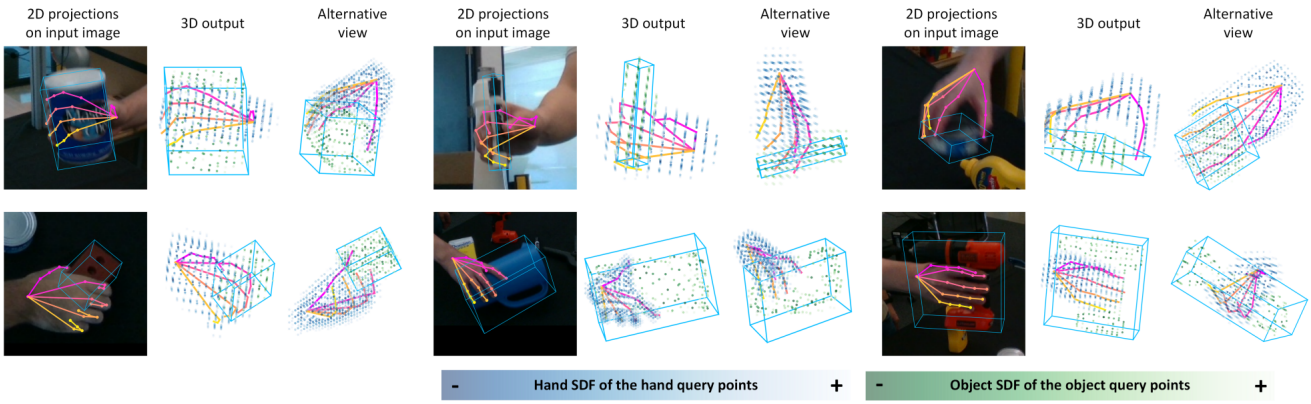
Addressing Limitations in 3D Hand-Object Interaction with HOISDF

David Robinson

Introduction

The HOISDF model [1] predicts the 3D hand and object keypoints from a single RGB image for hand-object interactions and was accepted by CVPR 2024. It first estimates signed distance fields [2] for both the hand and object to build the shape of each with sampled points. A signed distance field is a scalar field that represents the shortest difference from any point in a 3D space to the surface, where the sign indicates whether the point is inside (negative) or outside (positive) the object. The sampled points on the surface are then used to predict the hand joints and shape, as well as the object's position and orientation in the image.



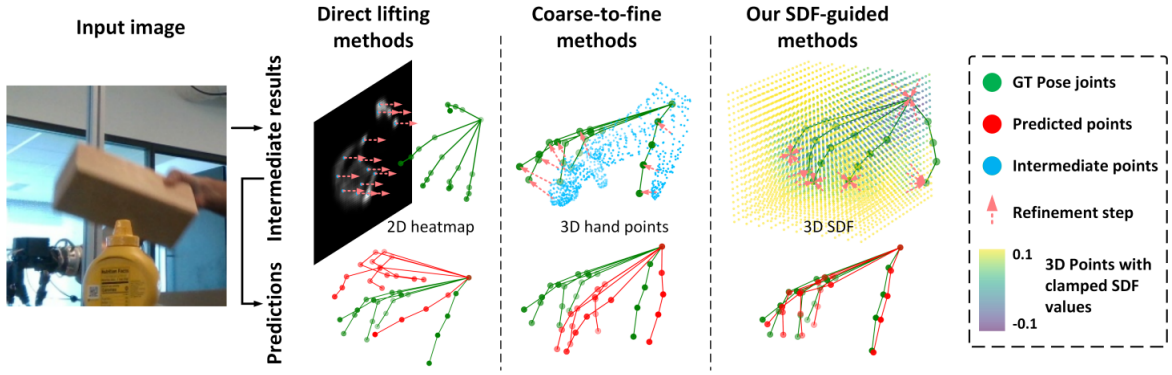
Background

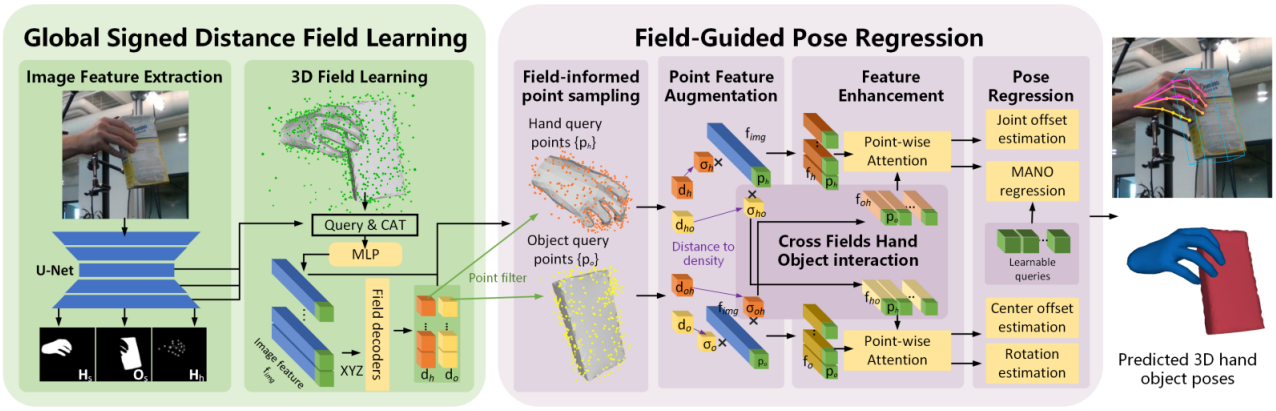
3D pose estimation is the field of estimating 3D coordinates of keypoints or joints from various input data, such as 2D images, depth maps, or WiFi signals. This technology plays a critical role in fields like robotics [3], augmented reality [4], and healthcare [5], where precise tracking and analysis of movement are essential.

One category of pose estimation with an RGB image is regression-based methods, where the 3D coordinates are regressed from the input image. For example, the ResNet [6] model passes the input image through convolution layers to generate features and then those features are regressed to the 3D coordinates using a fully connected layer. Heatmap-based methods, such as EgoTAP [7], are also very common where a heatmap of the 2D coordinates is predicted first and then a separate network directly lifts them to 3D. This method does not perform as well in occluded scenarios because it depends on the accuracy of the 2D coordinates and does not use context from the original image. Also, there is ambiguity in the projection as there are multiple possible ways to lift a set of 2D coordinates to 3D [8].

Motivation

Estimating the 3D poses of a hand and an object from a single camera is challenging due to the frequent occlusions during interaction. Most existing methods struggle in these scenarios because they rely on limited shape information or explicit representations like meshes, which only provide local context. Direct lifting methods predict 3D poses directly from 2D image features, without intermediate 3D representations, depending entirely on the network's ability to learn the 2D-to-3D mapping, making them less robust to occlusions [7]. Coarse-to-fine methods refine an initial prediction of the 3D joints or shape, but the reliance on explicit 3D representations often fails to capture global context [9]. Unlike explicit representations, signed distance fields provide a global, continuous, and implicit representation of the 3D shapes across the entire 3D volume. Signed distance fields not only encode the shape of the hand and object, but also their interactions and collision points, providing a way to address occlusions and provide global constraints for the 3D pose.





Method

Signed Distance Field

A U-Net [10] model extracts the hand and object segmentation masks and 2D keypoint predictions from the image. The decoder part of the U-Net and a multi-layer perceptron (MLP) is used to convert a 3D point into an image feature vector. Fourier positional encoding [11] is also incorporated to expand the 3D point to a larger space. The signed distance decoder is a fully connected neural network that predicts the signed distance to the hand or object surface from the 3D point, image features, and positional encodings.

Feature Extraction

The 3D space is divided into a set of 3D bins, each containing a query point. These points are filtered in the 2D space using the hand and object bounding boxes. The filtered set of points are sorted based on their distances to the surface and only the closest points are kept. For each query point, its volume density $\sigma = \alpha^{-1}\text{sigmoid}(-d/\alpha)$, where α is a learned parameter, is calculated where a higher density is closer to the surface, giving more importance to those points. These densities are combined with the query point, positional encodings, and image features, to form the final hand feature vector.

To incorporate hand-object interactions, the object query points are passed through the hand signed distance field to calculate the hand-object distances. These distances, along with the object query points, positional encodings, and image features, are used to form the cross-field hand feature vector.

Since both the final and cross-field feature vectors mainly contain local information, they are passed through six Multi-Head Self-Attention (MHSA) [12] layers to incorporate global shape information and form the enhanced hand features. The same process is repeated with the object query points to generate enhanced object features.

Pose Regression

The enhanced hand features are passed through Cross-Attention layers with 17 learned hand pose queries to regress the 3D hand joints and shape [13]. Since the object is more rigid, the rotation and translation vectors, that define the orientation and position of the object, are directly regressed from the enhanced object features.

Metrics in [mm]	MJE	PAMJE	OCE	MCE	ADD-S	Object
Lin *et al.* [32]	15.2	6.99	-	-	-	No
Spurr *et al.* [44]	17.3	6.83	-	-	-	No
Liu *et al.* [34]	15.2	6.58	-	-	-	Yes
Park *et al.* [39]	14.0	5.80	-	-	-	No
Chen *et al.* [9]	14.2	6.40	-	-	-	No
Xu *et al.* [52]	14.0	5.70	-	-	-	No
Lin *et al.* [33]	12.6	5.47	42.7	48.0	33.8	Yes
HOISDF (ours)	**10.1**	**5.13**	**27.6**	**35.8**	**18.6**	Yes

This table compares HOISDF [1] against previous state-of-the-art (SOTA) models on the DexYCB [14] dataset. The two major metrics are Mean Joint Error (MJE) and Poscrustes-Alignment Mean Joint Error (PAMJE) [15], where HOISDF had the best accuracy at the time of publication. MJE computes the average euclidean distance between the ground truth and predicted 3D joint positions. PAMJE first translates, scales, and rotates the skeletons to align them and focus on relative positioning, rather than absolute alignment.

Applications

3D pose estimation models like HOISDF that estimate hand and object keypoints have applications in a variety of fields, such as augmented reality, robotics, and healthcare. AR applications can leverage the 3D hand and object coordinates for gesture-based controls or hand-object interactions [4]. For example, users with limited mobility could perform gestures to take actions in an application or game and pose estimation could be used to recognize these gestures. Pose estimation can aid in enabling robots to interpret human hand movements and object manipulations [3]. The robot can then use this information for interactions, such as shaking the person's hand or picking up the object. Especially in areas like physical therapy, hand and object pose estimation can be used to analyze patient movements [5]. The Box and Block Test (BBT) [16] is a common assessment in rehabilitation, where a participant will move small blocks from one compartment to another. HOISDF can record the hand pose and hand-object interactions, which can then be passed through another network to be scored.

Limitations

The paper explicitly states that the hand and object meshes might intersect with each other in severely occluded scenarios. A possible solution is to include a weighted loss value for mesh intersections in the SDF decoder.

$$\mathcal{L}_{\text{intersection}} = \frac{1}{N} \sum_{p \in P} \text{ReLU}(-\text{SDF}_{\text{hand}}(p)) \cdot \text{ReLU}(-\text{SDF}_{\text{object}}(p))$$

The two datasets this model was trained on, DexYCB [14] and HO3D [17], only include objects that are bigger than the hand, which are much less occluded than smaller objects. The model can be trained and optimized for smaller objects to make better predictions about occluded scenarios.

Future Work

Temporal Consistency for Videos

While the model is designed for single images, the accuracy can be increased for videos by leveraging temporal constraints. One possible solution is to just pass the final keypoint from the previous frame along with the query point and features to the SDF decoder. This restricts the model from predicting the keypoints in parallel as each frame is dependent on the previous one.

Another solution is to train a separate temporal transformer encoder module, such as TCPFormer [18], that will receive the keypoint predictions for the video once all the frames have been calculated, and then shifts the keypoints to ensure smooth and consistent transitions. TCPFormer uses an implicit pose proxy to represent pose information, and utilizes self-attention to capture relationships between keypoints across frames.

By incorporating temporal context, keypoints are smoother and more consistent across consecutive frames. Temporal context can also enhance the model's ability to handle occlusions, as information from when a keypoint was visible can help infer the location of the occluded keypoint.

Occupancy Fields

Occupancy fields [19] model 3D shapes by representing the probability that a given a 3D point is inside or outside an object. Unlike signed distance fields, which encode the distance to the surface, occupancy fields classify points based on their occupancy probability, which could handle occlusions better. This approach could replace the SDF decoder in the HOISDF architecture, as it still represents the shape geometry but avoids relying on the exact surface distance. The surface of the hand or object can still be inferred with query points from the occupancy field by determining the decision boundary, where the occupancy probability is at a specific threshold.

References

[1] Haozhe Qi, Chen Zhao, Mathieu Salzmann, and Alexander Mathis, "HOISDF: Constraining 3D Hand-Object Pose Estimation with Global Signed Distance Fields," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[2] Baorui Ma, Zhizhong Han, Yu-Shen Liu, and Matthias Zwicker, "Neural-Pull: Learning Signed Distance Functions from Point Clouds by Learning to Pull Space onto Surfaces," in *International Conference on Machine Learning (ICML)*, 2021.

[3] Aude Billard and Danica Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, eaat8414, 2019.

[4] Yunqiang Chen, Qing Wang, Hong Chen, Xiaoyu Song, Hui Tang, and Mengxiao Tian, "An overview of augmented reality technology," in *Journal of Physics: Conference Series*, IOP Publishing, 2019, p. 022 082.

[5] Jan Stenum, Kendra M. Cherry-Allen, Connor O. Pyles, Rachel D. Reetzke, Michael F. Vignos, and Ryan T. Roemmich, "Applications of Pose Estimation in Human Health and Performance across the Lifespan," *Sensors*, vol. 21, p. 7315, 2021.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep Residual Learning for Image Recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] Taeho Kang and Youngki Lee, "Attention-Propagation Network for Egocentric Heatmap to 3D Pose Lifting," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[8] Feng Zhou, Jianqin Yin, and Peiyang Li, "Lifting by Image – Leveraging Image Cues for Accurate 3D Human Pose Estimation," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2024.

[9] Wencan Cheng, Hao Tang, Luc Van Gool, and Jong Hwan Ko, "HandDiff: 3D Hand Pose Estimation with Diffusion on Image-Point Cloud," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[11] Matthew Tancik et al., "Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[12] Ashish Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[13] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit, "Keypoint Transformer: Solving Joint Identification in Challenging Hands and Object Interactions for Accurate 3D Pose Estimation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[14] Yu-Wei Chao et al., "DexYCB: A Benchmark for Capturing Hand Grasping of Objects," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[15] Christian Zimmermann and Thomas Brox, "Learning to Estimate 3D Hand Pose from Single RGB Images," in *International Conference on Computer Vision (ICCV)*, 2017, pp. 4903–4911.

[16] Virgil Mathiowetz, Gloria Volland, Nancy Kashman, and Kathleen Weber, "Adult Norms for the Box and Block Test of Manual Dexterity," *American Journal of Occupational Therapy*, vol. 39, pp. 386–391, 1985.

[17] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit, "HOnnotate: A method for 3D Annotation of Hand and Object Poses," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[18] Jiajie Liu, Mengyuan Liu, Hong Liu, and Wenhao Li, "TCPFormer: Learning Temporal Correlation with Implicit Pose Proxy for 3D Human Pose Estimation," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2025.

[19] Jiahao Li, Haoyan Zhang, Jiawei Zhou, Zhaoyang Fan, and Zhiqiang Zhang, "Diffusion-fof: Single-view clothed human reconstruction via diffusion-based fourier occupancy field," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.