# gSDF: Geometry-Driven Signed Distance Functions for 3D Hand-Object Reconstruction


Zerui Chen, Shizhe Chen, Cordelia Schmid, Ivan Laptev

INRIA – WILLOW / École Normale Supérieure, Paris, France


## Goal

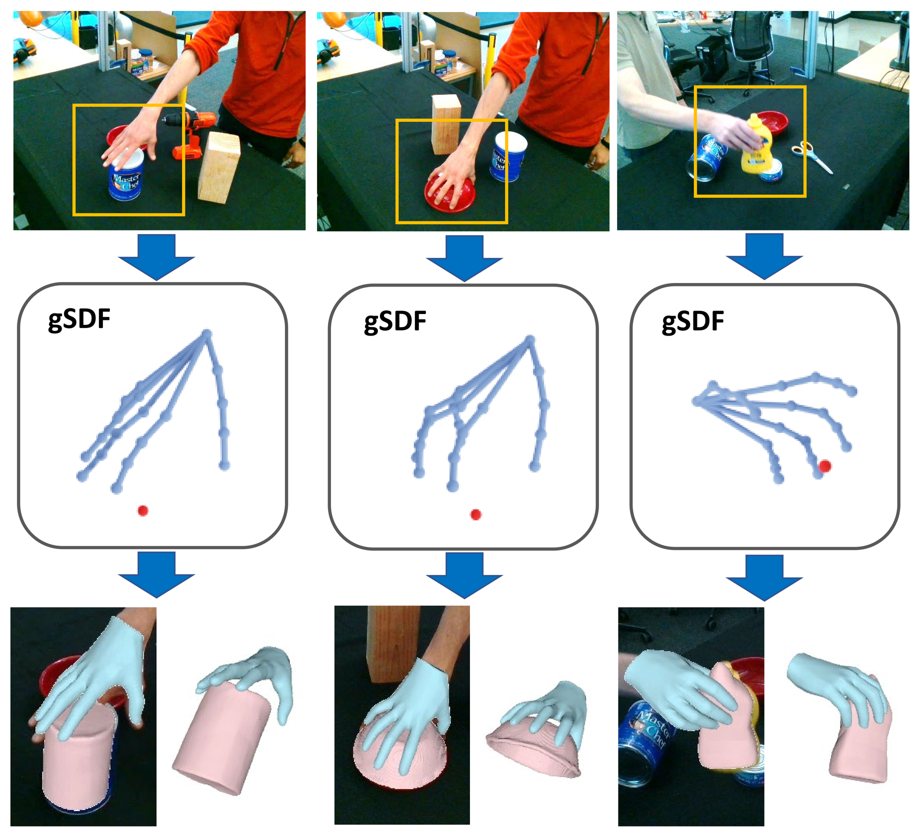
**Realistic** 3D hand-object reconstruction from **monocular images**.

## Contribution

- Leverage **hand pose estimation** as a guidance for SDF-based 3D hand-object reconstruction.
- Use monocular **videos** to alleviate **occlusion and motion blur** issues and improve the performance.

## Motivation

- Deep SDFs can generalize to different shape resolutions but lack explicit modeling of the underlying 3D geometry.
- 3D hand-object reconstruction from a single RGB image is intrinsically hard, especially under occlusion or motion blur.




## Related work


[1] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. Black, I. Laptev, and C. Schmid. Learning joint reconstruction of hands and manipulated objects. *In Proc. CVPR, 2019.*

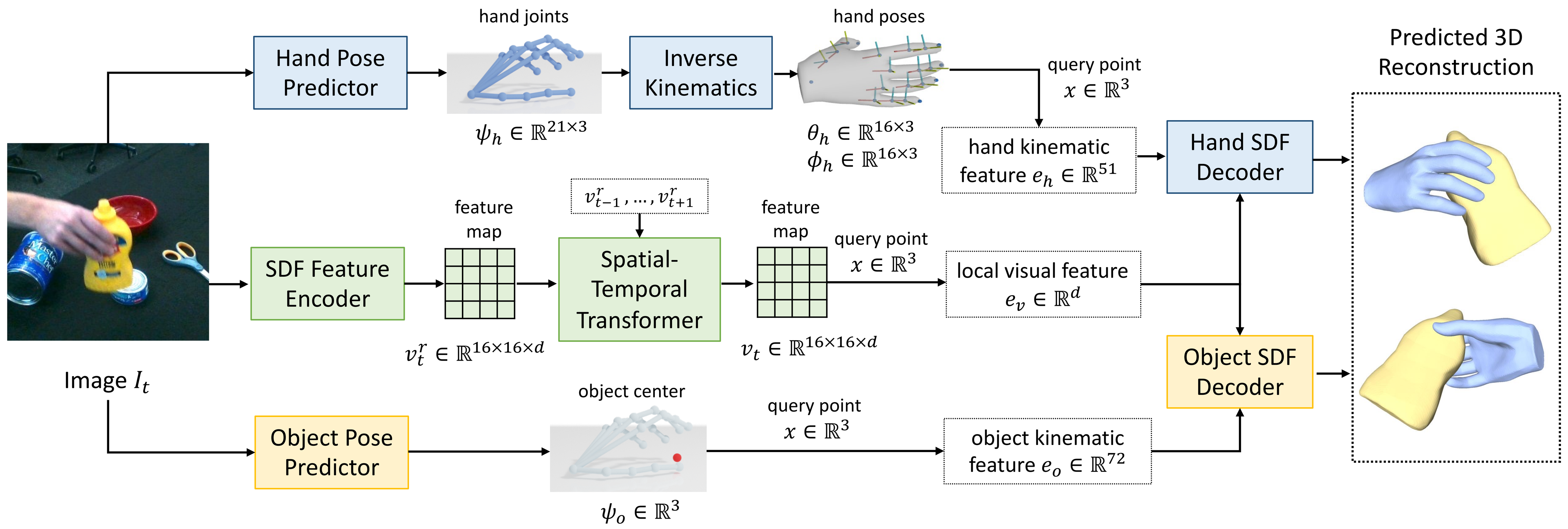
[2] K. Karunratanakul, J. Yang, Y. Zhang, M. Black, K. Muandet, and Siyu Tang. Grasping Field: Learning Implicit Representations for Human Grasps, *In Proc. 3DV, 2020.*

[3] Y. Ye, A. Gupta, and S. Tulsiani. What's in your hands? 3D Reconstruction of Generic Objects in Hands. *In Proc. CVPR, 2022.*

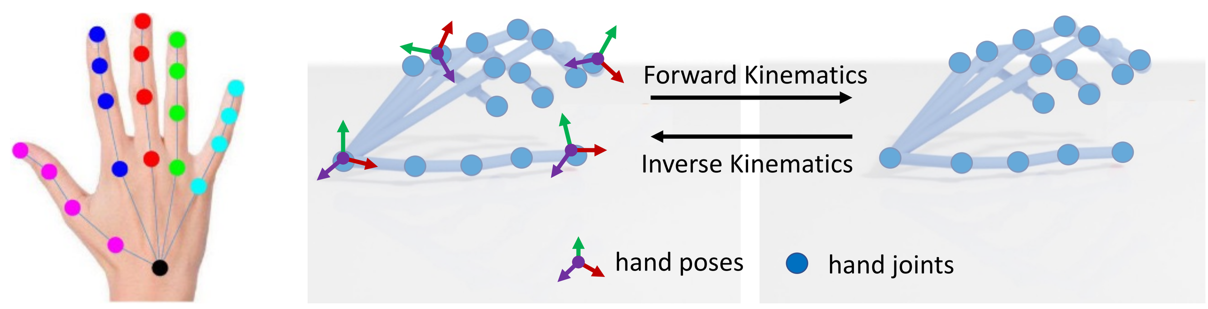
[4] Z. Chen, Y. Hasson, C. Schmid, I. Laptev. AlignSDF: Pose-Aligned Signed Distance Fields for Hand-Object Reconstruction. *In Proc. ECCV, 2022.*


## Approach

We formulate the joint hand-object 3D reconstruction task as a multi-task learning framework.



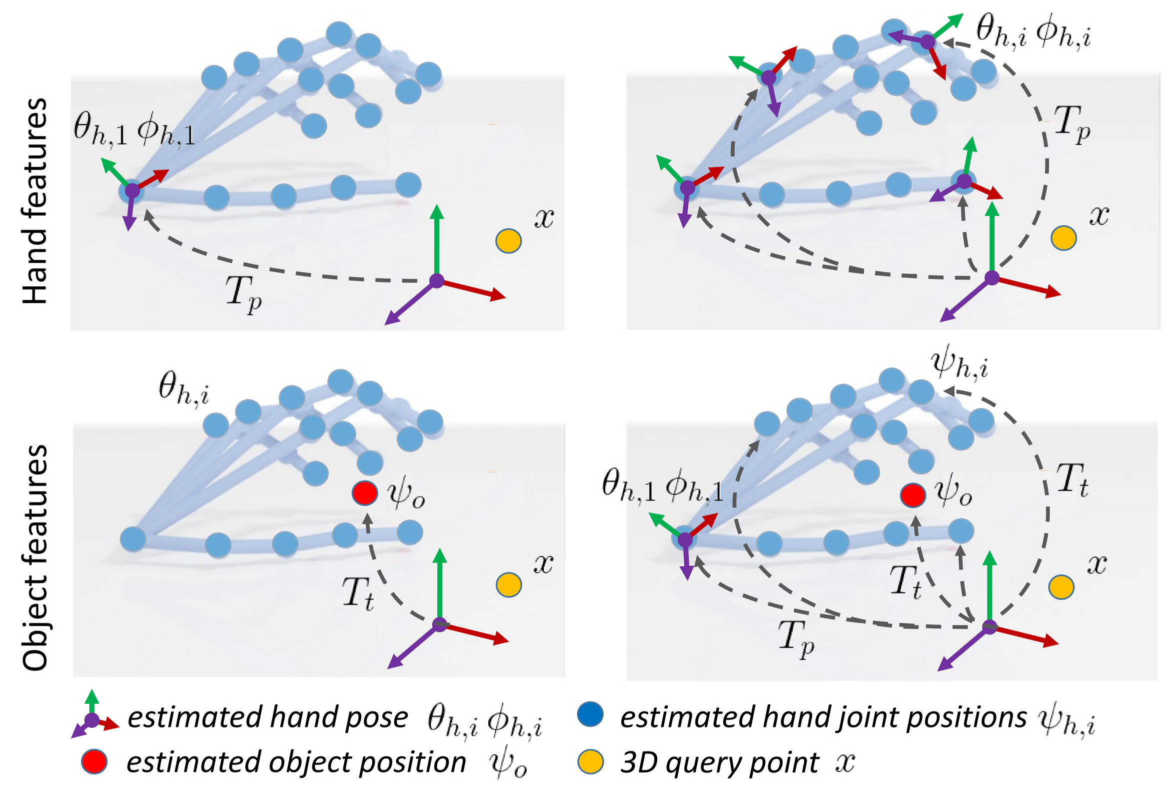

### Pose Estimation




(a) Hand Joints (b) Hand Kinematics

- As shown in (a), since deep CNNs is good at detecting interested points, we first use neural networks to predict 3D hand joint locations from single-view images.
- As shown in (b), we use inverse kinematics to recover the pose transformations for each hand bone.

### Kinematic Features

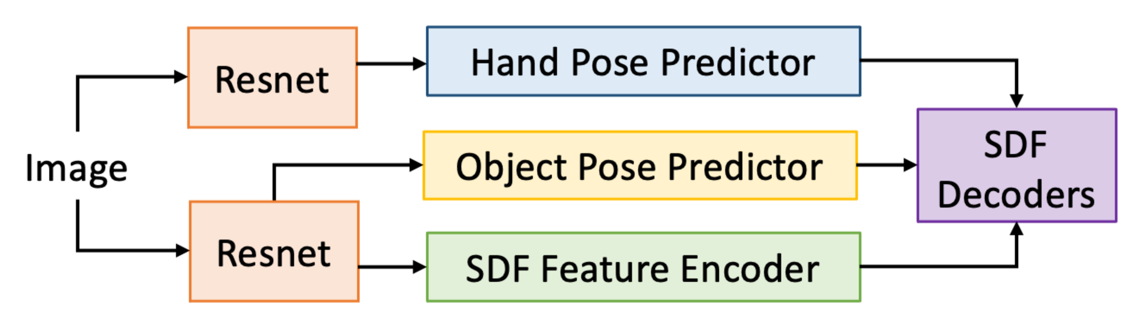



- For hand kinematic features, compared with a recent work [4], we use full kinematic chains of pose transformations.
- For object kinematic features, instead of only considering the object translation as in [4], we additionally consider relative positions between the query point *x* and each hand joint.

### Visual Feature

- We use the spatial and temporary transformer to aggregate features from multiple frames.
- Then, we project the query point *x* onto the plane of the feature map and obtain the refined local feature for the shape reconstruction.

### Network Architecture




- Our model consists of hand pose predictor, object pose predictor, SDF feature encoder and SDF decoders.
- We use two backbones to handle the task of 3D shape reconstructions and the task of pose predictions separately.
- We observe that our model can achieve the best performance when the object pose predictor and SDF feature encoder shares the same backbone.

## Results

We validate the method by conducting experiments on ObMan and DexYCB benchmarks. We employ metrics including Chamfer Distance (CD) and F-score (FS) to evaluate the quality of results.

### Quantitative Comparison on ObMan

| Methods | $CD_h \downarrow$ | $FS_h@1 \uparrow$ | $FS_h@5 \uparrow$ | $CD_o \downarrow$ | $FS_o@5 \uparrow$ | $FS_o@10 \uparrow$ | $E_h \downarrow$ | $E_o \downarrow$ |
|---|---|---|---|---|---|---|---|---|
| Hasson *et al.* [1] | 0.415 | 0.138 | 0.751 | 3.60 | 0.359 | 0.590 | 1.13 | - |
| Karunratanakul *et al.* [2] | 0.261 | - | - | 6.80 | - | - | - | - |
| Ye *et al.* [3] | - | - | - | - | 0.420 | 0.630 | - | - |
| Chen *et al.* [4] | 0.136 | 0.302 | 0.913 | 3.38 | 0.404 | 0.636 | 1.27 | **3.29** |
| gSDF (Ours) | **0.112** | **0.332** | **0.935** | **3.14** | **0.438** | **0.660** | **0.93** | 3.43 |

### Quantitative Comparison on DexYCB

| Methods | $CD_h \downarrow$ | $FS_h@1 \uparrow$ | $FS_h@5 \uparrow$ | $CD_o \downarrow$ | $FS_o@5 \uparrow$ | $FS_o@10 \uparrow$ | $E_h \downarrow$ | $E_o \downarrow$ |
|---|---|---|---|---|---|---|---|---|
| Hasson *et al.* [1] | 0.537 | 0.115 | 0.647 | 1.94 | 0.383 | 0.642 | 1.67 | - |
| Karunratanakul *et al.* [2] | 0.364 | 0.154 | 0.764 | 2.06 | 0.392 | 0.660 | - | - |
| Chen *et al.* [4] | 0.358 | 0.162 | 0.767 | 1.83 | 0.410 | 0.679 | 1.58 | **1.78** |
| Chen *et al.* [4] [1†] | 0.344 | 0.167 | 0.776 | 1.81 | 0.413 | 0.687 | 1.57 | 1.93 |
| gSDF (Ours) | **0.302** | **0.177** | **0.801** | **1.55** | **0.437** | **0.709** | **1.44** | 1.96 |

### Qualitative Results