JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Neural Voting Field for Camera-Space 3D Hand Pose Estimation

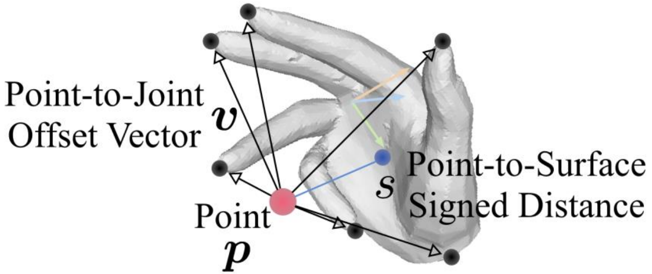
Lin Huang, Chung-Ching Lin, Kevin Lin, Lin Liang,
Lijuan Wang, Junsong Yuan, Zicheng Liu

Project website: https://linhuang17.github.io/NVF/ Poster: WED-AM-071

University at Buffalo
The State University of New York

Microsoft

Overview



Task:

- Absolute 3D hand pose estimation given a single RGB image

Assumption:

1. Camera intrinsic parameters are known
2. Optional:
    - Hand scale if provided

Existing Methods:

- First adopt holistic or pixel-level dense regression to obtain relative 3D hand pose and then follow with complex second-stage operations for 3D global root or scale recovery

Contributions:

1. Building on the recent progress in implicit representation learning, we propose Neural Voting Field (NVF), as the first 3D implicit representation-based unified solution to estimate camera-space 3D hand pose

2. NVF follows a novel unified 3D dense regression scheme to estimate camera-space 3D hand pose via dense 3D point-wise voting in camera frustum

3. NVF outperforms baseline methods based on holistic and 2D dense regression and achieves state-of-the-art results on absolute and relative hand pose estimation

Background: Challenges & Significance

General Task:
Monocular 3D hand pose estimation generally aims to recover 3D locations of hand joints from an RGB image

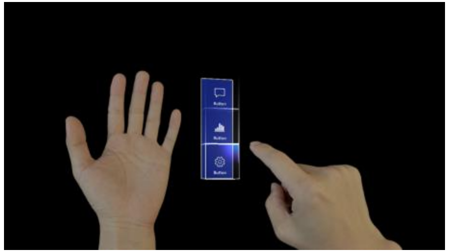
Common Challenges:
- Highly articulated structure
- Large variations in global orientations
- Severe (self-)occlusion

Challenges for RGB Input:
- 2D-to-3D depth and scale ambiguities
- Self-similarity and uniform hand texture
- Cluttered Background
- Lighting

Significance:
- Most existing works focused on root-relative 3D hand pose estimation. Having root-relative hand joint coordinates alone is insufficient for various interactive tasks.

- Being able to recover camera-space 3D hand joint coordinates in an AR view enables the user to directly use hands to manipulate virtual objects moving in 3D space.


Hand-Object Interaction


Mixed Reality [Microsoft]

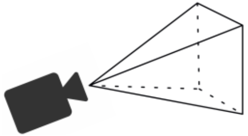

Teleoperation [Internet]

Considered Task: Monocular Absolute 3D Hand Pose Estimation

<u>Input</u>

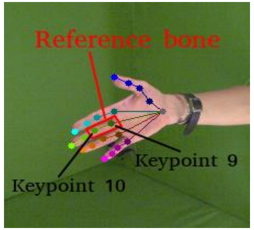
Single Hand RGB Frame

<u>Required prior:</u>



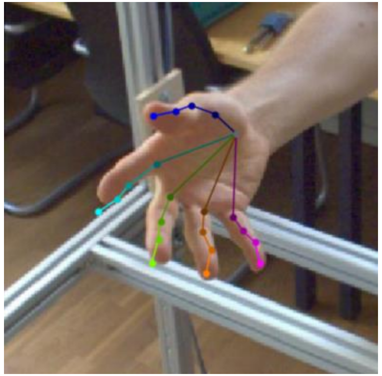
Camera Intrinsics

<u>Optional prior:</u>



Hand Scale

<u>Output</u>



3D Hand Pose

The 3D hand pose is defined as:
<u>hand joint locations in camera space</u>

Related Works: Monocular Absolute 3D Hand Pose Estimation

Comparison of representative absolute 3D hand pose estimation schemes

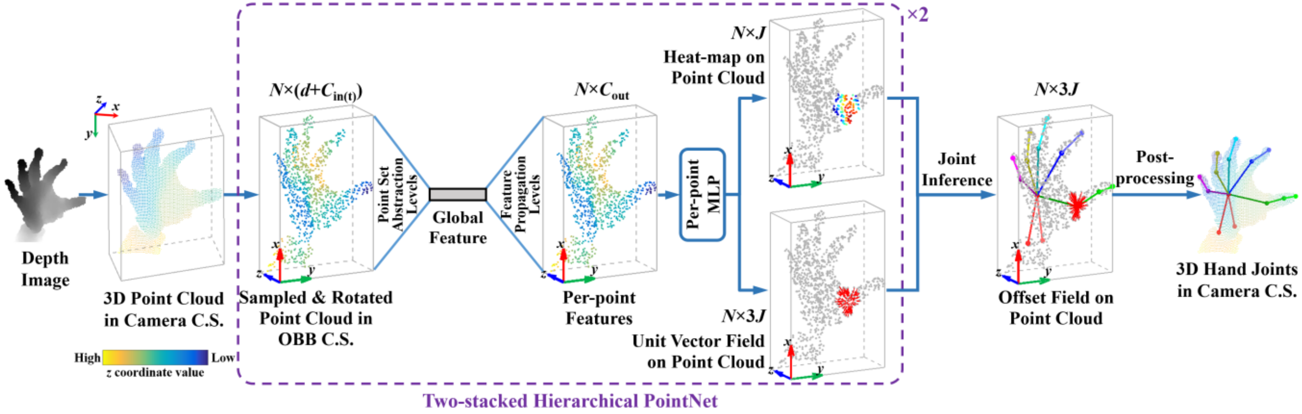
Method	First Stage	Second Stage
Iqbal *et al*. [29]	2D-Dense	Scale Estimation
ObMan [22]	Holistic	Root Estimation
I2L-MeshNet [42]	1D-Dense	Root Depth Estimation
CMR [8]	2D-Dense+SpiralConv	Registration
Hasson *et al*. [21]	Holistic	Model Fitting
NVF (Ours)	Unified 3D-Dense	Weighted Average

Motivation

Our Goal:
A unified framework for robust camera-space 3D hand pose estimation from an RGB image

Two Key Design Elements:
1) The ability to exploit dense local evidence:

   Dense regression-based methods are more effective than holistic regression-based counterparts for handling highly articulated 3D pose structure



[Point-to-Point'18] regresses dense 3D point-wise estimations directly from input 3D point cloud, showing superior performance improvements over holistic regression-based methods for 3D hand pose estimation

Motivation

Our Goal:

A unified framework for robust camera-space 3D hand pose estimation from an RGB image

Two Key Design Elements:

1) The ability to exploit dense local evidence:

   Dense regression-based methods are more effective than holistic regression-based counterparts for handling highly articulated 3D pose structure

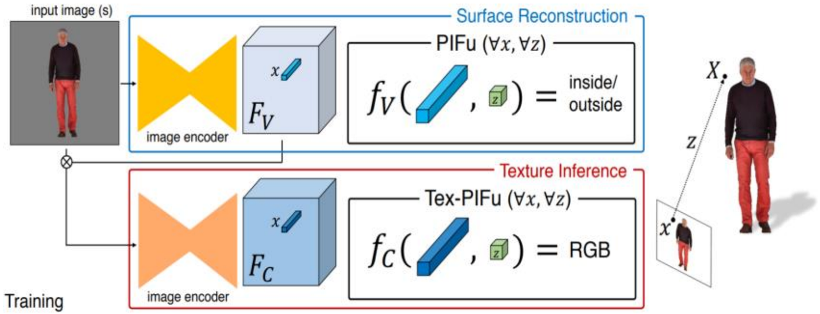
2) The ability to reason 3D hand global geometry:

   Given 2D evidence and camera intrinsic parameters, reasonable understanding towards target object 3D structure/geometry is crucial to alleviate 2D-to-3D depth ambiguity

Motivation

Our Goal:
A unified framework for robust camera-space 3D hand pose estimation from an RGB image

The Question:
How to fully integrate both elements into our algorithm design in a unified manner?



[PIFu’19]-based methods reconstruct highly detailed 3D human geometry from an RGB image in a unified way, showing its ability to model high frequency local details (e.g., clothing wrinkles) while generating complete global geometry including largely occluded region (e.g., back of a person)

The Proposed Method: Dense Offset-based Pose Re-Parameterization

Pose Re-Parameterization:

$$\psi : \mathbb{R}^3 \times \mathcal{J} \times \mathcal{M} \mapsto \mathbb{R}^{T\times 4} \text{ as } \psi(\boldsymbol{p}, J, M) = V.$$

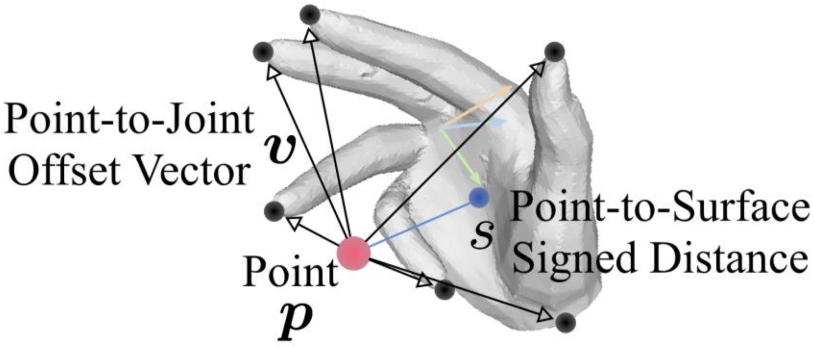
$$V = \{\boldsymbol{v}_t\}_{t=1}^{T}, \boldsymbol{v}_t \in \mathbb{R}^4$$

4D Offset Vector:

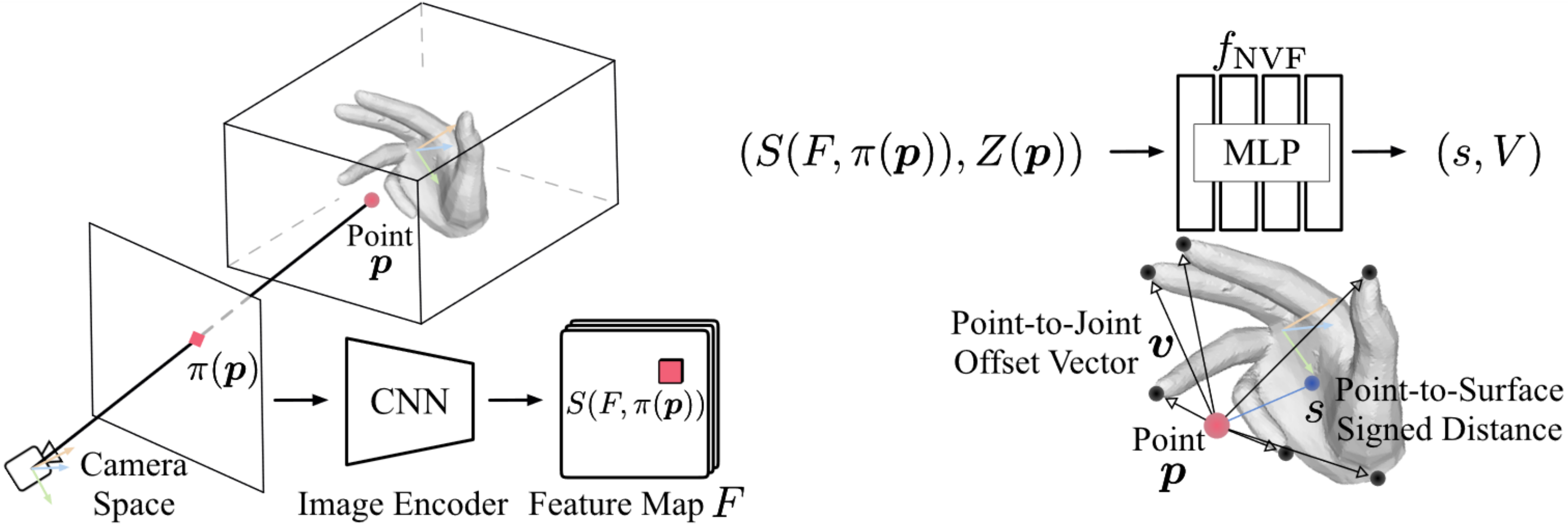
$$\boldsymbol{v}_t = (w_t, \boldsymbol{d}_t)$$

$$w_t = \begin{cases} 1 - \frac{\|\boldsymbol{j}_t - \boldsymbol{p}\|_2}{r} & |s| < \delta \text{ and } \|\boldsymbol{j}_t - \boldsymbol{p}\|_2 \le r \text{ and } \boldsymbol{p} \in B_t^K, \\ 0 & \text{otherwise;} \end{cases}$$

$$\boldsymbol{d}_t = \begin{cases} \frac{\boldsymbol{j}_t - \boldsymbol{p}}{\|\boldsymbol{j}_t - \boldsymbol{p}\|_2} & |s| < \delta \text{ and } \|\boldsymbol{j}_t - \boldsymbol{p}\|_2 \le r \text{ and } \boldsymbol{p} \in B_t^K, \\ \boldsymbol{0} & \text{otherwise;} \end{cases}$$



The Proposed Method: Neural Voting Field



Neural Voting Field (NVF):

$$f_{\text{NVF}} : \mathbb{R}^{C} \times \mathbb{R} \mapsto \mathbb{R} \times \mathbb{R}^{T \times 4} \quad \text{as}$$

$$f_{\text{NVF}}\big(S(F, \pi(\boldsymbol{p})), Z(\boldsymbol{p}); \boldsymbol{\theta}\big) = \big(s, V\big),$$

Optimization:

$$L_s = \frac{1}{N} \sum_{n=1}^{N} \left| \text{clamp}\left(\hat{s}_n, \delta\right) - \text{clamp}\left(s_n, \delta\right) \right|,$$

$$L_V = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\left(|\hat{s}_n| < \delta\right) H\left(\hat{V}_n, V_n\right),$$

$$\boldsymbol{\eta}^{\star}, \boldsymbol{\theta}^{\star} = \underset{\boldsymbol{\eta}, \boldsymbol{\theta}}{\arg\min}\, L_s + \lambda L_V .$$

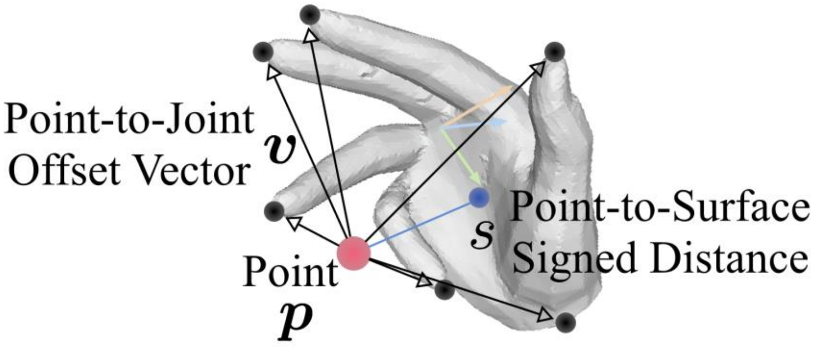
The Proposed Method: Dense 3D Point-to-Joint Voting

<u>4D Offset Vector to Actual 3D Offset:</u>

$$\boldsymbol{o}_t^n = \mathbb{1}\left(|s_n| < \delta\right)\left[r(1 - w_t^n)\boldsymbol{d}_t^n\right].$$

<u>Point-to-Joint Voting:</u>

$$\boldsymbol{j}_t = \sum_{n=1}^{N} \frac{\mathbb{1}\left(|s_n| < \delta\right) w_t^n (\boldsymbol{o}_t^n + \boldsymbol{p}_n)}{\sum_{n=1}^{N} \mathbb{1}\left(|s_n| < \delta\right) w_t^n}.$$



Experimental Setup: Baseline Methods

Sharing the same architecture of the Hourglass network and the MLP as NVF:

1) <u>Baseline-Holistic:</u>

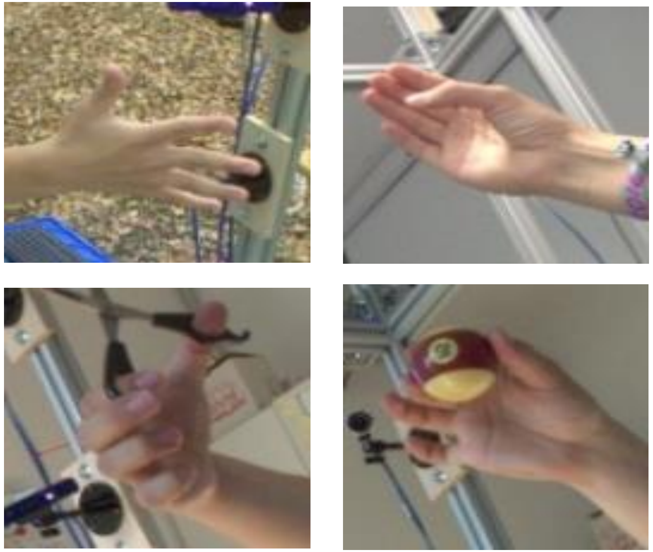
   We directly apply a global average pooling to the feature map extracted by the Hourglass network and use MLP to directly output the 3D hand pose

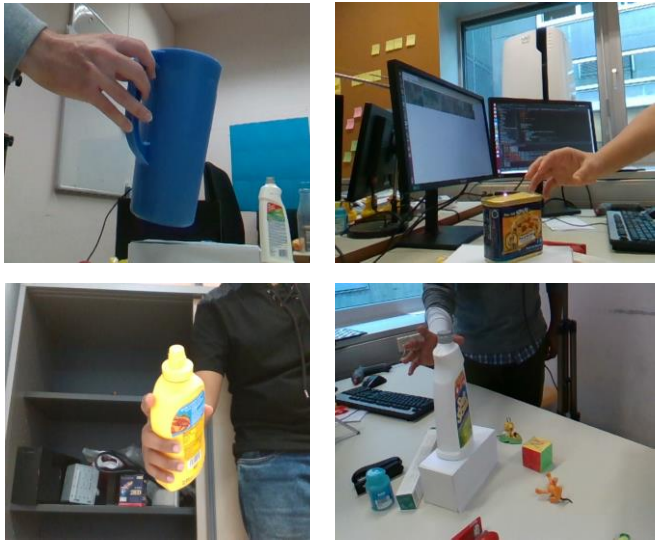
2) <u>Baseline-2D-Dense:</u>

   Given the feature map extracted by the Hourglass network, it uses MLP to predict for each pixel-aligned image feature:

   - Probability that the hand is present at each pixel
   - A set of 4D vectors (Each 4D vector consists of a 1D voting weight and 3D hand joint coordinate)

Experimental Setup: Datasets and Evaluation Metrics





FreiHAND:

- Camera-space 3D hand pose estimation

HO3D:

- Root-relative 3D hand pose estimation

Baseline Studies

Comparison with Baselines of CS-MJE for absolute 3D hand pose on FreiHAND

Method	Extra Data	Hand Crop	Hand Scale	CS-MJE↓
Baseline-Holisitc	-	✗	✗	54.5
Baseline-2D-Dense	-	✗	✗	53.2
CS-NVF (Ours)	-	✗	✗	**47.2**
Baseline-Holisitc	-	✗	✓	50.4
Baseline-2D-Dense	-	✗	✓	49.0
CS-NVF (Ours)	-	✗	✓	**42.4**
Baseline-Holisitc	Comp*	✗	✗	51.3
Baseline-2D-Dense	Comp*	✗	✗	50.9
CS-NVF (Ours)	Comp*	✗	✗	**44.6**
Baseline-Holisitc	Comp*	✗	✓	44.3
Baseline-2D-Dense	Comp*	✗	✓	43.4
CS-NVF (Ours)	Comp*	✗	✓	**39.3**

Comparison with Baselines of TE and DE for absolute 3D hand pose on FreiHAND

Method	Extra Data	Hand Scale	TE↓	DE↓
Baseline-Holisitc	-	✗	50.6	49.1
Baseline-2D-Dense	-	✗	49.2	47.9
CS-NVF (Ours)	-	✗	**43.6**	**42.4**
Baseline-Holisitc	-	✓	46.9	45.5
Baseline-2D-Dense	-	✓	45.3	43.9
CS-NVF (Ours)	-	✓	**38.9**	**37.8**
Baseline-Holisitc	Comp*	✗	48.7	47.1
Baseline-2D-Dense	Comp*	✗	47.9	46.4
CS-NVF (Ours)	Comp*	✗	**41.5**	**40.4**
Baseline-Holisitc	Comp*	✓	41.7	40.1
Baseline-2D-Dense	Comp*	✓	40.5	38.8
CS-NVF (Ours)	Comp*	✓	**36.5**	**35.5**

Main results

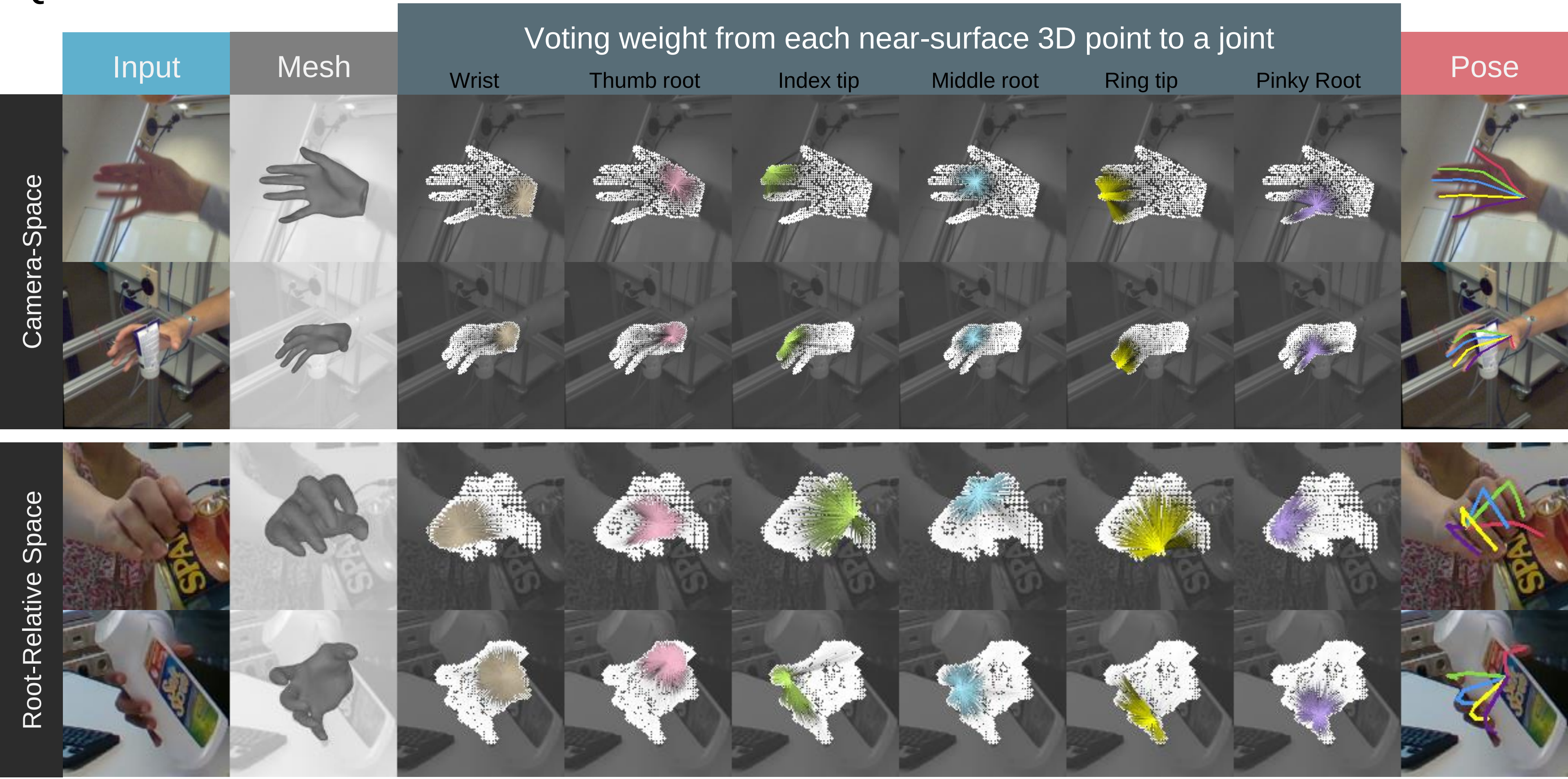
Comparison with SOTA methods for absolute 3D hand pose on FreiHAND

Method	Extra Data	Hand Crop	Hand Scale	CS-MJE↓
ObMan [22]	-	✓	✗	85.2
MANO CNN [64]	-	✓	✗	71.3
I2L-MeshNet [42]	-	✓	✗	60.3
CMR-SG-RN18 [8]	-	✓	✗	49.7
CMR-SG-RN50 [8]	-	✓	✗	48.8
Baseline-Holisitc	-	✗	✗	54.5
Baseline-2D-Dense	-	✗	✗	53.2
CS-NVF (Ours)	-	✗	✗	**47.2**
Baseline-Holisitc	-	✗	✓	50.4
Baseline-2D-Dense	-	✗	✓	49.0
CS-NVF (Ours)	-	✗	✓	**42.4**
CS-NVF (Ours)	Comp*	✗	✗	**44.6**
CS-NVF (Ours)	Comp*	✗	✓	**39.3**

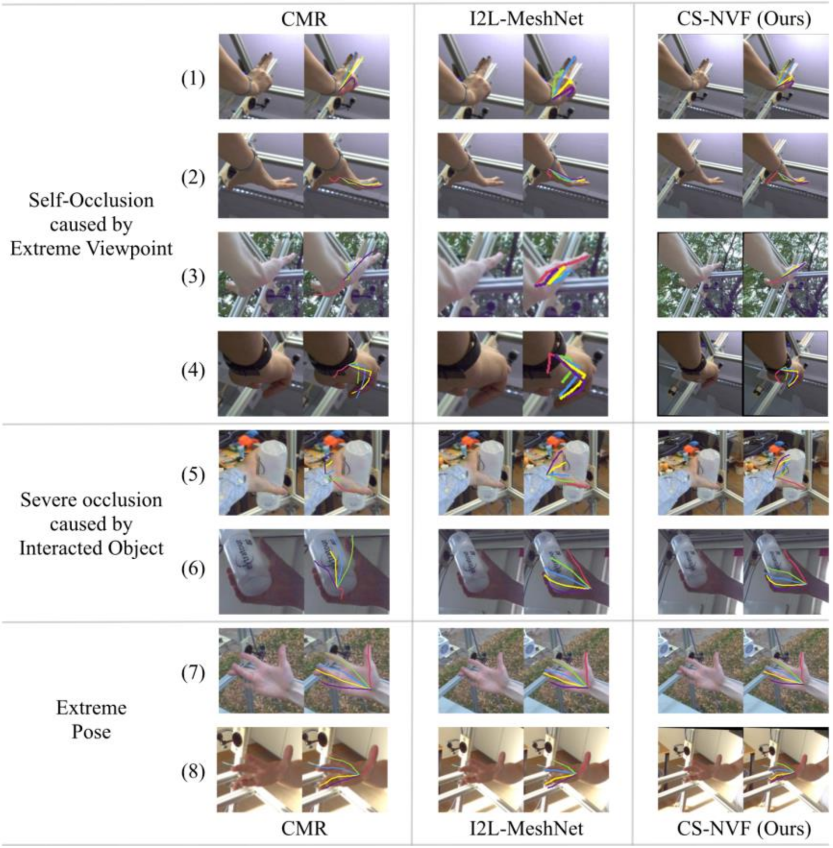
Comparison with SOTA methods for relative 3D hand pose on HO3D

Method	MJE↓	AUC↑	RS-MJE↓
Hasson et al. [20]	36.9	0.369	-
Pose2Mesh [11]	33.3	0.480	33.2
ObMan [22]	31.8	0.461	55.2
Liu et al. [38]	31.7	0.463	30.0
Hampali et al. [18]	30.4	0.494	-
METRO [35]	28.9	0.504	-
I2L-MeshNet [42]	26.0	0.529	26.8
Keypoint Trans. [19]	25.7	0.553	-
ArtiBoost [58]	25.3	0.532	-
Zheng et al. [61]	25.1	0.541	-
HandOccNet [45]	24.0	0.557	24.9
RS-NVF (Ours)	**21.8**	**0.610**	**23.2**

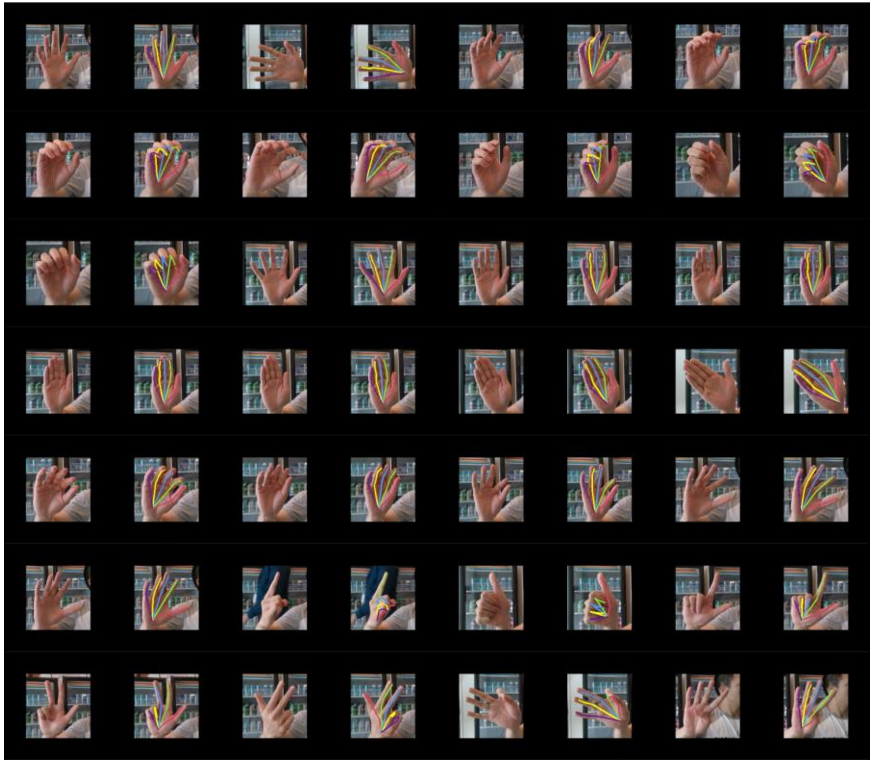
Qualitative Results



Additional Qualitative Results



Qualitative comparisons with SOTA methods on complex and failure cases



Qualitative results from another domain using NVF trained on FreiHAND only

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Neural Voting Field for Camera-Space 3D Hand Pose Estimation

Lin Huang, Chung-Ching Lin, Kevin Lin, Lin Liang,
Lijuan Wang, Junsong Yuan, Zicheng Liu

Project website: https://linhuang17.github.io/NVF/   Poster: WED-AM-071

University at Buffalo
The State University of New York

Microsoft