# High-Accuracy Tactile Pose Estimation for Connector Assembly

Antonia Bronars[1], Radu Corcodel[2], and Devesh Jha[2]

## I. INTRODUCTION

Existing industrial systems often rely on specialized end effectors that grasp objects in pre-defined poses. Designing systems that can solve high-precision tasks with simple grippers is an important goal, which high-accuracy in-hand pose estimation can ease. Image-based tactile sensors hold promise for this task, but high-accuracy tactile pose estimation from arbitrary grasps remains challenging for several reasons. First, many grasps are inherently ambiguous [1] without additional information from vision [2], extrinsic contacts [3], or multiple grasps [4]. Second, training pose estimation models from real data is expensive [5], whereas sim2real with RGB tactile images is difficult [6]. Motivated by these challenges, we present a solution for high-accuracy tactile pose estimation with the following contributions:

1) We use tactile depth images as an intermediate representation between binary masks [1] and RGB to regress discrete pose distributions.
2) We introduce a refinement network to improve the accuracy beyond the discrete pose resolution.
3) We introduce a suite of data augmentations that allow Depth2Pose to sim2real with high fidelity.
4) We introduce a simple ambiguity detection method to identify grasps that can be localized accurately.
5) We demonstrate Depth2Pose on a connector assembly task, and show that for some connectors, we achieve high success rates with a simple force controller.

## II. METHOD

We propose Depth2Pose for object pose estimation (Figure 1). First, we grasp the connector and use tactile depth images to estimate a discrete hand-object pose distribution. Next, a refinement network improves accuracy. Finally, we assess ambiguity in the grasp using the pose distribution.

**Inferring tactile depth images from RGB.** We infer tactile depth images from RGB tactile images using the GelSight Inc. implementation [7] of depth reconstruction in which a pre-trained neural network extracts RGB surface gradients, then Poisson reconstruction returns the depth images.

**Estimating discrete distributions over object pose.** We adapt the custom simulator in [1], [8] to return tactile depth images rather than binary masks over the region of contact. We then save a discrete set of possible grasps on the object, sampled with 2.5mm of translational resolution and 6 degrees of rotational resolution about the grasp axis. Each grasp is a tuple containing the rendered tactile depth images, the grasp width, and the object pose. We follow [1], [8] and learn an object-dependent function in simulation to match observed grasps against the precomputed set of grasps, using supervised learning. We leverage the same encoder architecture and loss function as [1], [8], and refer the reader to this seminal work for details. We design a suite of data augmentations for tactile depth images to ease sim2real transfer:

1) **Penetration depth.** We randomize the penetration depth, $\Delta d$, from 0.6mm to 1.5mm.
2) **Out-of-plane rotation.** We randomly tilt the object into and out of the plane of the sensor, by up to 8 degrees.
3) **Background noise.** We randomly reduce the brightness (by subtracting a random value from 0 to 50 from each pixel) to simulate background noise.
4) **Enhanced edges.** Because depth images are reconstructed from the gradients of RGB images, edges appear more pronounced than flat surfaces. We enhance the edges using the Sobel operator with kernel size 3, then scale the edge image by a random factor from 0 to 3, and superimpose it with the original image.
5) **Blurring.** We apply a Gaussian blur with a random kernel size sampled from 11 to 21.

**Pose Refinement Network.** We refine the pose estimate beyond the resolution of the discrete distribution using a pose refinement network trained per-object. After computing the discrete distribution, the pose refinement network takes the most likely grasp and the observed grasp as inputs, and

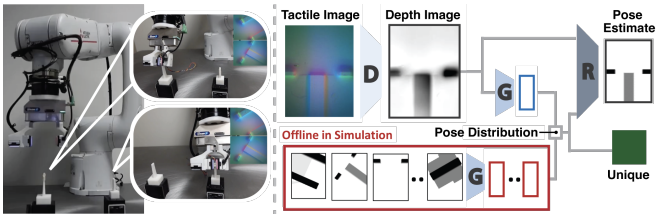


Fig. 1: **System and Depth2Pose overview.** We evaluate Depth2Pose in the real world through a connector assembly task (left). Depth2Pose (right) consists of several phases: converting RGB tactile images to depth images (D), estimating a discrete distribution over object pose (G), refining the pose estimate (R), and predicting whether the pose distribution is unique or ambiguous.

[1]Antonia Bronars is with Massachusetts Institute of Technology, Cambridge, MA, USA. and Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA. bronars@mit.edu

[2]Radu Corcobdel and Devesh K. Jha are with Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA. {corcodel,jha}@merl.com

This work was fully supported by Mitsubishi Electric Research Labs (MERL)

Fig. 2: **Depth2Pose is evaluated on four connectors.** We call them, from left to right, `Black`, `Six`, `Four`, and `M-connector`.

regresses the relative transformation in $SE(2)$. We use a ResNet-50 encoder with the average pooling layers removed to preserve spatial information, with weight sharing, to extract the image features. We then pass the concatenated feature vector through three fully connected layers, and output the relative transformation in $SE(2)$. We apply the same suite of data augmentations described above during training.

**Inferring Ambiguity.** To infer whether a grasp produces a unique pose estimate, we check for consensus between the two poses with the highest likelihood in the pose distribution; if the two most likely poses are adjacent, we consider the pose estimate as unique.

## III. EXPERIMENTS AND RESULTS

**Pose estimation experiments.** We conduct our experiments using a Mitsubishi Assista robot with a WSG-32 gripper and GelSight Mini tactile sensing fingers. We fix the connectors the world frame, and grasp in a range of gripper/object relative poses. We consider rotations of $\pm 40$ degrees, and translations in world X and Z until the robot loses contact with the connector. We sample grasps at 1mm of translational resolution and 1 degree of angular resolution, obtaining datasets of 80 to 300, depending on the object size. We measure pose error with the ADD metric [9], and find several key conclusions.

First, both the discrete pose estimation and refinement modules of Depth2Pose transfer well to the real world (Figure 3). Second, we break out the pose estimation results by the ambiguity detection criteria (Figure 4, top) and find that the metric is a practical approach for identifying grasps that can be localized accurately. Finally, we highlight the reliability of our pose refinement module for refining pose estimates from partial contacts. The refinement module improves the localization performance for every object, in sim and real (Figure 3), and for unique and ambiguous grasps (Figure 4).

**Connector assembly experiments.** We use a similar setup as in the pose estimation experiments, and deploy the robot to grasp the connector from an unknown and arbitrary pose. Depth2Pose estimates the discrete pose distribution, and predicts whether the grasp is unique. The robot releases the connector and samples new grasps until it finds a unique grasp. Then, Depth2Pose refines the pose. To assemble the connectors, the robot aligns the estimated frame of the grasped connector with the assembly axis of the fixed connector, and moves until a force threshold is reached.

We consider three outcomes for the connector assembly experiments: success, near success, and failure. Near successes are assembly attempts that fail by a small margin, rather than localization failures. For `M-connector` and `Black`, Depth2Pose with a simple force controller results in about 70% success (Figure 5, left). We also visualize the distribution of grasp attempts for each connector (Figure 5, right), and find that Depth2Pose can localize each connector with an average of two or less grasp attempts. Furthermore, for all four connectors, the majority of unsuccessful trials result in near successes, which can be resolved in future work with more sophisticated assembly controllers.

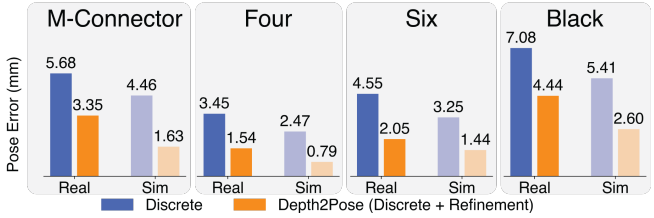


Fig. 3: **Depth2Pose retains performance in sim2real.** For all four connectors, Depth2Pose retains high performance in the real world for both discrete pose estimation and refinement.

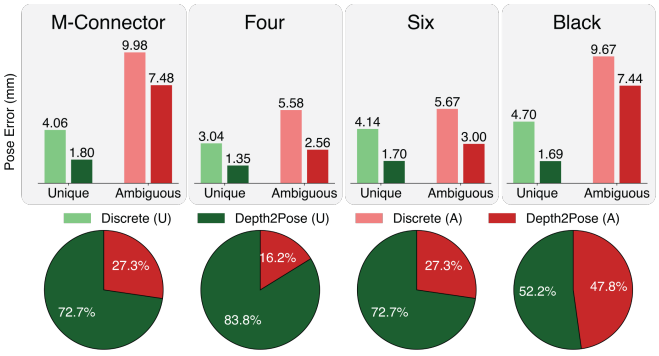


Fig. 4: **Depth2Pose distributions can detect unique grasps, which are easier to localize.** Depth2Pose achieves lower localization error for unique grasps than ambiguous ones (top). A meaningful fraction of grasps, between 52% and 84% depending on the object, are detected as unique (bottom).

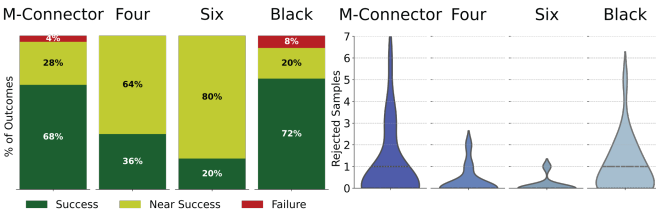


Fig. 5: **Connector assembly results.** Depth2Pose with a simple force controller achieves around 70% success for `M-connector` and `Black`. The success rate is lower for `Four` and `Six`, which have tigher assembly tolerance (left). Depth2Pose can localize each connector with an average of two or less grasp attempts (right).

## REFERENCES


[1] M. Bauza, A. Bronars, and A. Rodriguez, "Tac2pose: Tactile object pose estimation from the first touch," *arXiv preprint arXiv:2204.11701*, 2022.

[2] S. Suresh, H. Qi, T. Wu, T. Fan, L. Pineda, M. Lambeta, J. Malik, M. Kalakrishnan, R. Calandra, M. Kaess *et al.*, "Neuralfeels with neural fields: Visuotactile perception for in-hand manipulation," *Science Robotics*, vol. 9, no. 96, p. eadl0628, 2024.

[3] S. Kim, A. Bronars, P. Patre, and A. Rodriguez, "Texterity–tactile extrinsic dexterity: Simultaneous tactile estimation and control for extrinsic dexterity," *arXiv preprint arXiv:2403.00049*, 2024.

[4] S. Suresh, Z. Si, S. Anderson, M. Kaess, and M. Mukadam, "Midastouch: Monte-carlo inference over distributions across sliding touch," in *Conference on Robot Learning*. PMLR, 2023, pp. 319–331.

[5] J. Zhao, M. Bauza, and E. H. Adelson, "Fingerslam: Closed-loop unknown object localization and reconstruction from visuo-tactile feedback," *arXiv preprint arXiv:2303.07997*, 2023.

[6] C. Higuera, B. Boots, and M. Mukadam, "Learning to read braille: Bridging the tactile reality gap with diffusion models," *arXiv preprint arXiv:2304.01182*, 2023.

[7] D. Shure, "GelSight Robotics Software," Nov. 2021. [Online]. Available: https://github.com/gelsightinc/gsrobotics

[8] M. B. Villalonga, A. Rodriguez, B. Lim, E. Valls, and T. Sechopoulos, "Tactile object pose estimation from the first touch with geometric contact rendering," in *Conference on Robot Learning*. PMLR, 2021, pp. 1015–1029.

[9] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I 11*. Springer, 2013, pp. 548–562.