratory, Innovation Hub & Intellectual Property, Leonardo S.p.A., Genoa 16149, Italy.

Stefano Gasperini and Benjamin Busam are with the TUM School of Computation, Information and Technology, Technical University of Munich, 80809 Munich, Germany.

Lionel Ott is with the Department of Mechanical and Process Engineering, Autonomous Systems Lab, ETH Zürich, 8092 Zürich, Switzerland.

Stefan Thalhammer is with Industrial Engineering Department, UAS Technikum Vienna, 1040 Vienna, Austria.

Federico Tombari is with the TUM School of Computation, Information and Technology, Technical University of Munich, 80809 Munich, Germany, and also with Google Zürich, 8092 Zürich, Switzerland.

This article has supplementary downloadable material available at https://doi.org/10.1109/LRA.2025.3562788, provided by the authors.

Digital Object Identifier 10.1109/LRA.2025.3562788

I. Introduction

CORRECTLY estimating the 6D pose of objects (i.e., position and orientation in 3D space) is crucial for understanding the robot's environment, enabling it to perform manipulation tasks and improve navigation [1]. To this end, the so-called instance-level estimation has had remarkable success but it relies on specific, known objects, usually provided by a 3D Computer-Aided Design (CAD) model [2], [3]. This dependency restricts their use in real-world applications where obtaining manual 3D models is expensive and infeasible given the plethora of objects encountered in the open world.

As an alternative, category-level pose estimation methods have gained traction [7], [8]. They eliminate the necessity for detailed knowledge of individual object instances and enable models to generalize to objects of the same category. While category-level pose estimation is an established field, it has recently seen renewed interest, aligning it with other thriving computer vision tasks like object detection and instance segmentation [8]. However, current methods still depend on complete scans [5], object models [4], a combination of both [9] or support posed image [6], [10] for each new instance, restricting their practical usability. Recently, state-of-the-art approaches have tried to improve the generalizability of category-level approaches

by semantically matching different instances belonging to the same category using a pre-trained feature extractor [11], [12]. Following this direction, Zero123-6D [13] has recently applied zero-shot novel-view synthesizers to both improve the performance of zero-shot category-level 6D pose estimation and lower the data requirements by using a small set of reference views with associated pose to be compared to the query view. However, obtaining a set of posed reference views can hamper these methods' applicability to real-world problems. In fact, approaches like Structure from Motion may involve a manual or at most semi-automatic alignment of objects' frames [14].

To this end, this letter presents InstantPose, a novel approach inspired by improvements introduced in the 3D reconstruction and training-free pose estimation. InstantPose employs a pre-trained foundation model to estimate the 6D pose of objects accurately and subsequently grasp them in real-world scenarios. As shown in Fig. 1, a key advantage of InstantPose is its ability to eliminate the need for geometrically accurate object models or the availability of multiple or single posed reference images.

The proposed zero-shot pipeline is entirely training-free and requires only a monocular RGB-D query of the object and a monocular RGB reference view. The RGB reference is processed by a pre-trained single-view Large Reconstruction Model (LRM) [15], [16], yielding a noisy 3D model within seconds. Unlike methods designed for multi-view inputs [17], [18], the LRM is tailored for single-view reconstruction. Novel images are rendered in pre-defined poses, and the best match to the query is selected based on feature similarity. While the LRM output is a coarse 3D representation with geometric and texture distortions, category-level semantic feature correspondences are used to align texture and pose. The initial estimate is further refined through an online optimization technique to improve accuracy.

Overall, this pipeline shows better results in zero-shot instance-level pose estimation, surpassing the accuracy of models that rely on geometrically accurate 3D information. This leads, to the best of our knowledge, to the first work that employs Large Reconstruction Models for accurate single reference view zero-shot 6D pose estimation of a query object with downstream robotic applications.

In summary, the main contributions of this work can be outlined as follows:

- The formalization of a new challenging setting for zero-shot pose estimation of never-seen-before objects from one unposed single RGB reference view;
- We propose InstantPose, a novel method addressing 6D pose estimation in the (unposed) single-view zero-shot scenario;
- Robotic grasping of objects from single view 6D pose estimation and the extracted imperfect object priors.

II. Related Work

Instance- and category-level works are presented in the same section while a dedicated section presents various 3D content generation approaches, which are used to explain the rationale behind the choices made for the zero-shot reconstruction model. Table I briefly summarizes key novelties introduced in InstantPose with respect to relevant state-of-the-art instance-level works.

TABLE I
Comparison With Some State-of-The-Art Instance-Level Pose Estimation Methods

Method	Train	CAD	Rescan	Pose
OnePose [5]	once	no	yes	no
POPE [10]	no	no	no	yes
NOPE [6]	once	no	no	yes
Oryon [22]	once	no	no	yes
FreeZe [23]	no	yes	no	no
GigaPose [24]	once	yes	no	no
FoundationPose [9]	once	yes	no	no
InstantPose (ours)	no	no	no	no

InstantPose minimizes the assumptions since it is training-free, 3D-model-free (CAD or full point cloud), and does not rely on reference-posed images ("Pose" column).

A. 6D Pose Estimation

Monocular 6D pose estimation inherently requires the integration of 3D information to address the ill-posed nature of the task. This information typically comes as 3D CADs, which provide detailed vertex and face data. These 3D models represent a dense 3D input exploited by instance level pose estimator, which is generally challenging to obtain at inference time. Instance-level pose estimation requires the model of the specific object since it prioritizes accuracy by leveraging precise 3D structure during training [2], [3] and/or inference [19], achieving remarkable performance in recent years [20]. However, such methods struggle with generalization when confronted with slightly different or unseen samples, limiting their effectiveness in autonomous and robotic systems operating in unstructured environments.

On the other hand, category-level pose estimation [7], [8] aims to improve generalizability across multiple objects within a semantic category, even in the presence of substantial intra-class variation. Methods like Zero-Shot-Pose (ZSP) [11] leverage semantic features [21] to find correspondences between images from the same object category. This approach estimates a 6D pose from posed images of the category rather than each individual instance. However, ZSP's reliance on a reference set without enough viewpoint's variation can hinder accurate rotation estimation. To address this, generative models like Zero123-6D [13] synthesize novel views from a sparse RGB-only reference set (ranging from one to five images). This allows for the generation of additional viewpoints to aid reference selection, pose assignment, and 3D reconstruction [18], thus offering a more favorable starting point for refinement. However, the assumption that poses are linked to images is not always valid in real-world settings, and furthermore, the performance of category-level approaches still lag behind instance-level because the category-level strategy often sacrifices accuracy to address intra-class variability [7].

Zero-shot instance-level methods, such as GigaPose [24] and FreeZe [23], fundamentally rely on the assumption of having geometrically accurate 3D object models available, which significantly limits their practical applicability. In contrast, FoundationPose [9] aims to address this constraint. FoundationPose demonstrates superior results when compared to other leading methods, like Gen6D [25] or OnePose++ [5]. Nevertheless,

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

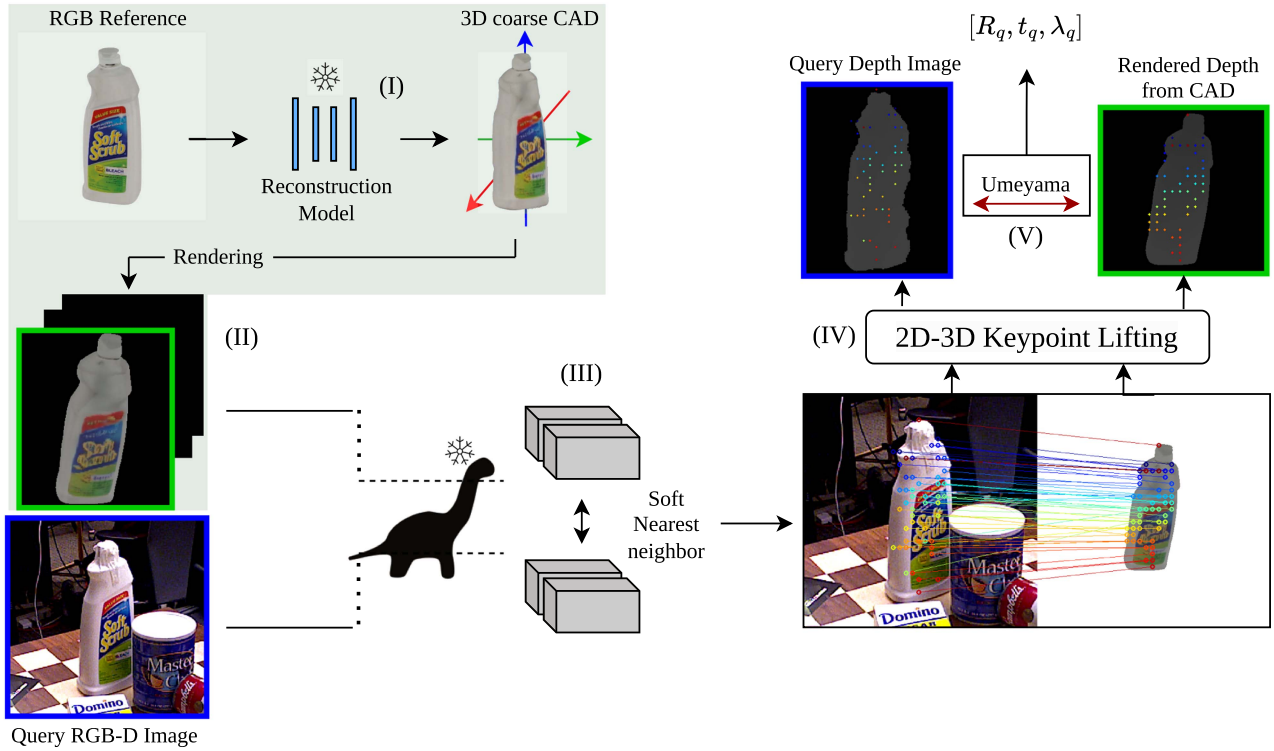

Fig. 2. **InstantPose** generates a 3D noisy object's model from a single RGB unposed reference view using a Large Reconstruction Model (I). Multiple renderings are then created from this model (II). For pose estimation, features of a query RGB-D view are compared with the rendered views using DINO [12] and a soft nearest neighbor algorithm (III), providing an initial pose estimate. 3D matching points are then recovered from both the query depth data and the object's model (IV). Finally, a deformation-aware optimization refines the pose, yielding the final 6D pose estimation (V). Output is R_q, t_q, λ_q. This approach enables efficient pose estimation from a single reference while accommodating potential discrepancies between the reconstructed and actual object.

FoundationPose still requires a sufficient number of posed reference views when the CAD model of the instance is not provided. In contrast to these works, InstantPose advances the state-of-the-art by utilizing a single RGB reference view of the object to estimate an accurate 6D pose and achieve zero-shot generalization to novel object instances while also retrieving and exploiting the object's 3D geometry.

In this direction, other state-of-the-art works have been proposed to relieve the multi-view constraint. In particular, Pope [10], Nope [6], and Oryon [22], which perform monocular pose estimation, assume the availability of a posed image for each new object to obtain the absolute pose. This assumption is too restrictive for most practical applications, as it limits the task to purely relative pose estimation. Additionally, Pope and Nope estimate translation only up-to-scale, while InstantPose automatically handles translation management during the refinement stage, enabling direct pose transfer to robotic grasping.

B. 3D Content Generation

Research in 3D scene representation has evolved from implicit models like NeRF [26] to more grounded representations such as 3D Gaussian Splatting [27], though they struggle with limited views [18]. Inspired by the success of 2D generative models like Stable Diffusion, methods have been adapted for 3D content generation, but challenges like slow optimization and hallucinations in symmetrical parts persist [28]. Zero123 [29] and subsequent works focused on enhancing single-view novel-view synthesis, while EscherNet [18] and SpaRP [17] addressed multi-view input setups to improve performance. Similar to InstantPose, SpaRP also tries to address the pose estimation problem. However, they rely on assumptions about multiple input views. In contrast, InstantPose specifically addresses the pose estimation challenge from a single-view perspective without relying on multiple input views.

These methods excel in RGB-only settings but struggle with monocular inputs and lack depth information. In contrast, depth-completion approaches [30] offer more accuracy potential but are limited to a narrower training set. Large Reconstruction Models (LRM) [15] and related approaches [16], [31] have shifted towards mesh optimization using multi-view diffusion models. However, these methods generally do not address pose estimation, while InstantPose utilizes them to derive object models from a single RGB view and estimates 6D pose, accounting for geometric inconsistencies.

III. Methodology

As shown in Fig. 2, InstantPose employs a three-step approach for zero-shot instance-level pose estimation. The process begins with acquiring a single RGB reference view of the target object. This reference view is then fed into a Large Reconstruction Model to generate a coarse 3D model of the object, from which multiple template views are rendered corresponding to assigned template poses (Section III-A). In the second step, a vision foundation model [12] establishes semantic correspondences between reference and query views. Based on these correspondences, the best-matching reference view is selected along with

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

its associated pose, providing an initial coarse pose estimation (Section III-B). The final step involves refining this coarse pose through an online optimization process. This refinement exploits 3D-3D correspondences [32], which are obtained by rendering the depth of the reference view in the pose of the best-matching view (Section III-C).

A. Reference Set Generation

Given one single unposed view $\mathbf{I}_1 \in \mathbb{R}^{H\times W\times 3}$ of a query object, a 3D large reconstruction model can generate a 3D object's model. Commonly [16], [31], LRMs first employ zero-shot novel-view-synthesizers to generate consistent views of the query object, and then they train a transformer-based sparse-view large reconstruction model to obtain a 3D object's model. Once the model is generated, it is used to render a set of reference RGB views $\tilde{\mathbf{I}}_{1:M} \in \mathbb{R}^{H\times W\times 3}$, in pre-defined template poses $\mathbf{P}_{1:M}$. The main issue to be addressed with LRM employment is that they provide a coarse 3D model of the object, often lacking of geometric consistency as explained in Section III-C, the geometric inconsistency has to be addressed to obtain an accurate pose estimation of the query object.

B. Feature and View Extraction

After rendering novel RGB views from the coarse object's model, it becomes feasible to derive semantic features by leveraging various foundation models. An essential characteristic of the semantic feature extractor is its ability to extract feature correspondences across images of varying object instances [21]. This capability is valuable for category-level tasks [13], but becomes crucial when reference images obtained from objects' models may present geometric inconsistencies between 2D query and reference view. As investigated in ZSP [11], semantic features can be derived from a query image $\mathbf{I_q} \in \mathbb{R}^{H\times W\times 3}$ and a reference image $\mathbf{I_r} \in \mathbb{R}^{H\times W\times 3}$, using the operator $\mathbf{F} : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{h'\times w'\times c}$, where h', w', and c represent the height, width, and channel size of the resulting feature map. These features can be utilized to establish correspondences between a query and reference using a soft nearest neighbor operator (SNN):

$$\text{SNN}\left(\mathbf{I_q}, \mathbf{I_r}\right) := \left\{\left(\mathbf{x}_q^k, \mathbf{x}_r^k\right) \mid \mathbf{x}_q^k \leftrightarrow \mathbf{x}_r^k\right\}, \tag{1}$$

where $\mathbf{x}_q^k \leftrightarrow \mathbf{x}_r^k, k \in \{1, \ldots, K\}$ denote the 2D pixel coordinates of the top K correspondences identified through SNN, in the query and reference images, respectively. These K correspondences are computed using the extracted query features $\mathbf{F}_q := \mathbf{F}(\mathbf{I}_q) \in \mathbb{R}^{h'\times w'\times c}$ and reference features $\mathbf{F}_r := \mathbf{F}(\mathbf{I}_r) \in \mathbb{R}^{h'\times w'\times c}$. The ranking is established based on the cyclical distance d between features, calculated as follows:

$$\bar{\mathbf{x}}_r = \Gamma(\mathbf{F}_r, \mathbf{F}_q[\mathbf{x}_q])) \in \mathbb{R}^2 \tag{2}$$

$$\bar{\mathbf{x}}_q = \Gamma(\mathbf{F}_q, \mathbf{F}_r[\bar{\mathbf{x}}_r]) \in \mathbb{R}^2 \tag{3}$$

$$d\left(\mathbf{x}_q, \mathbf{F}_q, \mathbf{F}_r\right) = \left\|\mathbf{x}_q - \bar{\mathbf{x}}_q\right\|_2 \in \mathbb{R} \tag{4}$$

where $\mathbf{F}_j[\mathbf{x}_j]$ represents the feature $\mathbf{F}_j$ corresponding to coordinate $\mathbf{x}_j$, and the function $\Gamma(\mathbf{F}_i, \mathbf{F}_j[\mathbf{x}_j]) =: \bar{\mathbf{x}}_i$ identifies the coordinate $\bar{\mathbf{x}}_i$ on the feature map $\mathbf{F}_i$ that is nearest (with respect the cyclical distance map introduced in ZSP [11]) in feature space to $\mathbf{F}_j[\mathbf{x}_j]$. As depicted in ZSP [11], the SNN operators ensures that K optimal correspondences are drawn between each pair of images: this is crucial to ensure that a sufficiently large number of correspondences is given.

We arrange $\mathbf{x}_q^k$ in ascending order based on the values of d. These K optimal correspondences between the query view and reference set can then be employed to identify the most suitable RGB image $\tilde{\mathbf{I}}_r^*$ in the reference set using:

$$\tilde{\mathbf{I}}^* = \underset{i\in\{1:N\}}{\text{argmin}} \sum_{k=1}^{K} d\left(\mathbf{x}_q^k, \mathbf{F}_q, \mathbf{F}_i\right), \tag{5}$$

where the aggregate cyclic distance for the top K feature matches is minimized. This selection process is crucial for identifying the reference view most similar to the given query, along with its associated pose. The pose is then refined through an iterative process to obtain the final 6D pose estimation.

C. Pose Refinement

In the context of 6D pose estimation within the outlined framework, a primary challenge emerges from the imprecise object model generated by the pre-trained large reconstruction model. Despite the hallucinated model's semantic consistency, it lacks perfect geometric fidelity, which precludes conventional instance-level pose estimation methods. To achieve accurate 6D pose estimation, employing refinement techniques that account for geometric discrepancies would be advantageous. In this vein, the challenge of instance-level pose estimation from approximate object models could draw insights from category-level pose estimation research, which has often utilized alternative approaches to address similar issues. This study adopts a robust least-squares method based on singular value decomposition named Umeyama [11], [32]. This refinement approach leverages 3D-3D correspondences by lifting the matching points identified during the semantic feature correspondence phase. Furthermore, RANSAC handles the presence of outliers in correspondences that are obtained. Umeyama algorithm is well suited for category-level tasks to handle geometric variability. For this reason, this work is adopted to address the issue of geometric and texture inconsistency induced by imprecise model generation. To this end, the depth $\mathbf{D}_r$ corresponding to the optimal RGB image $\tilde{\mathbf{I}}_i^*$ can be extracted by rendering the object's model at the appropriate reference pose. With the depth $\mathbf{D_r}$ of the reference image $\tilde{\mathbf{I}}_r^*$, its intrinsic matrix $\mathbf{K}_r$, and extrinsic parameters $\mathbf{R}_r, \mathbf{t}_r$, each 2D point $\mathbf{x}_r$ from the set of reference matches is back-projected to a 3D point $\mathbf{X}_r$ using:

$$\mathbf{X}_r = \mathcal{P}^{-1}\left(\mathbf{x}_r, D_r, \mathbf{K}_r, \mathbf{R}_r, \mathbf{t}_r\right), \tag{6}$$

where $\mathcal{P}$ represents the projection operator for a pinhole camera model. On the other hand, given depth information from camera input query view and camera intrinsics, 2D query matching points $\mathbf{x}_q$ are unprojected to corresponding 3D query points $\mathbf{X}_q$. Formally, given at each time step of the optimization process $\mathbf{R}$ and $\mathbf{t}$ the goal is to find optimal $\mathbf{R}_q, \mathbf{t}_q$ given each correspondence $\mathbf{X}_r \leftrightarrow \mathbf{X}_q$. The translation $\mathbf{t}_q \in \mathbb{R}^3$ is represented as a vector in Euclidean space while the rotation is represented as a 3×3 matrix.

The refinement procedure detailed in Algorithm 1 aims to recover the 6D pose transformation $\mathbf{R}_q, \mathbf{t}_q$ and scale λ_q to automatically adjust the scaling factor of the reference depth with respect to the query depth.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE II
COMPARISON OF POSE ESTIMATION METHODS

Method	Model-free	Single-view	Average Recall (%)
GigaPose-refined [24]	✗	✗	25.03
GigaPose [24]	✗	✗	12.72
FoundationPose [9]	✓	✗	30.26
Ours	✓	✓	**36.83**

Algorithm 1: Pose Refinement for 6D Pose Estimation.

Require: K Reference 3D points $\mathbf{X}_r$, K Query 3D points $\mathbf{X}_q$, T_s max number of RANSAC [33] iterations
Ensure: Optimized Rotation $\mathbf{R}_q$, Translation $\mathbf{t}_q$, Scale λ_q
Initialize $\mathbf{R}_q$ and $\mathbf{t}_q$ from the initial pose estimate
for $t = 1 \dots T_s$ **do**
  Pts centroids: $\bar{\mathbf{X}}_r = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_{r,i}$, $\bar{\mathbf{X}}_q = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_{q,i}$
  Recenter pts: $\mathbf{X}'_{r,i} = \mathbf{X}_{r,i} - \bar{\mathbf{X}}_r$, $\mathbf{X}'_{q,i} = \mathbf{X}_{q,i} - \bar{\mathbf{X}}_q$
  Compute covariance: $\mathbf{H} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}'_{r,i}\mathbf{X}'^{T}_{q,i}$
  Compute SVD: $\mathbf{H} = \mathbf{U}\mathbf{S}\mathbf{V}^T$
  if $\det(\mathbf{U}\mathbf{V}^T) < 0$ **then**
    Adjust $\mathbf{V}$: $\mathbf{V}[:,3] = -\mathbf{V}[:,3]$
  end if
  Compute Rotation: $\mathbf{R}_q = \mathbf{U}\mathbf{V}^T$
  Compute Scale: $\lambda_q = \mathrm{tr}(\mathbf{S})/(\frac{1}{n}\sum_{i=1}^{n} |\mathbf{X}'_{r,i}|^2)$
  Compute translation: $\mathbf{t}_q = \bar{\mathbf{X}}_q - \lambda_q \mathbf{R}_q \bar{\mathbf{X}}_r$
  RANSAC step for outlier rejection in $\{\mathbf{X}_r, \mathbf{X}_q\}$
end for
return $\mathbf{R}_q$, $\mathbf{t}_q$, λ_q

IV. EXPERIMENTS

A comparative analysis evaluates InstantPose against FoundationPose and GigaPose, focusing on challenging single-view pose estimation scenarios with imperfect 3D object models. The study highlights the method's performance in addressing the limitations of existing techniques. The performance of InstantPose on the YCB-V dataset [2] is comprehensively demonstrated in Table II, with Tables III and IV providing detailed object-specific performance insights. Qualitative evaluations are conducted to demonstrate the effectiveness of the approach, utilizing both the YCB-V dataset and real-world grasping experiments. The experiments on the YCB-V dataset correspond to the quantitative evaluations performed on the same dataset, providing complementary insights. The real-world experiments are carried out using a Franka Emika Panda manipulator to validate the practical applicability of the approach and assess the influence of reconstruction quality.

A. Dataset and Metrics

For the analysis of InstantPose's quantitative performance, the Average Recall (AR) metric [20] is evaluated on the YCB-Video (YCB-V) dataset [2]. The YCB-V dataset is characterized by complex scenarios, including cluttered, multi-object scenes with diverse object attributes, and is widely used in instance-level pose estimation studies [2], [24]. The AR metric is computed over three pose-error functions: VSD (Visible Surface Discrepancy), MSSD (Maximum Symmetry-Aware Surface Distance), and MSPD (Maximum Symmetry-Aware Projection Distance).

TABLE III
COMPARISON ON YCB-V DATASET [2] WITH STATE-OF-THE-ART METHODS ON THE AR METRICS USED IN THE BOP CHALLENGE, BOLD REPRESENTS BEST VALUES, UNDERLINED IS SECOND-BEST

	Average Recall (%)			
Instance	GP[24]	GP-R[24]	FP[9]	**Ours**
master_chef_can	4.19	11.51	<u>21.81</u>	**31.80**
cracker_box	7.31	<u>20.50</u>	5.58	**33.07**
sugar_box	1.67	11.18	<u>12.36</u>	**36.98**
tomato_soup_can	10.52	15.89	<u>30.95</u>	**33.60**
mustard_bottle	23.32	<u>47.72</u>	21.55	**62.12**
tuna_fish_can	12.77	4.96	**45.15**	<u>24.72</u>
pudding_box	10.05	<u>42.63</u>	12.17	**43.43**
gelatin_box	20.60	**53.08**	21.46	<u>43.84</u>
potted_meat_can	20.33	18.73	**24.57**	<u>21.04</u>
banana	6.46	**11.45**	0.15	<u>10.56</u>
pitcher_base	7.94	16.99	<u>17.81</u>	**27.48**
bleach_cleanser	20.34	38.77	<u>49.67</u>	**50.90**
bowl	25.57	57.65	**81.38**	<u>63.35</u>
mug	17.54	**35.35**	<u>27.98</u>	15.86
power_drill	12.00	59.18	**84.75**	<u>72.25</u>
wood_block	9.39	35.80	**88.17**	<u>80.01</u>
scissors	0.06	<u>4.59</u>	0.00	**4.72**
large_marker	32.22	**56.04**	0.84	<u>49.01</u>
large_clamp	9.36	**19.07**	13.55	<u>18.91</u>
extra_large_clamp	1.90	5.42	**19.51**	<u>11.96</u>
foam_brick	22.04	<u>26.48</u>	15.73	**40.12**
1st Places	0	5	6	**10**
2nd Places	0	5	6	**10**

TABLE IV
COMPARISON ON AR METRIC FOR POWER DRILL INSTANCE OF YCB-VIDEO DATASET [2]

	Average Recall [%] ↑			
CAD Quality	GP [24]	GP-R [24]	FP [9]	**Ours**
High-quality	12.00	59.18	**84.75**	72.25
Low-quality	5.09	14.04	58.73	**64.71**

A pose is considered correct with respect to an error function $e \in \{\mathrm{VSD}, \mathrm{MSSD}, \mathrm{MSPD}\}$ if $e < t_e$, where t_e is the threshold. The AR for an error function, AR_e, is calculated by averaging Recall across thresholds t_e and tolerances τ (specific to VSD). The dataset accuracy is computed as $AR_D = \frac{AR_{VSD}+AR_{MSSD}+AR_{MSPD}}{3}$ and overall accuracy is AR_C, the mean AR_D across core datasets.

B. Quantitative Results

The proposed InstantPose implementation utilizes InstantMesh as pre-trained LRM for its superior performance in single-view 3D reconstruction [16], while employing DINOv2 [12] for feature extraction. Object detection is performed with the use of pre-trained Grounding-Sam [34]. The employed InstantPose implementation considers 400 template views rendered from the 3D object model. Baseline approaches, that are not designed to estimate poses from a single reference (as shown in Table II) are more prone to fail when faced with this challenging scenario. To ensure a fair comparison with these baselines, the generated noisy object's model (computed by the pre-trained large reconstruction model) is provided as input to both FoundationPose and GigaPose after its creation. Another

challenge in comparison arises from the fact that neither baseline method addresses the issue of scaling factors: FoundationPose assumes the availability of ground-truth CAD models or multiple posed views (for experiment purposes, the model-based version is considered), while GigaPose operates solely at the RGB level. Consequently, these baselines struggle to accurately estimate poses when the object model requires appropriate rescaling for real-world applications. To address this limitation, the generated 3D object models are manually rescaled before being input to FoundationPose and GigaPose.

As a technical note, it is worth adding that, in real-world applications, our method does not need any manual or semi-automatic preliminary alignment, unlike approaches exploiting EscherNet [18], which requires input poses to be consistent with the ones used at training time. However, only for the evaluation, the 3D assets obtained from the LRM have to be aligned to the canonical frames in YCB-V; this is equally applied to all the reported methods.

Regarding GigaPose, the original work [24] proposes an approach that addresses pose estimation through a dual refinement process. This process incorporates a 2D refinement step, followed by an additional refinement utilizing MegaPose [35]. For the purposes of comparison in this chapter, both methods will be considered, even if InstantPose relies solely on one refinement process. Throughout the remainder of this chapter, "GigaPose" will refer to the version employing only the 2D refinement, while "Gigapose+Megapose" will denote the version that implements both refinement stages.

From Table II it can be seen how baseline approaches struggle to estimate correct poses on YCB-V dataset. Gigapose+Megapose remarkably improves the GigaPose predictions. However, all these methods are outperformed by InstantPose, which is able to produce accurate pose estimation even with geometric inconsistencies.

Table III extends the results reported in Table II over the 21 objects of the YCB-V dataset. As it appears from the last two rows, InstantPose reports the highest number of first and second places in the ranking against the other methods.

InstantPose does not deliver the best performance for every object because it depends on the mesh's quality. The quality of the mesh reconstructed by the LRM (in terms of both the texture and geometric consistency) is strictly influenced by the initial viewpoint [16]. Depending on the object's input view information, the quality can increase or decrease (Fig. 3). As expected, the quality of the generated mesh in turn affects the performance of InstantPose and baseline methods. In fact high-quality object models create a scenario that is more favorable for instance-level methods designed to address contexts with better geometries like FoundationPose [9] and GigaPose+MegaPose [24], as for *power-drill*, *tuna-fish-can* and *wood-block*. Conversely, InstantPose is optimized for scenarios with less precise meshes. These high-quality reconstructions are beneficial for such baseline approaches, allowing them to perform well despite the general advantage of our method.

To highlight such a behavior, Table IV has to be regarded together with Fig. 3. Here, it is possible to see how FoundationPose beats InstantPose when the *power_drill* instance is reconstructed better, while its performance drops when a more challenging perspective is considered as a reference image, this poses a great advantage for InstantPose since in general, it is not possible to choose in advance the viewpoint from which to acquire the reference image, used for the 3D reconstruction.

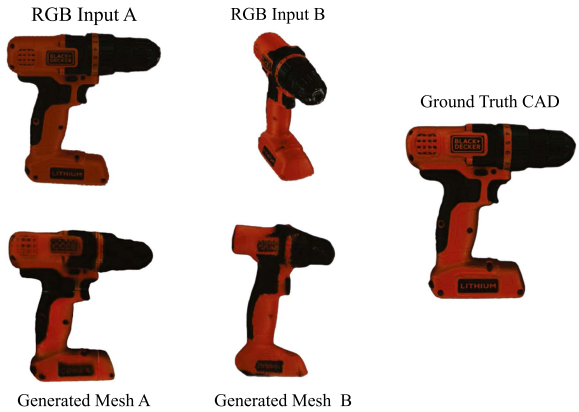

Fig. 3. Comparative visualization of different mesh generations for *power_drill* from different input viewpoints (input A and input B in black). The input viewpoint provided to the reconstruction model significantly influences the quality of the generated mesh [16]. Generated Mesh from A can be considered a high-quality reconstruction while the one from B is a low-quality one, compared to Ground Truth CAD on the right.

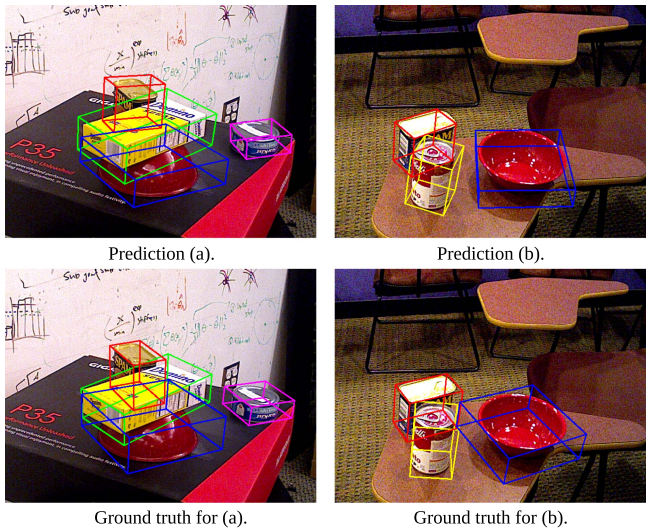

Fig. 4. Qualitative results on the YCB-V dataset, emphasizing the method's robustness. The first row shows predictions from InstantPose, and the second row shows the corresponding ground truth images. For objects with a cylindrical symmetry axis, rotational errors in the estimation may occur along this axis without however impacting subsequent robotic tasks.

From a runtime perspective, the pose estimation process takes 3.9 seconds for inference on an NVIDIA A4500 GPU. In comparison, on the same hardware, FoundationPose [9] requires 1.8 seconds, while GigaPose [24] takes approximately 8 seconds. Future work could further reduce this inference time by addressing performance bottlenecks, such as optimizing the soft nearest neighbor computation, which currently represents the most time-consuming step.

C. Qualitative Results

Fig. 4 showcases qualitative results from two scenes in the YCB-V dataset, aligning with the previously discussed quantitative analysis. The figure highlights the method's capability to handle the diverse shapes and textures of objects in the dataset. As evidenced, InstantPose accurately estimates poses even in complex multi-object scenes. Moreover, the method

Fig. 5. Grasping of different objects of the YCB-V [2] dataset, from the provided reference image (top-left in each picture), using the presented InstantPose method.

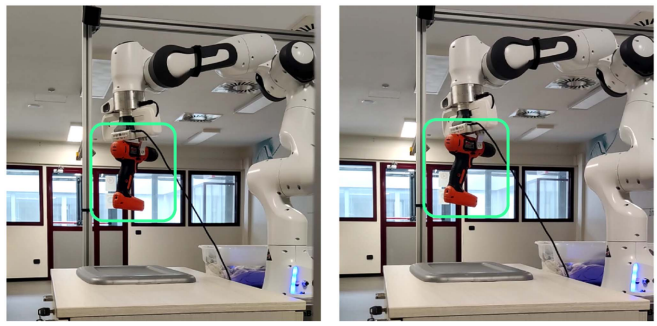

Fig. 6. Comparison between the grasping of YCB-V's *power_drill* object based on different 3D model generations. On the left, a firm grasp is granted by a proper reconstruction. On the right, despite the lower 3D quality, the grasping is quasi-identical with respect to to the left. InstantPose can compensate for geometrical imperfections and successfully catch the object.

demonstrates resilience to occlusions, as illustrated by objects such as the *bowl* in prediction (a) and *spam_can* in prediction (b), underscoring the robustness of the overall pipeline.

InstantPose can still face challenges when dealing with heavy occlusions beyond a certain threshold. For instance, in the case of the *spam_can* object, prediction (b) is well aligned with the ground truth, while prediction (a) suffers due to a significant occlusion, which impacts the method's performance. As seen in predictions and ground truth (b), reducing occlusion levels can lead to better performance. Additionally, it is important to note that the visual disparity between the 3D bounding boxes of the *red_bowl* object is attributed to the cylindrical symmetry property of this object. However this does not affect real-world application as discussed in Section IV-D and supplementary video.

D. Qualitative Grasping Results

The proposed approach demonstrates robust (with respect to mesh quality) and accurate pose estimation capabilities, even in challenging scenarios. This section aims to showcase the practical applicability of our approach through open-loop grasping experiments, as illustrated in Fig. 5. The emphasis lies not on the intricacies of the grasping strategy, but to qualitatively show InstantPose's robustness and accuracy in real-world scenarios.

Since InstantPose inherently also estimates the 3D model of the object, the grasping point can be computed based on the 3D bounding box of the mesh while exploiting the estimated pose to instruct the motion routine in cartesian coordinates. For each object one RGB reference view template has been previously acquired for the generation of the noisy object model, performed through the pre-trained InstantMesh [16].

The analysis demonstrates that our approach is capable of performing grasping from just one single RGB reference view and to manage the geometric noise induced in the generated object model from the pre-trained reconstructor. As illustrated in Fig. 5, the pose estimated by InstantPose is sufficiently accurate to enable the grasping of various types of objects, even in multi-object scenarios and under occlusions. This capability, previously highlighted qualitatively in Section IV-C, demonstrates the method's robustness.

As highlighted in Table III and in the supplementary video, quasi-symmetries can pose challenges in certain cases, such as with the mug object, where few correspondences are established on the thin handle compared to the larger body. Unlike Foundation-Pose [9] and GigaPose-refined [24], which avoid correspondence-based methods during the refinement stage (thus sidestepping this issue), InstantPose and GigaPose remain susceptible to this limitation. However, grasping remains feasible if the rotation error is confined to the cylindrical axis of the mug's body. As shown in Fig. 5, the estimated pose is still accurate enough to mug's enable successful grasping. To further assess the robustness of our method qualitatively in the presence of noisy object models, grasping experiments are conducted on the *power_drill* object under high and low quality conditions. As illustrated in Fig. 3, the quality of the mesh can vary depending on the initial viewpoints used for reconstruction. Table IV highlights that the quality of the reconstructed mesh directly influences the overall performance of the method. However, InstantPose exhibits strong robustness to varying mesh qualities, with an 8% performance loss, in contrast to the baseline methods up to 45%. For the grasping experiments, the *power_drill* was placed in the same location in both experiments. As depicted in Fig. 6, the robotic arm successfully grasped the object in both scenarios. While the grasp achieved with the high-quality mesh appears more secure compared to the grasp performed with the low-quality mesh, the outcomes are visually similar, demonstrating the method's robustness in handling disturbances.

All these qualitative experiments show the applicability of the method that could be potentially improved with specific grasping pipelines like [36] taking the pose as an input, which, however, is beyond the scope of this letter and left as further research direction.

V. Conclusion and Future Work

This work presents InstantPose, the first single-view 6D pose estimation pipeline, which outperforms state-of-the-art approaches by eliminating the need for pre-defined 3D models or multiple reference views, making it suitable for real-world applications like manipulation. It bridges instance and category-aware pose estimation by effectively managing photo-geometric imperfections during reconstruction using category-level methods. Future research could focus on the application of the method to more complex robotics tasks and addressing real-time applications while enhancing feature extraction for better geometric consistency. Additionally, eliminating depth requirements through RGB stereo cameras and extending the approach to multi-view sparse inputs could improve accuracy and consistency in 3D reconstruction.

REFERENCES

[1] F. D. Felice, S. D'Avella, A. Remus, P. Tripicchio, and C. A. Avizzano, "One-shot imitation learning with graph neural networks for pick-and-place manipulation tasks," *IEEE Robot. Automat. Lett.*, vol. 8, no. 9, pp. 5926–5933, Sep. 2023.

[2] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," in *Proc. Robot., Sci. Syst.*, 2018.

[3] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "GDR-Net: Geometry-guided direct regression network for monocular 6D object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16606–16616.

[4] E. P. Örnek et al., "FoundPose: Unseen object pose estimation with foundation features," in *Proc. 18th Eur. Conf. Comput. Vis.*, 2024, pp. 163–182.

[5] X. He, J. Sun, Y. Wang, D. Huang, H. Bao, and X. Zhou, "OnePose++: Keypoint-free one-shot object pose estimation without CAD models," in *Adv. Proc. Neural Inf. Process. Syst.*, 2022, pp. 35103–35115.

[6] V. N. Nguyen et al., "NOPE: Novel object pose estimation from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 17923–17932.

[7] A. Remus, S. D'Avella, F. D. Felice, P. Tripicchio, and C. A. Avizzano, "i2c-net: Using instance-level neural networks for monocular category-level 6D pose estimation," *IEEE Robot. Automat. Lett.*, vol. 8, no. 3, pp. 1515–1522, Mar. 2023.

[8] Y. Chen et al., "SecondPose: SE(3)-consistent dual-stream feature fusion for category-level pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 9959–9969.

[9] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "FoundationPose: Unified 6D pose estimation and tracking of novel objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 17868–17879.

[10] Z. Fan, P. Pan, P. Wang, Y. Jiang, D. Xu, and Z. Wang, "POPE: 6DoF promptable pose estimation of any object in any scene with one reference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2024, pp. 7771–7781.

[11] W. Goodwin, S. Vaze, I. Havoutis, and I. Posner, "Zero-shot category-level object pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 516–532.

[12] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res.*, 2024, [Online]. Available: https://openreview.net/forum?id=a68SUt6zFt

[13] F. D. Felice et al., "Zero123-6D: Zero-shot novel view synthesis for RGB category-level 6D pose estimation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2024, pp. 14204–14211.

[14] Y. Wang, X. He, S. Peng, H. Lin, H. Bao, and X. Zhou, "AutoRecon: Automated 3D object discovery and reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 21382–21391.

[15] Y. Hong et al., "LRM: Large reconstruction model for single image to 3D," in *Proc. Int. Conf. Learn. Representations*, 2024.

[16] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan, "InstantMesh: Efficient 3D mesh generation from a single image with sparse-view large reconstruction models," 2024. [Online]. Available: https://github.com/TencentARC/InstantMesh

[17] C. Xu et al., "SpaRP: Fast 3D object reconstruction and pose estimation from sparse views," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 143–163.

[18] X. Kong, S. Liu, X. Lyu, M. Taher, X. Qi, and A. J. Davison, "EscherNet: A generative model for scalable view synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 9503–9513.

[19] I. Shugurov, F. Li, B. Busam, and S. Ilic, "OSOP: A multi-stage one shot object pose estimation framework," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 6835–6844.

[20] T. Hodan et al., "BOP challenge 2023 on detection, segmentation and pose estimation of seen and unseen rigid objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 5610–5619.

[21] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9650–9660.

[22] J. Corsetti, D. Boscaini, C. Oh, A. Cavallaro, and F. Poiesi, "Open-vocabulary object 6D pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 18071–18080.

[23] A. Caraffa, D. Boscaini, A. Hamza, and F. Poiesi, "Freeze: Training-free zero-shot 6D pose estimation with geometric and vision foundation models," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 414–431.

[24] V. N. Nguyen, T. Groueix, M. Salzmann, and V. Lepetit, "GigaPose: Fast and robust novel object pose estimation via one correspondence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 9903–9913.

[25] Y. Liu et al., "Gen6D: Generalizable model-free 6-DoF object pose estimation from RGB images," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 298–315.

[26] B. Mildenhall, P.P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 99–106.

[27] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, 2023, Art. no. 139.

[28] C. H. Lin et al., "Magic3D: High-resolution text-to-3D content creation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 300–309.

[29] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3D object," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 9298–9309.

[30] Y. Zhang, X. Guo, M. Poggi, Z. Zhu, G. Huang, and S. Mattoccia, "CompletionFormer: Depth completion with convolutions and vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 18527–18536.

[31] S. Li et al., "Instant-3D: Instant neural radiance field training towards on-device AR/VR 3D reconstruction," in *Proc. 50th Annu. Int. Symp. Comput. Architecture*, 2023, pp. 1–13.

[32] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 04, pp. 376–380, Apr. 1991.

[33] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.

[34] S. Liu et al., "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in *Proc. Comput. Vis. – ECCV 2024*, 2025, pp. 38–55.

[35] Y. Labbé et al., "MegaPose: 6D pose estimation of novel objects via render & compare," in *Proc. 6th Conf. Robot Learn.*, 2022, pp. 715–725.

[36] H. Wen, J. Yan, W. Peng, and Y. Sun, "TransGrasp: Grasp pose estimation of a category of objects by transferring grasps from only one labeled instance," in *Proc. 17th Eur. Conf. Comput. Vis.*, 2022, pp. 445–461.