Training-Free 6D Robot Pose Estimation with Neural Memory Objects

Sebastian Jung*,1, Leonard Klüpfel*,1, Tjark Darius*,1,2, Rudolph Triebel1,3, and Maximilian Durner1

***Abstract*— Estimating the hand-eye transformation between a robot's base frame and a camera is a prerequisite for robot manipulation. Classical methods rely on dedicated calibration targets and controlled acquisition procedures; learning-based approaches remove the targets but require robot-specific training data that must be collected and a respective model trained for every robot. We present a training-free method for hand-eye transformation estimation based on just a single RGB image and the robot's kinematic description. Our key idea is that a robot arm at a fixed joint configuration is geometrically equivalent to a rigid object whose 3D shape is fully determined by its URDF and current forward kinematics. We exploit this to directly apply a training-free pose estimator: synthetic views of the arm are rendered from the URDF, efficiently encoded offline into a geometry-aware Neural Memory Object representation [1] and at inference a real query image is decoded to produce 2D–3D correspondences which together with PnP+RANSAC recovers the full camera-to-robot transformation from a single real RGB image. The approach requires no calibration targets, no robot-specific training data, and potentially generalises to any robot equipped with a URDF.**

I. INTRODUCTION

The manipulation of an object through a robotic system requires the precise knowledge of both 6D poses in a shared reference frame. Object poses are commonly estimated with dedicated perception algorithms in the camera coordinate frame. In contrast, the robot arm pose is analytically derived from its forward kinematics. The relationship to transform the two poses into a shared frame, which is also called *hand-eye transformation*, must thus be established such that a robot can reliably manipulate the desired object.

Classical calibration. Traditional methods estimate the hand-eye transformation by observing a known calibration target from multiple robot configurations [2], [3]. Tsai and Lenz [2] introduce the standard closed-form formulation; Strobl and Hirzinger [3] later propose an optimal reprojection-error formulation. Both require physical calibration objects and a dedicated acquisition procedure that must be repeated whenever this extrinsic transformation changes due relative movements which becomes especially challenging in dynamic or unstructured deployments.

Learning-based camera-to-robot pose estimation. More recent work eliminates calibration targets by regressing the pose directly from images of the robot arm. Lee et al. [4] train the keypoint network DREAM on synthetic renderings, demonstrating that domain randomisation bridges the sim-to-real gap. Klüpfel et al. [5] treat uncertain forward kinematics as a probabilistic prior in their PK-ROKED method, improving robustness on elastic robot arms with non-negligible kinematic errors. Lu et al. [6] address with CtRNet-X occlusion and varying illumination in real-world conditions. All three achieve strong results but share a fundamental limitation: a robot-specific network must be trained, requiring synthetic data generation for every new platform.

Fig. 1. We present a training-free approach for 6-DoF robot pose estimation that requires only a URDF model and joint configurations. Our method recovers the camera-to-robot transformation from a single real image without any robot-specific training. From left to right: input query image, ground-truth pose, and our estimated pose (rendered as URDF overlays).

Training-free 6-DoF object pose estimation. The object pose estimation community has meanwhile shown that 6-DoF pose can be recovered for *arbitrary* objects without per-object training. MegaPose [7] uses render-and-compare with a CAD model. FoundationPose [8] unifies pose estimation and tracking in a zero-shot framework trained once on large-scale synthetic data. NeMO [1] encodes a small set of RGB template views offline into a sparse, geometry-aware 3D point cloud; at query time a decoder produces dense 2D–3D correspondences from which EPnP+RANSAC recovers the full 6-DoF pose without per-object training or camera calibration during encoding.

Training-free calibration. Recent work applies similar ideas directly to robots. EasyHeC++ [10] initialises camera pose from pretrained features and refines it via differentiable rendering, but requires observing the robot across multiple joint configurations rather than a single image. FEEPE [11] matches foundation-model features between rendered CAD templates and a query image to estimate the end-effector pose online without training, using a multi-frame memory pool to handle occlusion. Neither method estimates the full camera-to-robot transformation from a single image.

Our contribution. A robot arm at a fixed joint configuration is geometrically a rigid object whose 3D shape is fully determined by its URDF and the current forward kinematics. This observation lets us recast deriving the hand-eye transformation as a training-free 6-DoF object pose

*Indicates equal contribution.

[1]Institute of Robotics and Mechatronics, German Aerospace Center (DLR), 82234 Wessling, Germany. firstname.lastname@dlr.de

[2]T. Darius is also with the Osnabrück University, 49074 Osnabrück, Germany.

[3]R. Triebel is also with the Institute for Anthropomatics and Robotics, Intelligent Robot Perception, Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany.

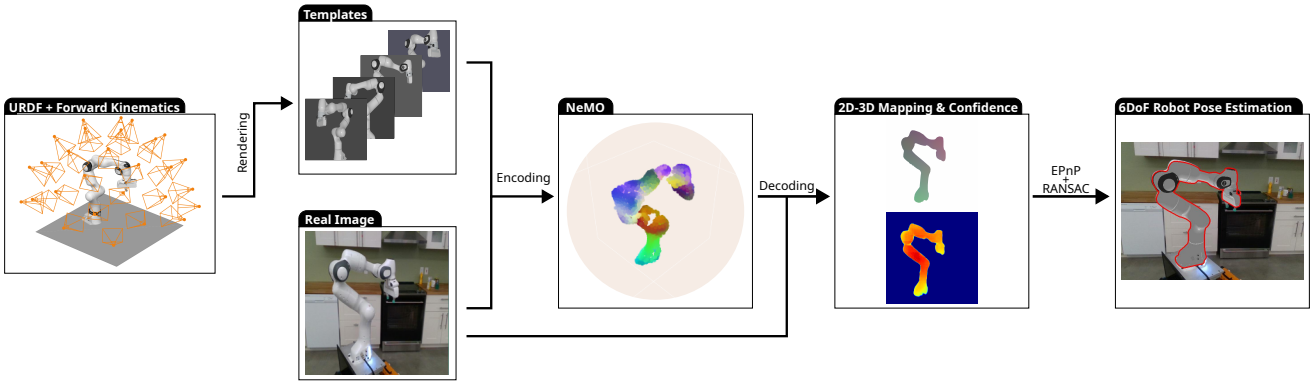

Fig. 2. **Method Overview.** Given the robot URDF and joint configuration $\mathbf{q}$, N synthetic views are rendered. These templates, together with the real query image $\mathbf{I}$, are encoded into a Neural Memory Object (NeMO) [1] to produce a sparse 3D point cloud $\mathcal{X}$ with associated features. The NeMO decoder maps the real query image $\mathbf{I}$ to dense 2D–3D correspondences and an estimated per-pixel confidence map. Finally, RANSAC and EPnP [9] are used to recover the 6-DoF camera-to-robot transformation $\mathbf{T}$ from the decoder output.

estimation: we render synthetic views of the whole arm from the URDF, encode them into a NeMO, and recover the full camera-to-robot transformation from a *single* RGB image via PnP+RANSAC. Unlike EasyHeC++, no multi-pose acquisition procedure is needed; unlike FEEPE, we estimate the complete rigid-body transformation from a single image rather than end-effector pose alone from a sequence of observations. We highlight that our approach is training-free, markerless, and has the potential to generalises to any robot with a URDF based on just one real RGB image. Additionally, we note that to the best of our knowledge our approach is the first to estimate the pose in such a context.

II. METHOD

A. Problem Formulation

Let $\mathbf{q} \in \mathbb{R}^n$ denote the joint angles of an n-DoF robot arm with a known URDF model. Forward kinematics $\mathrm{FK}(\mathbf{q})$ maps $\mathbf{q}$ to a 3D mesh of the arm expressed in the robot's base frame. Given a single RGB query image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and a camera with known intrinsic matrix $\mathbf{K}$, we seek the rigid transformation $\mathbf{T} = [\mathbf{R} \mid \mathbf{t}] \in SE(3)$ that maps points in the robot base frame to the camera frame, where $\mathbf{R} \in SO(3)$ and $\mathbf{t} \in \mathbb{R}^3$. step requires no camera parameters.

At a fixed configuration $\mathbf{q}$, the robot arm is geometrically equivalent to a rigid object whose 3D shape is fully determined by $\mathrm{FK}(\mathbf{q})$ and the URDF link geometries. We exploit this to treat hand-eye calibration as a 6-DoF pose estimation problem for a known rigid object, enabling the direct application of training-free pose estimation methods. An overview of our approach can be found in Fig 2.

B. URDF-Based Template Rendering

We render N synthetic views of the robot arm at configuration $\mathbf{q}$. Camera viewpoints are sampled uniformly on a sphere centered around the arm, covering a wide range of viewing directions. The rendering uses the textured URDF meshes directly, without additional material randomisation. The background colors are randomized for each template.

C. NeMO Encoding

The N rendered views are encoded into a Neural Memory Object (NeMO) [1]. NeMO constructs a sparse, geometry-aware 3D point cloud $\mathcal{X} = \{(\mathbf{s}_i, \hat{\mathbf{f}}_i)\}_{i=1}^{M}$, where $\mathbf{s}_i \in \mathbb{R}^3$ are surface points estimated jointly via an Unsigned Distance Field and $\hat{\mathbf{f}}_i$ are image patch features lifted into 3D by a multi-view transformer encoder. This multi-view encoding fuses information across all template views by cross attending each template to a pre-defined anchor image [1]. This anchor image also defines the object coordinate frame in which the 6D pose is expressed in. We chose the anchor image from a rendered image set with the highest visible surface score w.r.t. the look-at direction to the respective camera.

D. 6-DoF Pose Estimation via PnP+RANSAC

At inference, the NeMO decoder processes the query image $\mathbf{I}$ and produces two outputs: (i) a dense map of 2D–3D correspondences pairing each image pixel with a 3D object surface point, and (ii) a per-pixel confidence map $\mathbf{C} \in \mathbb{R}^{H\times W}$ indicating the reliability of each correspondence.

We apply RANSAC [12] to robustly identify the inlier set among the correspondences, discarding outliers caused by background clutter or partial occlusion. EPnP [9] then solves for the 6-DoF transformation $\mathbf{T} = [\mathbf{R} \mid \mathbf{t}]$ from the inlier 2D–3D pairs given the camera intrinsics $\mathbf{K}$. The output $\mathbf{T}$ is the hand-eye transformation: the pose of the robot base in the camera frame.

III. EXPERIMENTS

A. Setup

We evaluate on the *DREAM* dataset [4], which contains real-world images of the Franka Emika Panda arm in various configurations captured with a RealSense camera that is placed at 27 different position looking at the robot, yielding ≈32k frames total. Following prior work, we report the *ADD AUC* (area under the ADD curve, threshold 100 mm, higher is better) and mean *ADD* in mm (lower is better) [4]. For our experiments we render N=32 synthetic views with

TABLE I

ADD ON DREAM. RESULTS FOR SUPERVISED METHODS FROM [5] TABLE IV, EXCEPT ROBOPOSE FROM [6] TABLE II.

Method	Training-Free	AUC ↑	Mean (mm) ↓
DREAM [4]	✗	69.1	25
RoboPose [14]	✗	80.1	20
CtRNet [15]	✗	85.3	21
CtRNet-X [6]	✗	<u>86.2</u>	**14**
PK-ROKED + 1%	✗	**89.3**	<u>15</u>
PK-ROKED (synth.)	✗	75.5	34
Ours	✓	**46.8**	**69**

PyTorch3D [13] at the recorded joint configuration, encode them into a NeMO [1], and recover the 6-DoF pose with EPnP [9] inside RANSAC [12]. Note that in order to evaluate the aforementioned derived pose on the DREAM dataset, we need to align the respective coordinate systems. For details on this alignment step, we refer the reader to Jung et al. [1].

B. Comparison with State of the Art

Table I evaluates our approach against existing baselines on the DREAM benchmark. It is important to note that all current state-of-the-art methods rely on robot-specific training, whereas our method is the first to achieve 6-DoF robot pose estimation in a completely training-free manner with just a single real RGB image. We include these supervised results not as a direct performance competition, but to provide a context for the current upper bounds of the field.

We consider our performance on the DREAM dataset to be a strong initial baseline for this new "zero-training" paradigm. The performance gap between our method and supervised baselines is a logical consequence of the difference in problem setting. Supervised models are optimized for a narrow data distribution, whereas our approach leverages the general-purpose features of NeMO [1]. Since the domain of robotic arms differs significantly from the general-object training distribution of NeMO, our results represent an out-of-distribution application of neural memory representations with preliminary promising results.

We further analyze the translational error in Fig 3, where a significant peak is observed along the Z-dimension (depth). This localization error is a known characteristic of monocular RGB-based systems that lack explicit depth supervision or stereo geometric constraints.

To narrow the gap between training-free generalization and supervised precision, we plan to integrate a rigid-body refiner such as M3T [16] and incorporate outlier rejections. Such an extension remains strictly within our training-free constraints, as it only requires the kinematic chain and meshes already provided by the URDF, without necessitating any pre-training phase. Preliminary filtering already shows promising results by leveraging our confidence output to boost the ADD AUC to **52.0** with an improved mean of **50 mm** while keeping 80% of all samples.

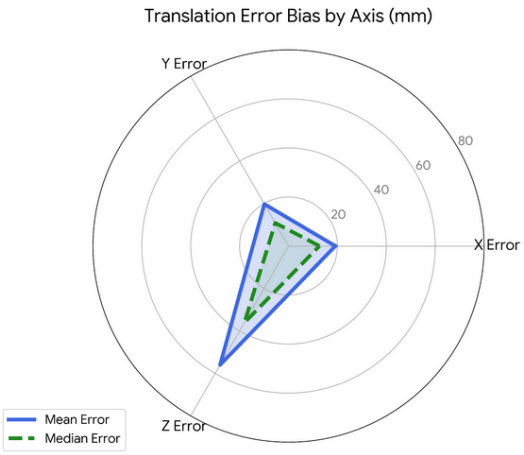

Fig. 3. **DREAM ADD Analysis.** Decomposition of the mean Average Distance (ADD) error across the X, Y, and Z camera axes (OpenCV convention, where the Z-axis represents depth). The Z-axis exhibits the highest error, which is expected as our method does not utilize explicit depth information during pose estimation.

error being greater on this direction makes sense as our approach is only RGB based.

C. Qualitative Results

Figure 4 presents additional qualitative results. Columns (a) and (b) demonstrate that our approach can produces accurate robot pose estimates. Column (c) illustrates a minor rotation error; these slight deviations can potentially be improved using local refinement methods [16]. Finally, column (d) highlights an instance of catastrophic failure. Investigating the root causes of these significant misalignments is a primary focus for future research. To ensure safe robotic deployment, we aim to develop automated failure-detection mechanisms. Specifically, we will investigate evaluating the reliability of the pose estimation by leveraging the model's predicted confidence maps, alongside solver-specific metrics such as the RANSAC inlier ratio and the mean reprojection error from the PnP algorithm.

IV. CONCLUSION

We demonstrate a training free 6D robot pose estimation pipeline that we apply to solve the hand-eye transformation. Our key idea is to frame this unknown transformation as a 6-DoF object pose estimation problem which we propose to solve by applying a NeMO [1] based pipeline. The key enabler is the observation that a robot arm at a known joint configuration is a rigid object with a fully specified 3D model, making URDF-driven template rendering a direct substitute for per-object training data. The resulting pipeline is immediately applicable to any robot equipped with a URDF, with no data collection, and no fine-tuning as demonstrated on the DREAM dataset. Additionally, our approach is markerless and requires only one real-world image captured with a standard RGB camera.

Limitations. The current approach requires a separate NeMO encoding for each distinct joint configuration, since

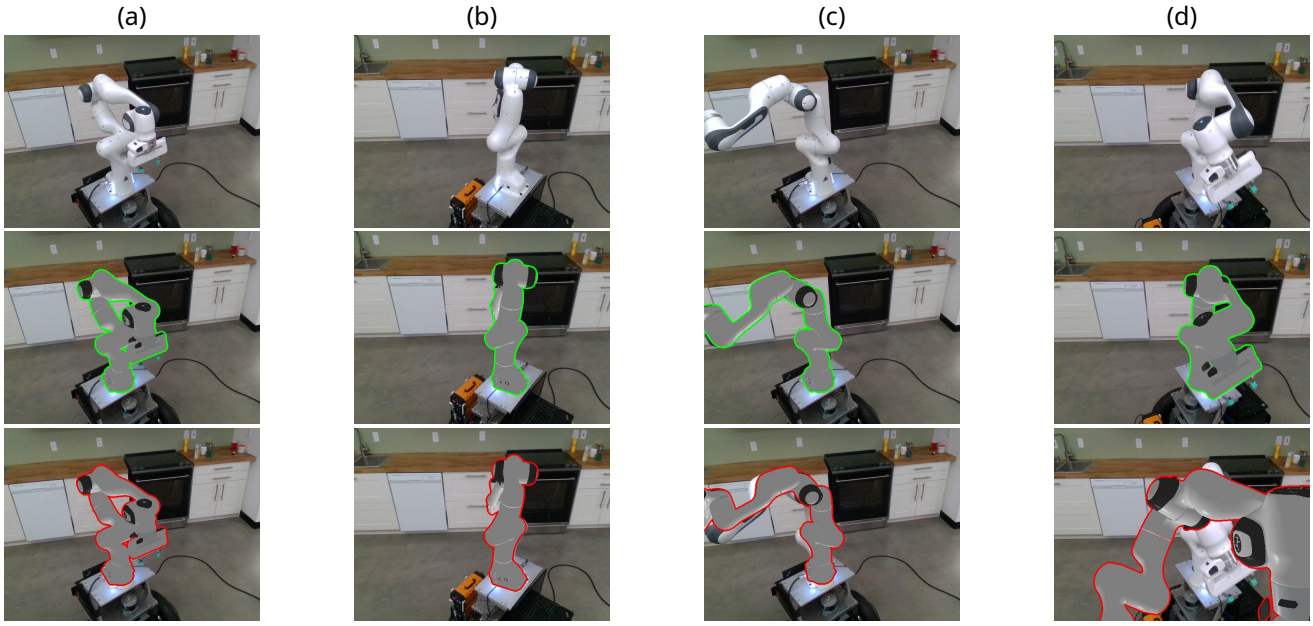

Fig. 4. **Qualitative results on DREAM.** We show the input query images (top row), ground-truth pose overlays (middle row), and our estimated robot poses rendered as URDF mesh overlays (bottom row). Columns (a) and (b) demonstrate accurate estimation across different configurations. Column (c) illustrates a minor alignment deviation, while (d) showcases a catastrophic failure where the estimated pose is significantly misaligned.

the arm's geometry changes with joint angles. Although encoding is a one-time offline step, it may be prohibitive when many configurations must be supported in real time. Additionally, NeMO inherits known limitations of foundation models in texture discrimination, which may reduce correspondence quality on uniformly coloured or symmetric robot links.

Future work. We plan to integrate a rigid body pose refiner, such as M3T [16], to boost the precision of our approach while staying training-free. Additionally, we investigate filtering and post-processing steps by leveraging our confidence output to further improve our performance. Eventually, we intend to conduct a full quantitative evaluation on other real robot platforms in addition to further robotic 6D pose estimation benchmarks.

References

[1] S. Jung, L. Klüpfel, R. Triebel, and M. Durner, "Finding NeMO: A geometry-aware representation of template views for few-shot perception," in *International Conference on 3D Vision*, 2026.

[2] R. Y. Tsai and R. K. Lenz, "A new technique for fully autonomous and efficient 3D robotics hand/eye calibration," *IEEE Transactions on Robotics and Automation*, vol. 5, no. 3, pp. 345–358, 1989.

[3] K. H. Strobl and G. Hirzinger, "Optimal hand-eye calibration," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006, pp. 4647–4653.

[4] T. E. Lee, J. Tremblay, T. To, J. Cheng, T. Mosier, O. Kroemer, D. Fox, and S. Birchfield, "Camera-to-robot pose estimation from a single image," in *IEEE International Conference on Robotics and Automation*, 2020.

[5] L. Klüpfel, L. Burkhard, A. E. Reichert, M. Durner, and R. Triebel, "Seeing through uncertainty: Robot pose estimation based on imperfect prior kinematic knowledge," *IEEE Transactions on Robotics*, vol. 41, pp. 4459–4478, 2025.

[6] J. Lu, Z. Liang, T. Xie, F. Richter, S. Lin, S. Liu, and M. C. Yip, "Ctrnet-x: Camera-to-robot pose estimation in real-world conditions using a single camera," in *2025 IEEE International Conference on Robotics and Automation*. IEEE, 2025, pp. 1914–1920.

[7] 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*, 2023, pp. 715–725.

[8] 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*, 2024, pp. 17 868–17 879.

[9] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An accurate O(n) solution to the PnP problem," *International Journal of Computer Vision*, vol. 81, no. 2, pp. 155–166, 2009.

[10] Z. Hong, K. Zheng, and L. Chen, "EasyHeC++: Fully automatic hand-eye calibration with pretrained image models," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2024.

[11] T. Wu, J. Zhang, S. Liang, Z. Han, and H. Dong, "Foundation feature-driven online end-effector pose estimation: A marker-free and learning-free approach," in *2025 IEEE International Conference on Robotics and Automation*. IEEE, 2025, pp. 1921–1928.

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

[13] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari, "Accelerating 3d deep learning with pytorch3d," *arXiv:2007.08501*, 2020.

[14] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, "Single-view robot pose and joint angle estimation via render & compare," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8456–8465.

[15] J. Lu, F. Richter, and M. C. Yip, "Markerless camera-to-robot pose estimation via self-supervised sim-to-real transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 296–21 306.

[16] M. Stoiber, M. Sundermeyer, W. Boerdijk, and R. Triebel, "A Multi-body Tracking Framework - From Rigid Objects to Kinematic Structures," *arXiv*, Feb. 2023.