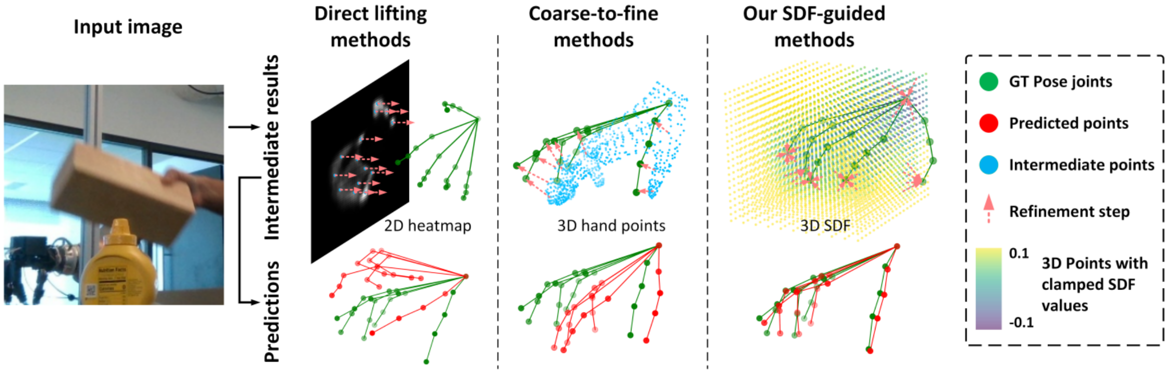
HOISDF: Constraining 3D Hand-Object Pose Estimation with Global Signed Distance Fields

Haozhe Qi, Chen Zhao, Mathieu Salzmann, Alexander Mathis

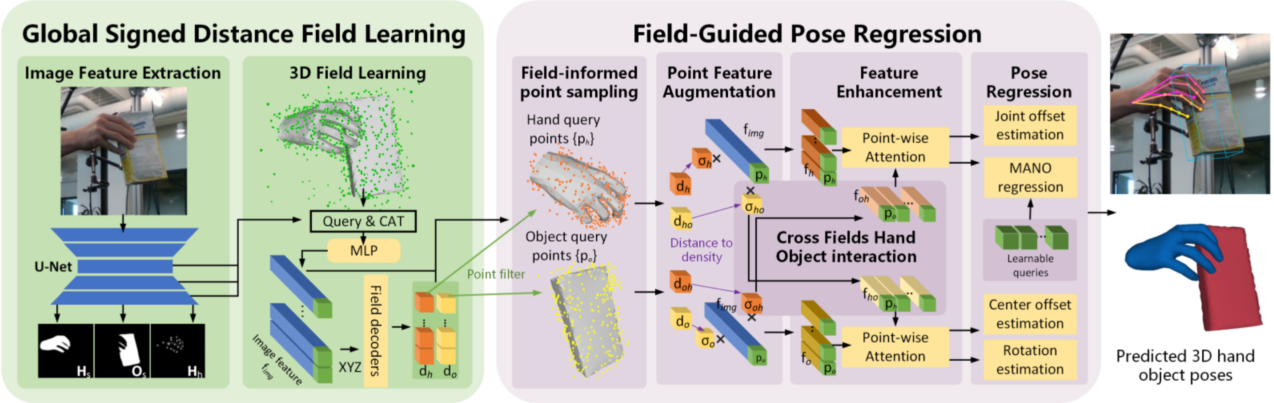
Accepted for CVPR 2024 (Paper) (GitHub)

The HOISDF model estimates signed distance fields for both the hand and object to build the shape of each with sampled points. The sample points on the surface are then used to predict hand joints and shape, as well as object rotation and translation.



Motivation

This model encodes the hand shape as a continuous function instead of a surface mesh with predefined vertices and edges. Also, since the pose keypoints are determined from the signed distance field, you have more freedom around the keypoint prediction by modifying the signed distance field decoder, such as joint or collision constraints.



Method

Global Signed Distance Field Learning

1. **Image Feature Extraction**

   A U-Net architecture extracts hand and object segmentation masks and 2D keypoint predictions from the image.

2. **3D Signed Distance Field Learning**

   The decoder part of the U-Net is then used to convert a 3D point into an image feature vector. Fourier positional encoding is also incorporated to expand the 3D point to a larger space. The 3D point, image features, and positional encodings are passed through a fully connected neural network to predict the signed distance to the hand or object surface with positive values for outside of the hand and negative for inside.

Field-Guided Pose Regression

1. **Field-informed Point Sampling**

   The 3D space is split into uniform bins, each with a query point, and then the points are filtered in the 2D space using the hand and object bounding boxes. The points are sorted using their distances to the surface and the first set of points are kept.

2. **Field-based Point Feature Augmentation**

   For each query point, its volume density is calculated where a higher density is closer to the surface, giving more importance to those points. This is concatenated with the features from before to form the final query point feature vector.

3. **Cross Fields Hand-Object Interaction**

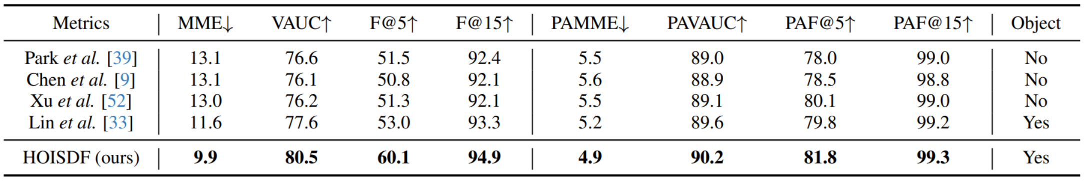
   The query points are passed through the other signed distance field so hand query points are passed through object signed distance field and vice versa. We then repeat what we did before but with these distances instead.

4. **Feature Enhancement with Point-wise Attention**

   The final feature and cross-field feature vectors are passed through six multi head self-attention layers to compute the enhanced point features.

5. **Point-wise Pose Regression**

   Cross attention layers are used with 17 learned hand pose queries to compute the 3D joint angles and the 10D mano shape parameters. For each joint, distance vectors are created from each nearby query point to the joint to fine-tune that joint's 3D position. Since the object is more rigid, the rotation and translation vectors can be directly regressed from the enhanced object features.



Metrics	MME↓	VAUC↑	F@5↑	F@15↑	PAMME↓	PAVAUC↑	PAF@5↑	PAF@15↑	Object
Park *et al.* [39]	13.1	76.6	51.5	92.4	5.5	89.0	78.0	99.0	No
Chen *et al.* [9]	13.1	76.1	50.8	92.1	5.6	88.9	78.5	98.8	No
Xu *et al.* [52]	13.0	76.2	51.3	92.1	5.5	89.1	80.1	99.0	No
Lin *et al.* [33]	11.6	77.6	53.0	93.3	5.2	89.6	79.8	99.2	Yes
HOISDF (ours)	**9.9**	**80.5**	**60.1**	**94.9**	**4.9**	**90.2**	**81.8**	**99.3**	Yes

Limitations

1. 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 with SDF learning.

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

2. The two datasets this model was trained on only include objects that are bigger than the hand, which are much less occluded than smaller objects.

Enhancements

1. Train a separate temporal transformer encoder module that will receive the keypoint predictions for the video once all of the frames have been calculated and then shifts the keypoints to ensure smooth and consistent transitions.

   - Leverages the Multi-Head Self Attention to capture global relationships

- Allows the frames to be still processed in parallel by not making each frame dependent on previous frames

2. Pass the previous keypoints along with the query point to the current SDF decoder.