HAND SENSING FOR AUGMENTED INTERACTION

Junsong Yuan

University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences

Outline

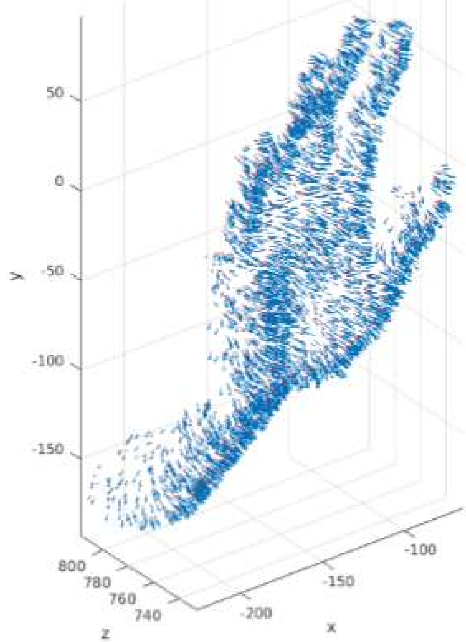
- Hand pose estimation and augmented interaction via depth cameras
- Hand pose estimation and augmented interaction via RGB cameras

Recent progress

The recent several years have witnessed a surging market of depth cameras and wearable devices.

Hand sensing from a depth camera

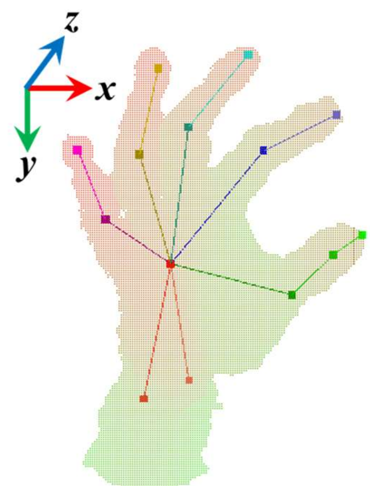
2.5-D data



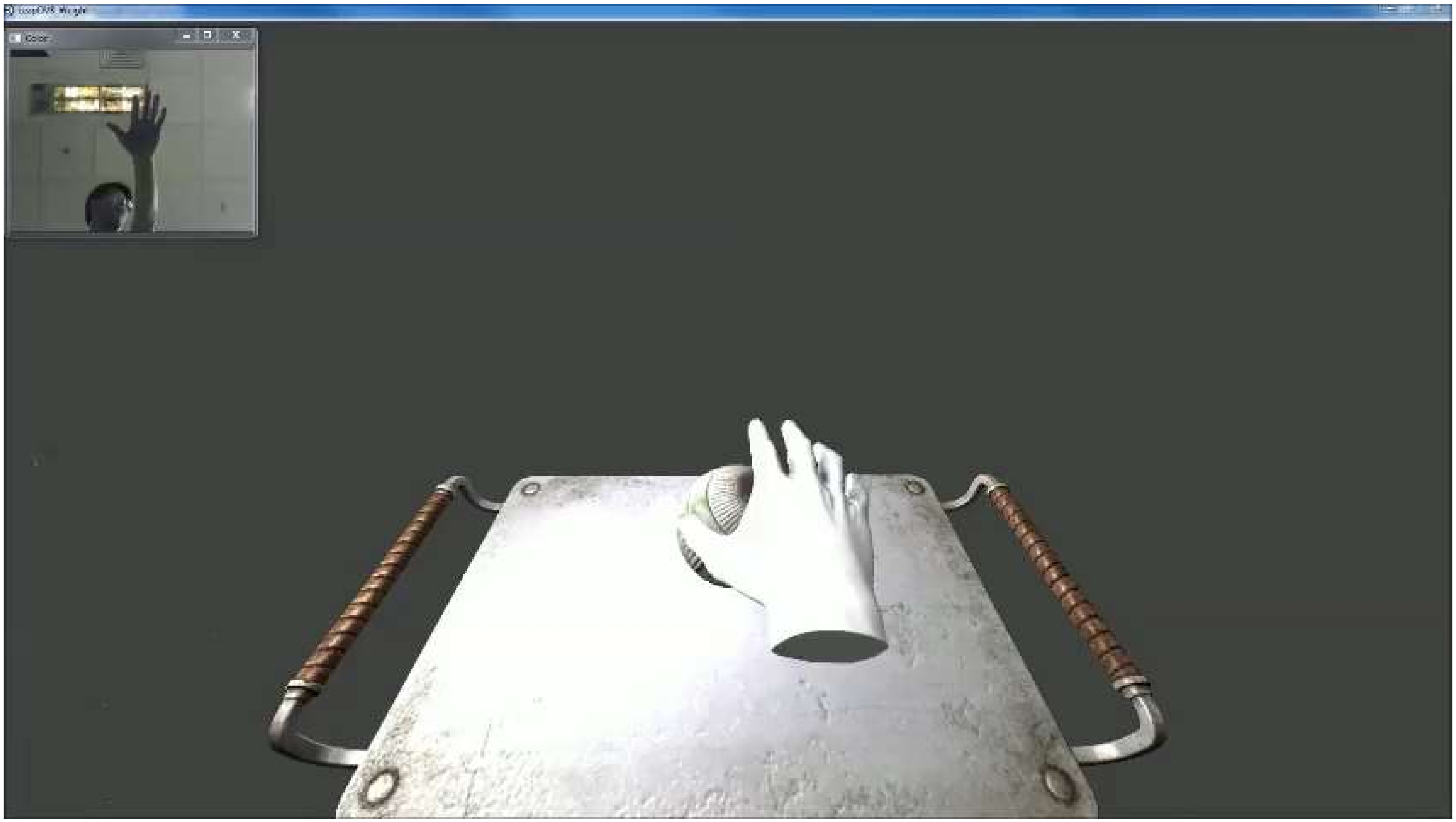
Hand MoCap: Problem Description

Input: a **depth image** containing a human hand

Output: estimated **3D** hand joint locations (in total 21 joints) which represent the hand pose



Virtual Reality

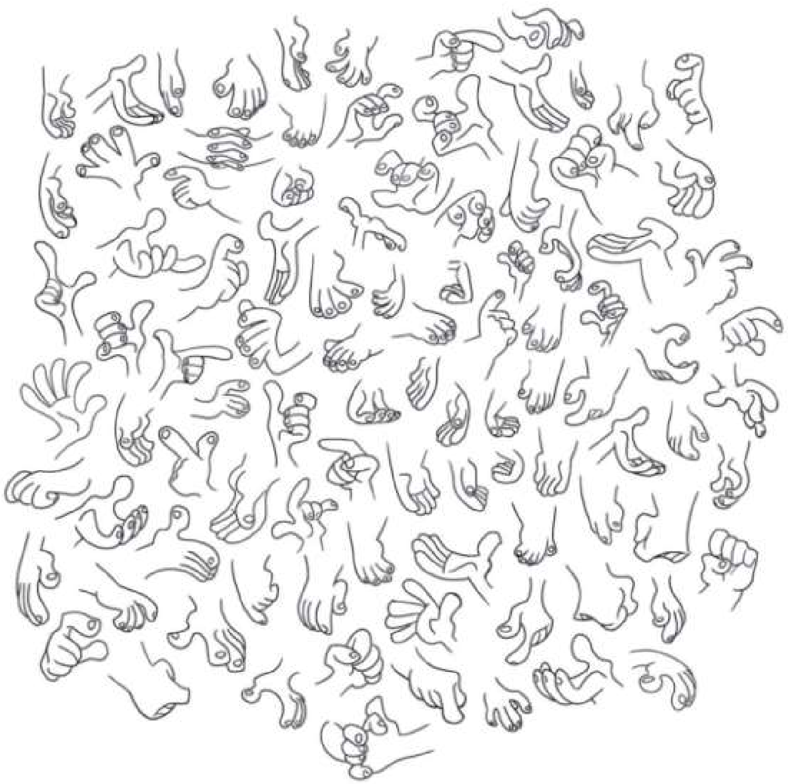


Augmented Reality

Hui Liang, Junsong Yuan, Daniel Thalmann,
IEEE Trans. on Cybernetics 2019

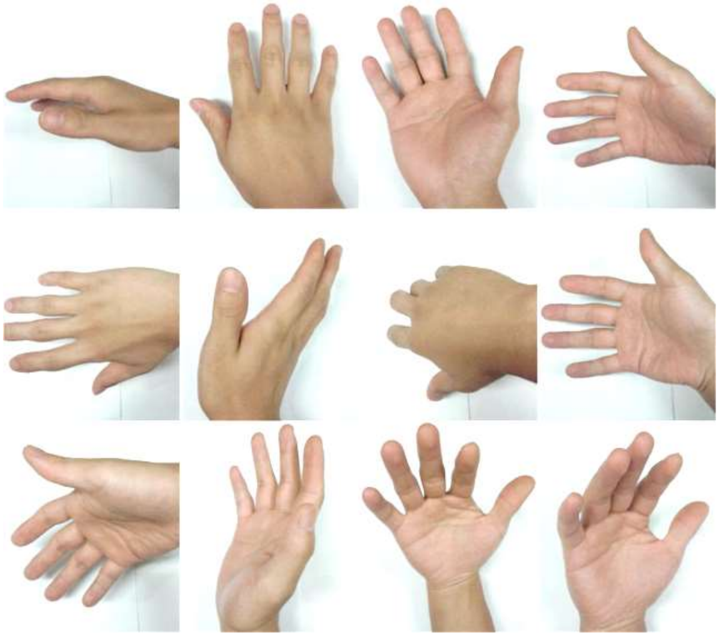
Challenges

- Hand pose has high degree of freedom



Challenges

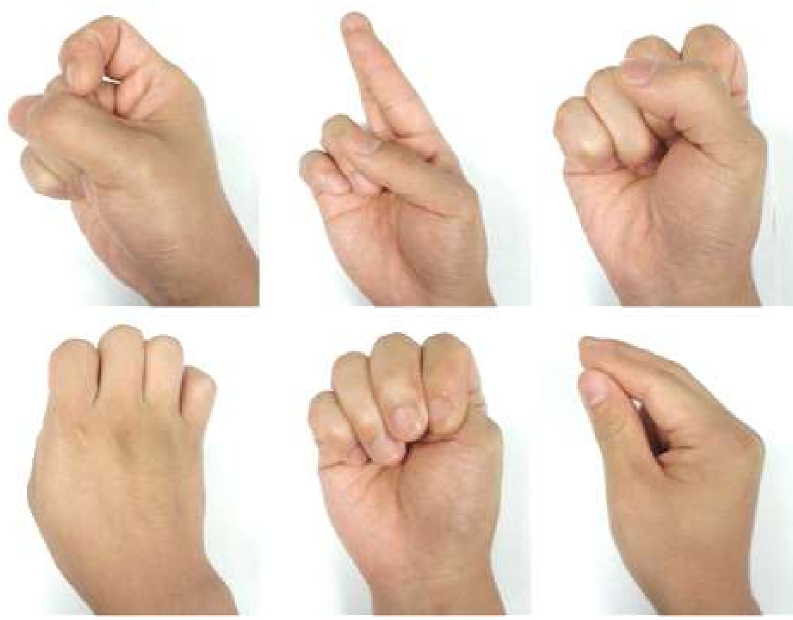
- View-point variations



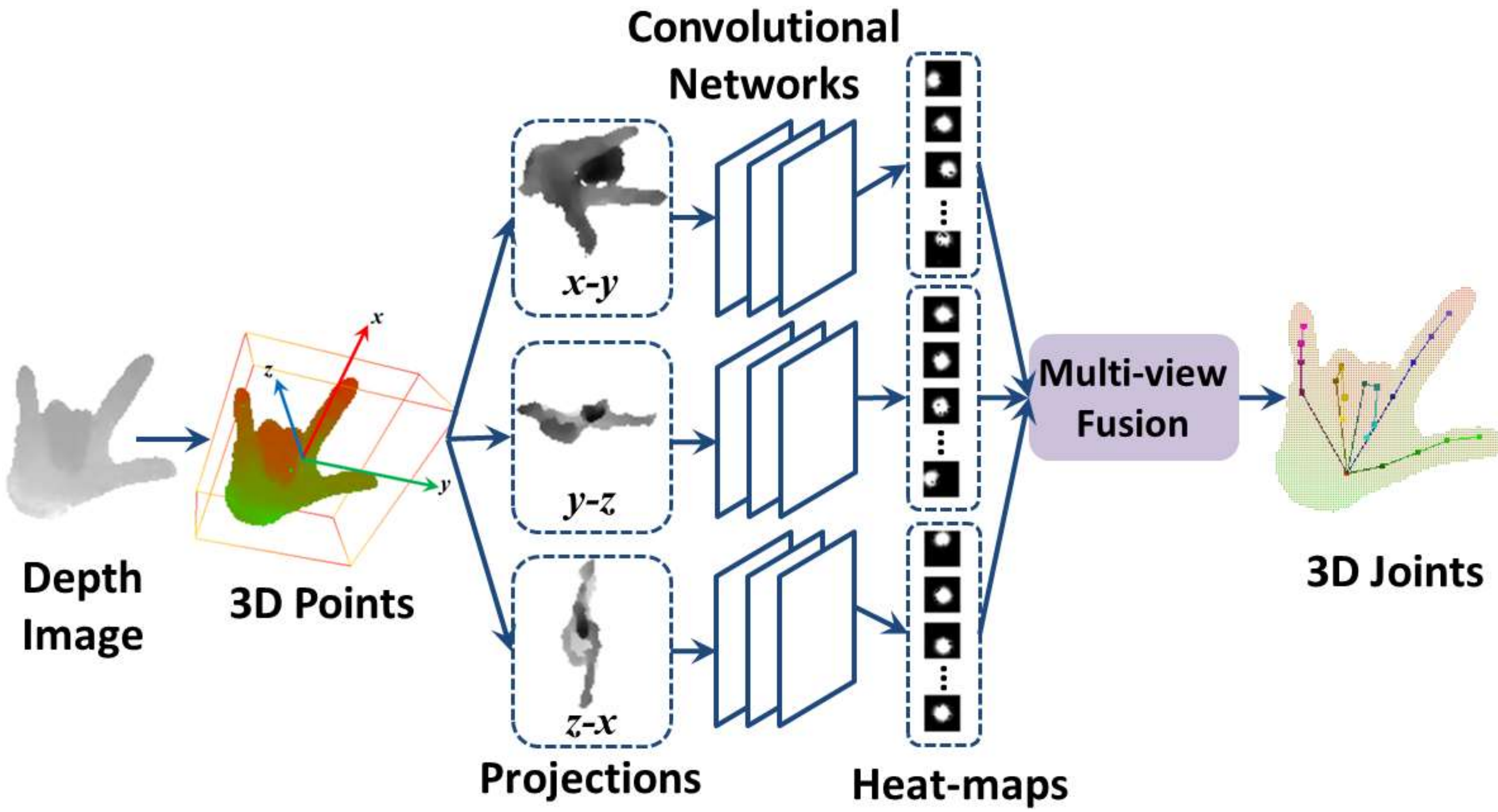
Same gesture from different view

Challenges

- Self-occlusions

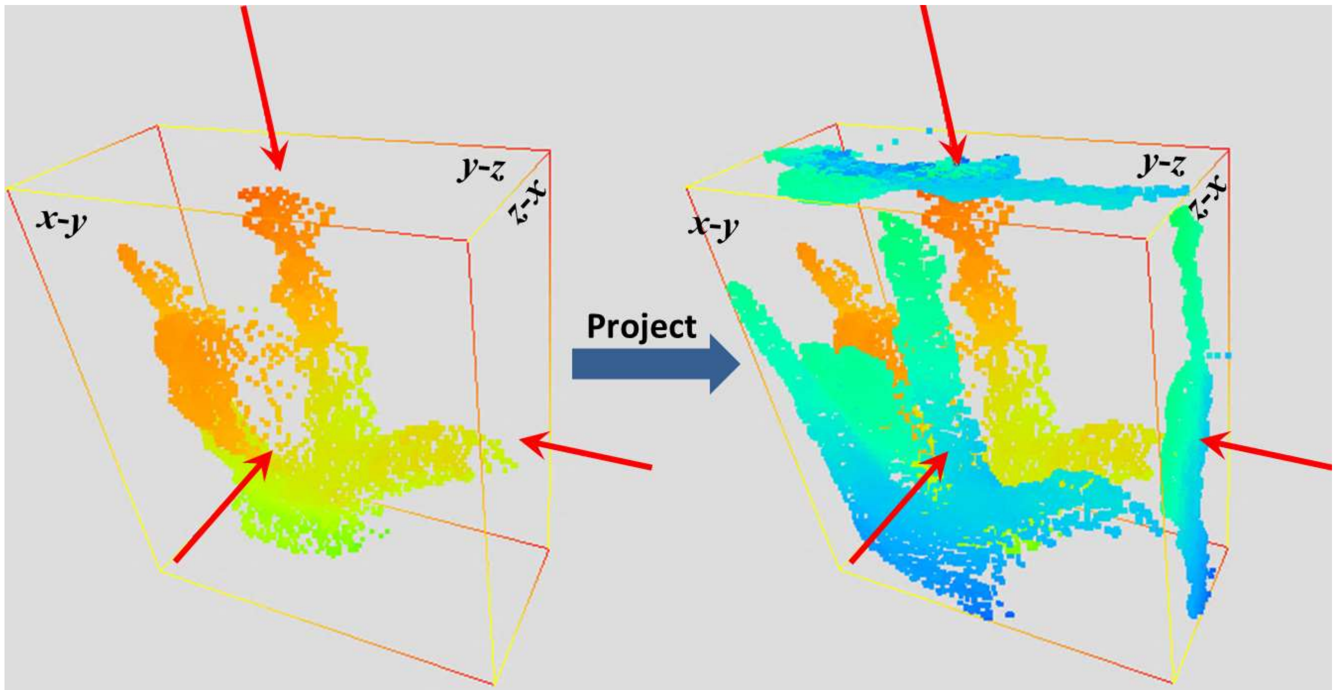


Multi-view CNNs based Method



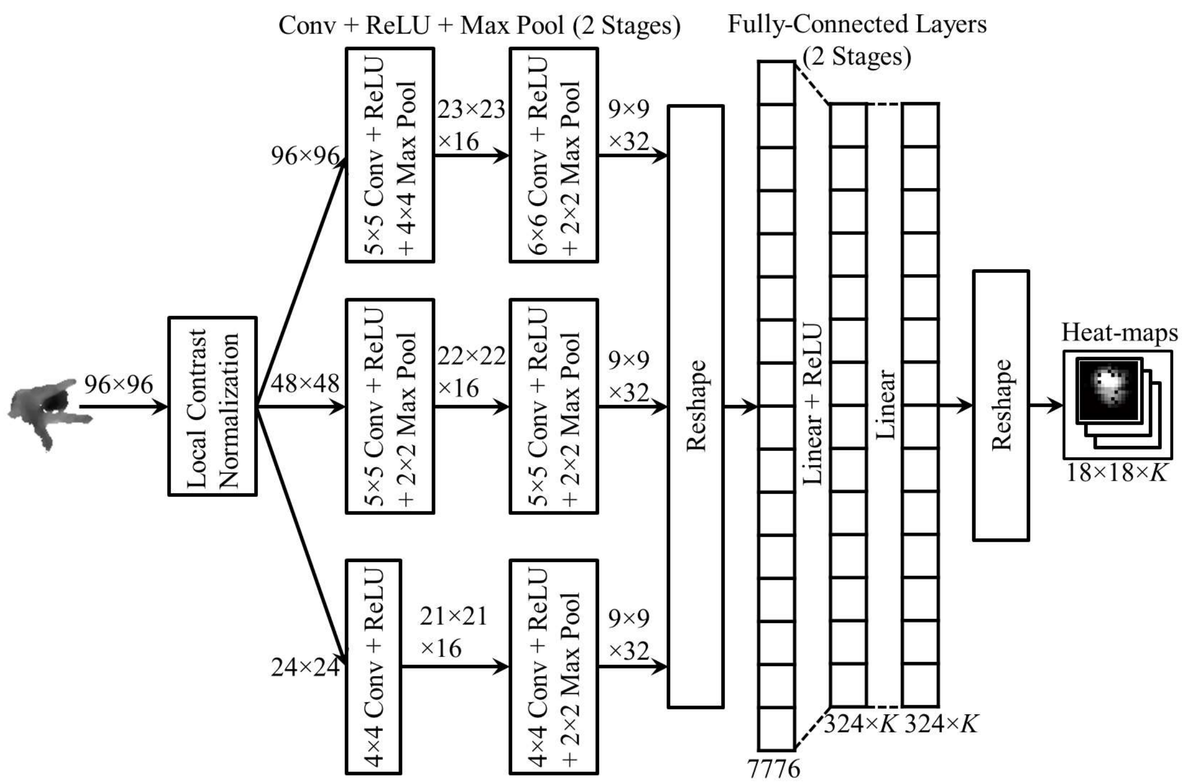
L. Ge, H. Liang, J. Yuan, and D. Thalmann. Robust 3D Hand Pose Estimation in Single Depth Images: From Single-View CNN to Multi-View CNNs. In *CVPR*, 2016.

Multi-view Projection



- **The pixel values on projection images represent the normalized projection distances of 3D points.**

Architecture of CNNs



- **The network generates *K* heat-maps with the size of 18×18 pixels. All of the six views have the same network architecture and the same architectural hyperparameters.**

Multi-view Fusion

Objective: estimate *K* objective hand joint 3D locations

$$\mathbf{\Phi} = \{\boldsymbol{\phi}_k\}_{k=1}^{K} \in \mathbf{\Lambda}$$

Maximum a posteriori estimation

$$\mathbf{\Phi}^* = \arg\max_{\mathbf{\Phi}} P(\mathbf{\Phi} | I_1, I_2, \cdots, I_N)$$ posterior probability

$$= \arg\max_{\mathbf{\Phi}} P(I_1, I_2, \cdots, I_N | \mathbf{\Phi})$$ maximum likelihood

(assume equal a priori probability)

$$= \arg\max_{\mathbf{\Phi}} \prod_{n=1}^{N} P(I_n | \mathbf{\Phi})$$ (assume conditional independence)

$$= \arg\max_{\mathbf{\Phi}} \prod_{n=1}^{N} P(\mathbf{\Phi} | I_n)$$

related with heat-map

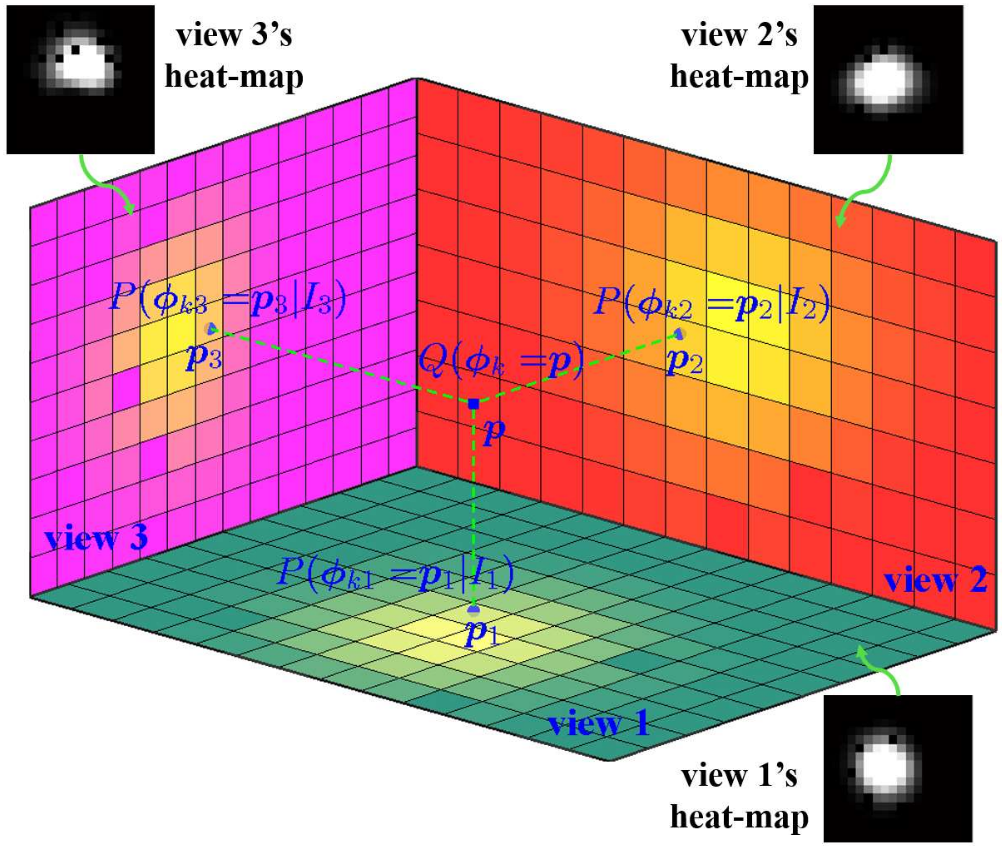
$$= \arg\max_{\mathbf{\Phi}} \prod_{k=1}^{K} \prod_{n=1}^{N} \underline{P(\boldsymbol{\phi}_k | I_n)}$$

$$s.t.\ \mathbf{\Phi} \in \mathbf{\Omega},$$

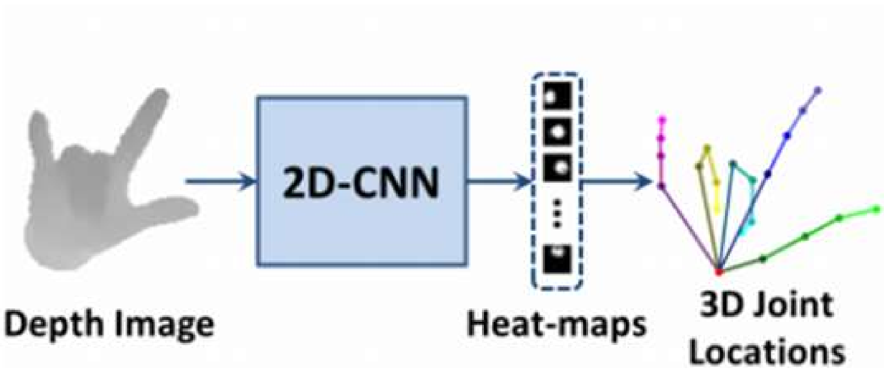
constrained to a low dimensional subspace in order to resolve ambiguous joint estimations

Multi-view Fusion

$$Q\left(\boldsymbol{\phi}_k = \boldsymbol{p}\right) = \prod_{n=1}^{N} P\left(\boldsymbol{\phi}_{kn} = \boldsymbol{p}_n \,|\, I_n\right)$$

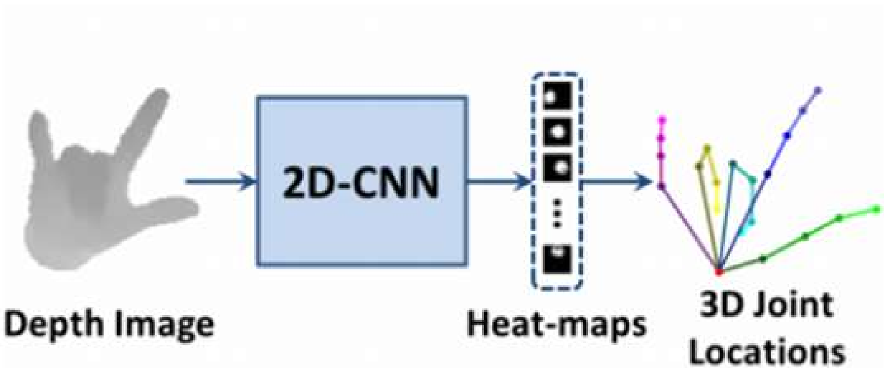


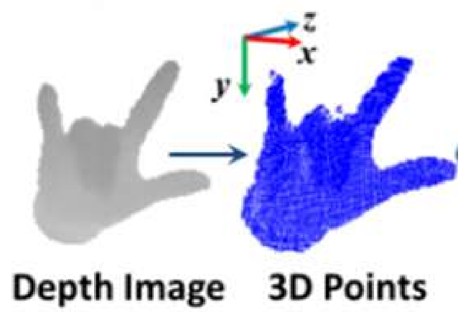
From 2D convolution to 3D convolution?



2D convolution for depth image

From 2D convolution to 3D convolution?





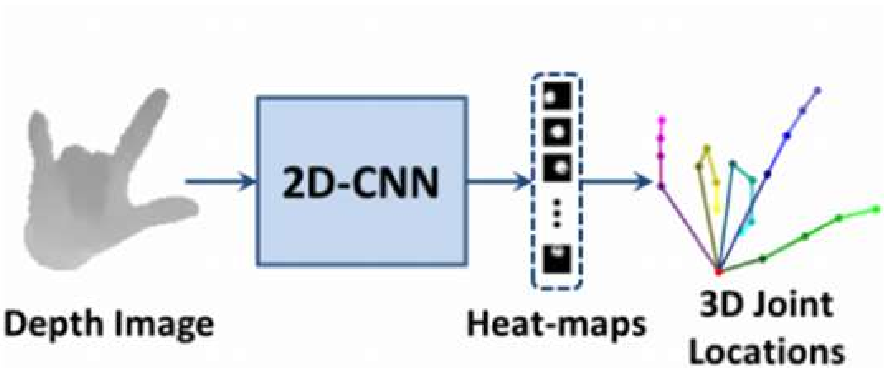
But 3D points are sparse data

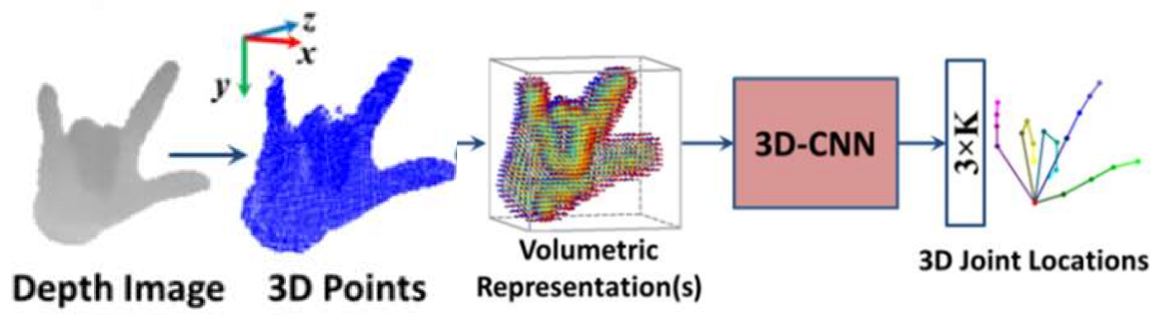
Dense 3D convolution on sparse point clouds will fail

2D convolution for depth image

depth image can be transferred to 3D points

From 2D convolution to 3D convolution?

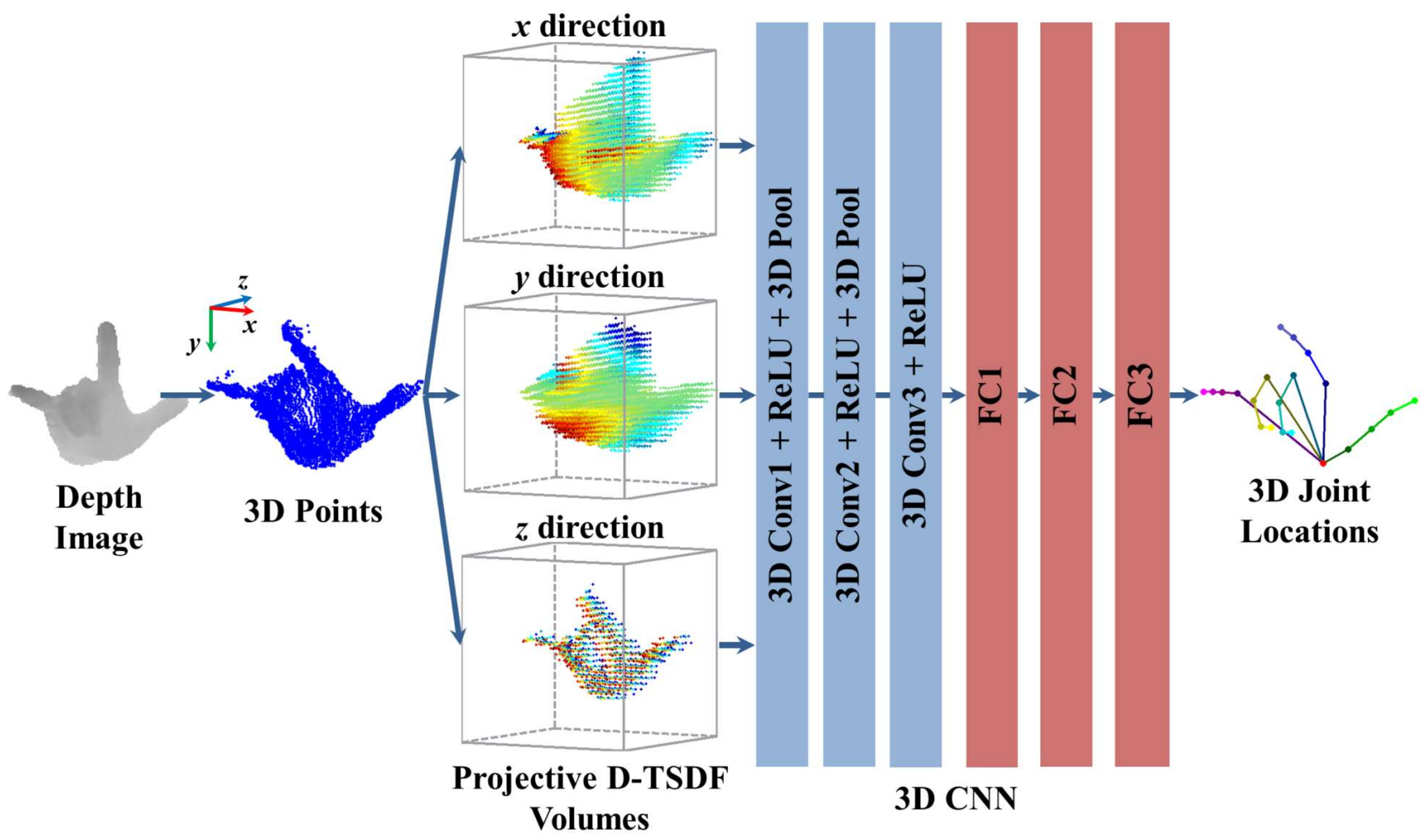




2D convolution for depth image

Transfer sparse 3D points to dense volumetric representation

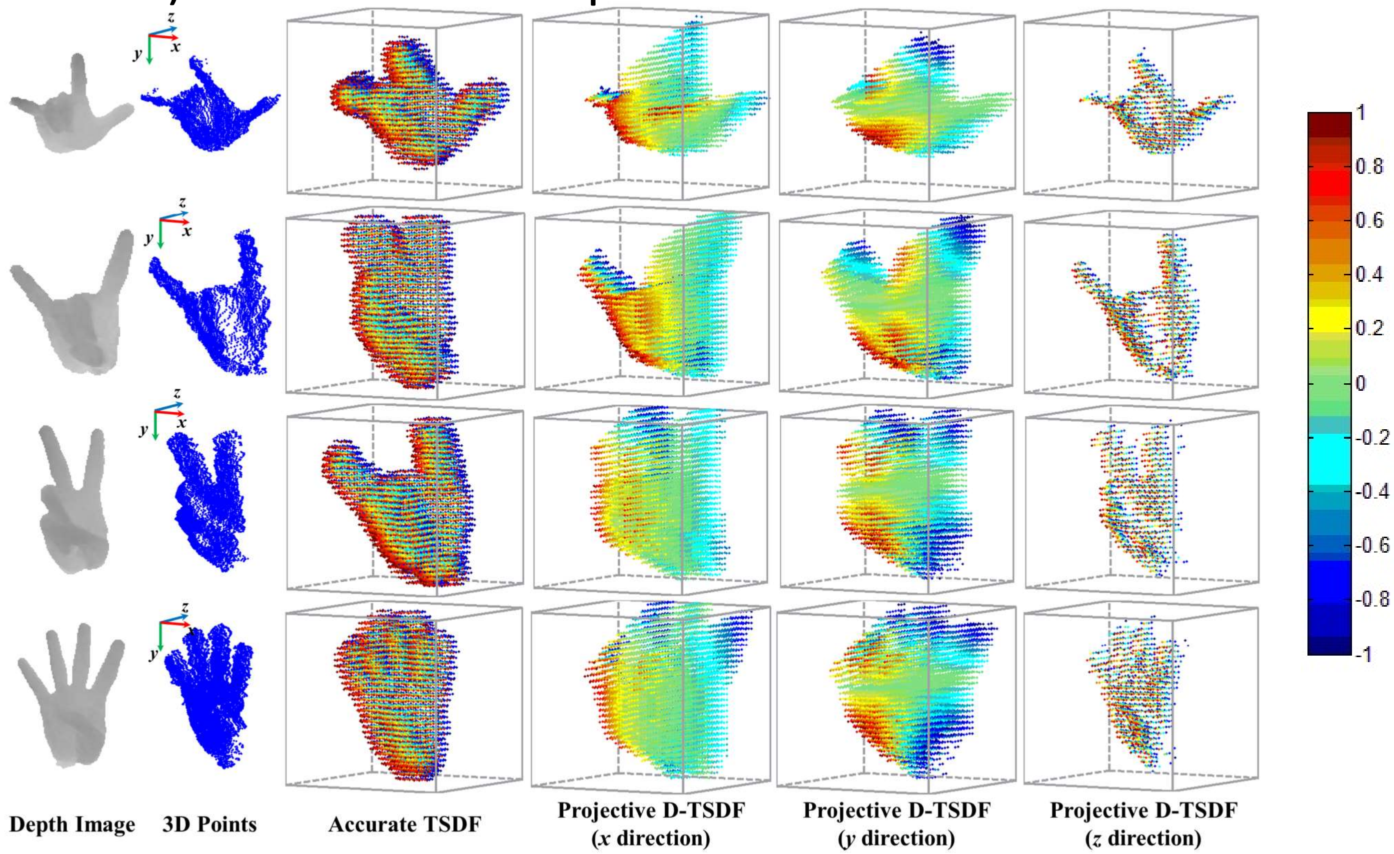
3D CNN for hand pose estimation



- L. Ge, H. Liang, J. Yuan, and D. Thalmann. 3D Convolutional Neural Networks for Efficient and Robust Hand Pose Estimation from Single Depth Images. *CVPR* 2017 and TPAMI 2019.

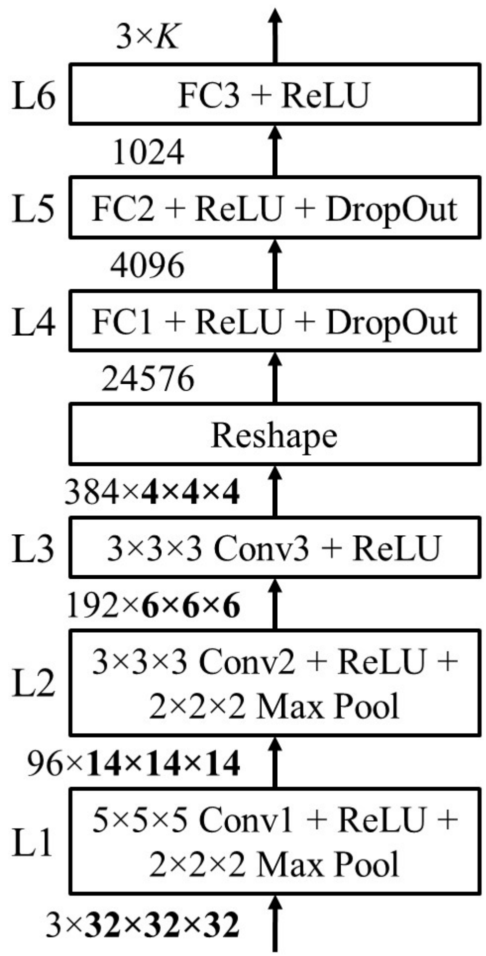
Volumetric Representation

- Projective Directional Truncated Signed Distance Function (D-TSDF) for volumetric representation



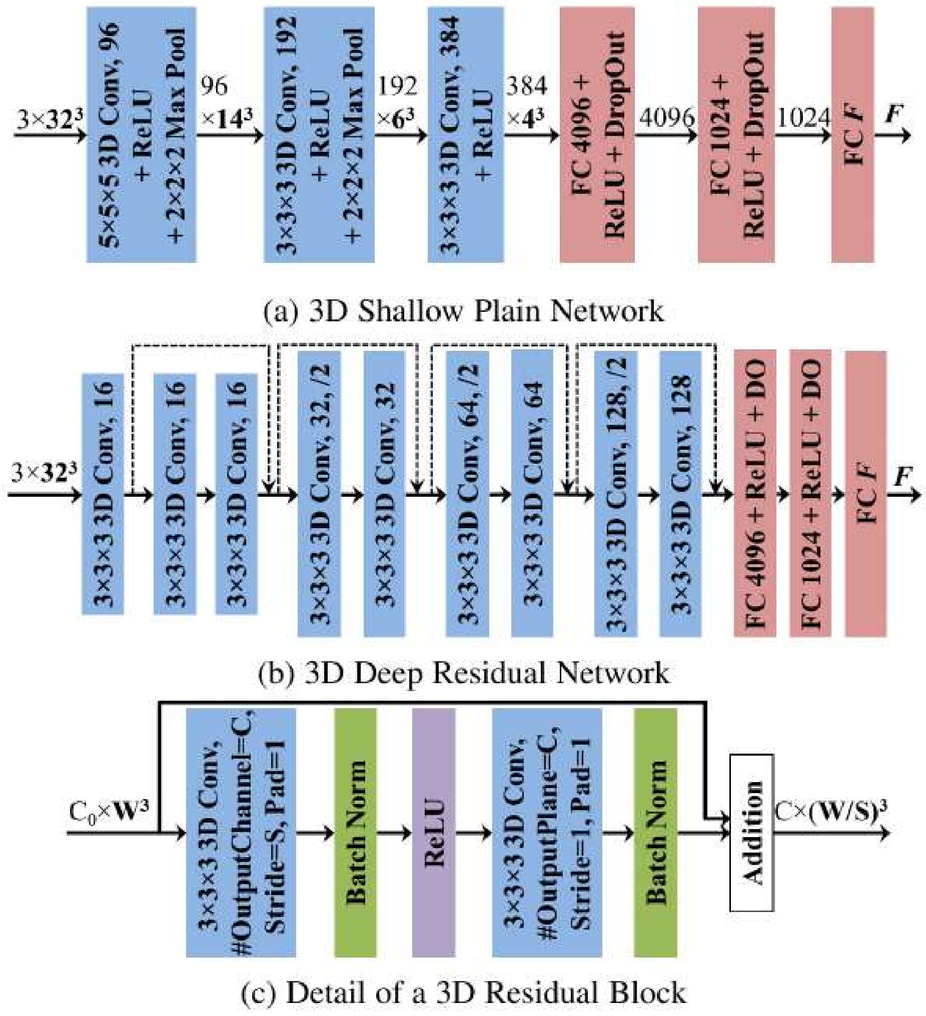
S. Song and J. Xiao, Deep Sliding Shapes for A modal 3D Object Detection in RGB-D Images, CVPR 2016

Network Architecture



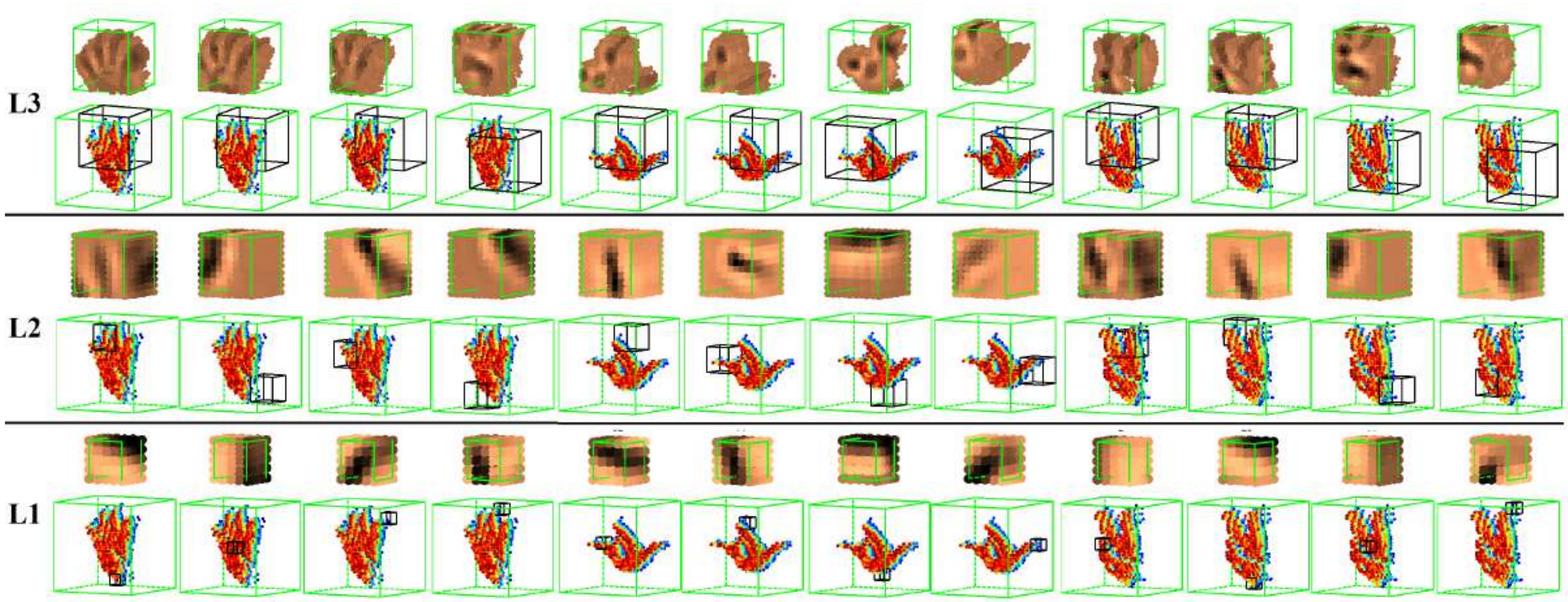
- **Input: three volumes of the projective D-TSDF**
- **Output: a column vector containing 3×*K* elements corresponding to the *K* 3D hand joint relative locations in the volume.**
- **Three 3D convolutional layers and three fully-connected layers**

Network Architecture



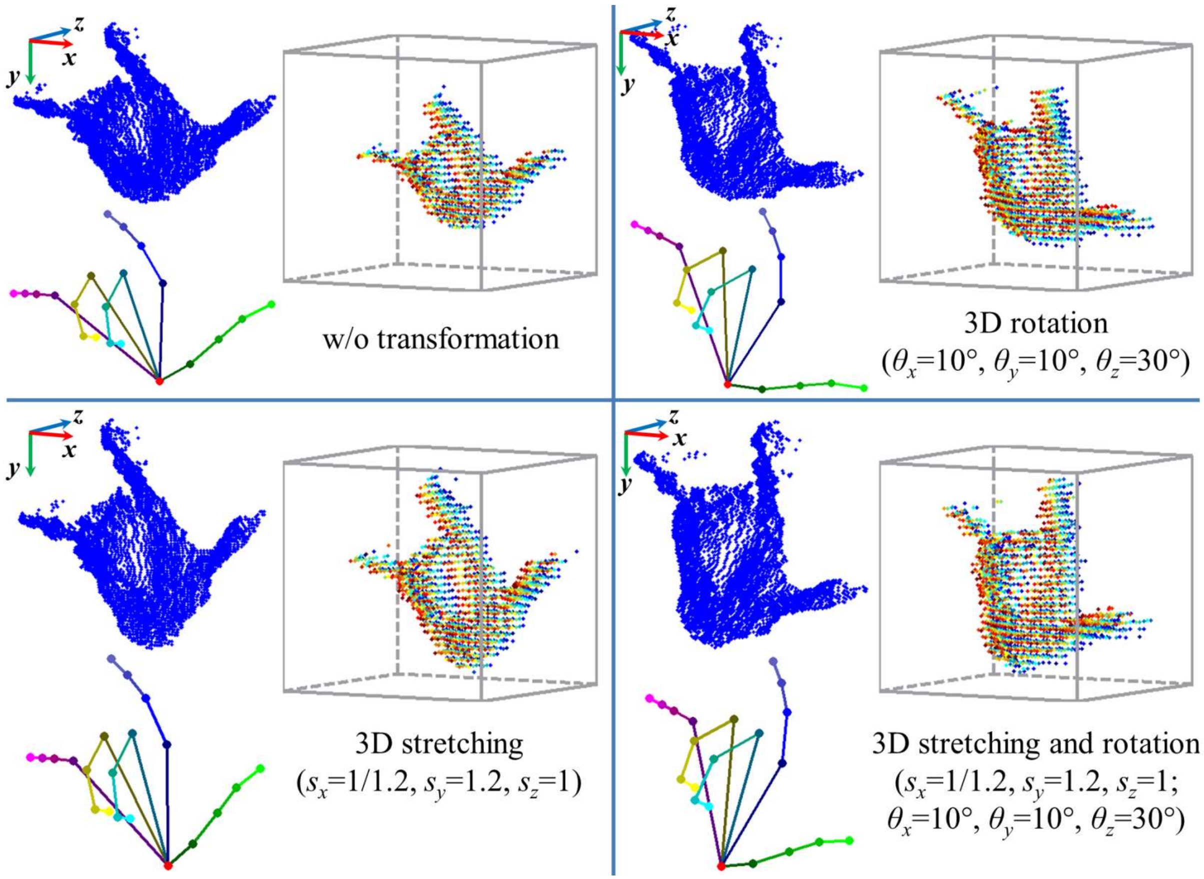
- **Input**: three volumes of the projective D-TSDF
- **Output**: a column vector containing 3×*K* elements corresponding to the *K* 3D hand joint relative locations in the volume.

Patterns learned in 3D shallow network



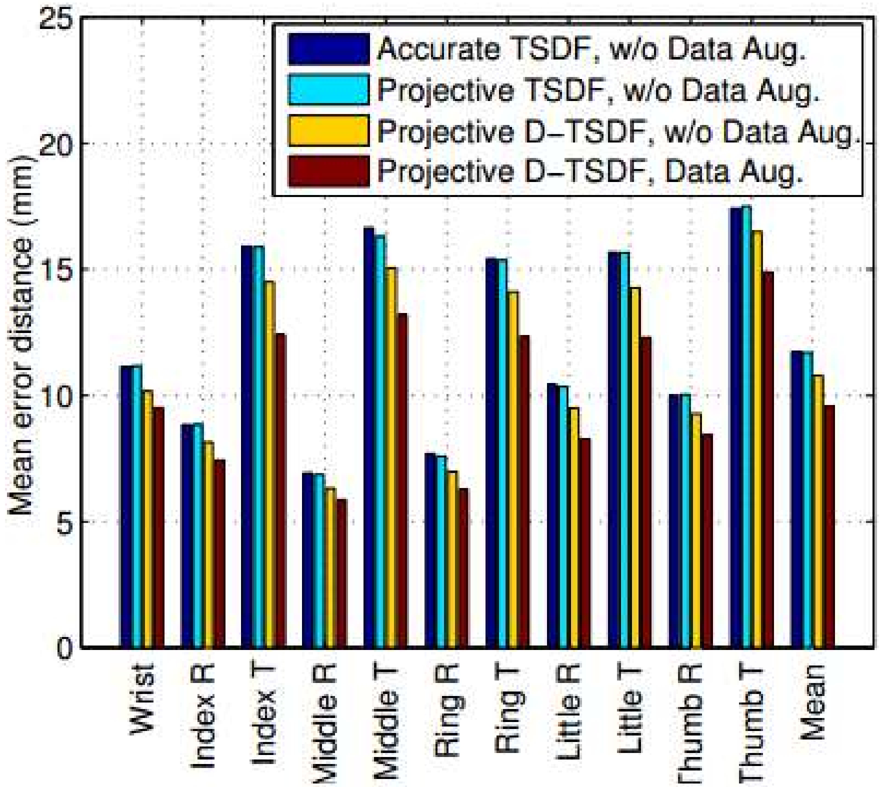
- Neurons in layer 1 (L1) can capture local structures, such as corners and edges;
- neurons in layer 2 (L2) can capture structures of hand part, such as fingers;
- neurons in layer 3 (L3) can capture global structures of hand.

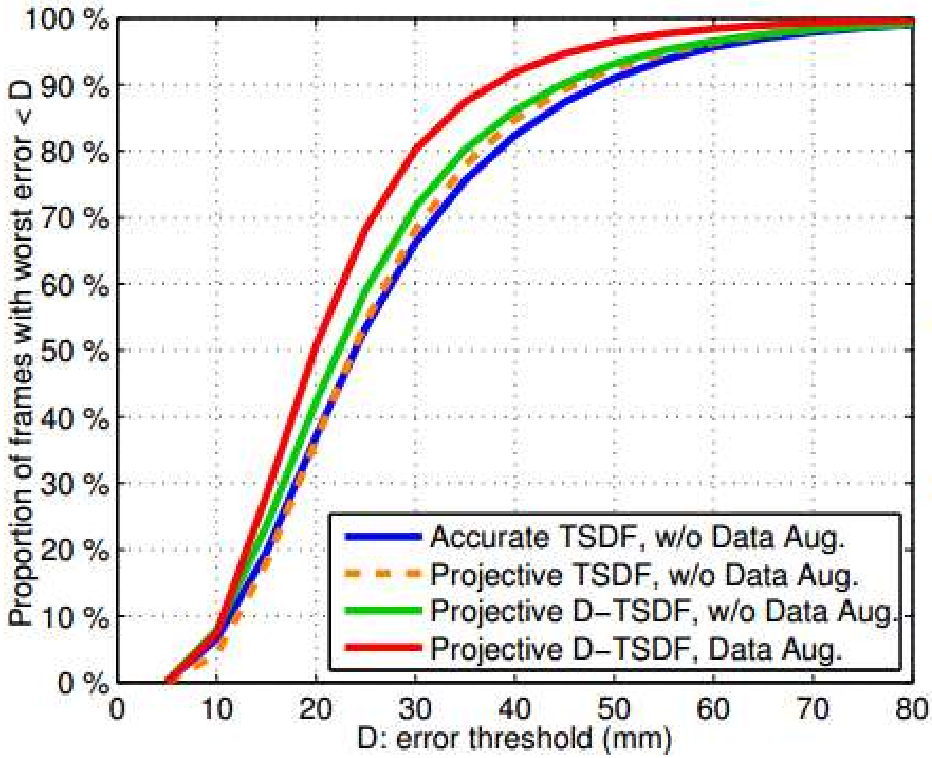
3D Data Augmentation



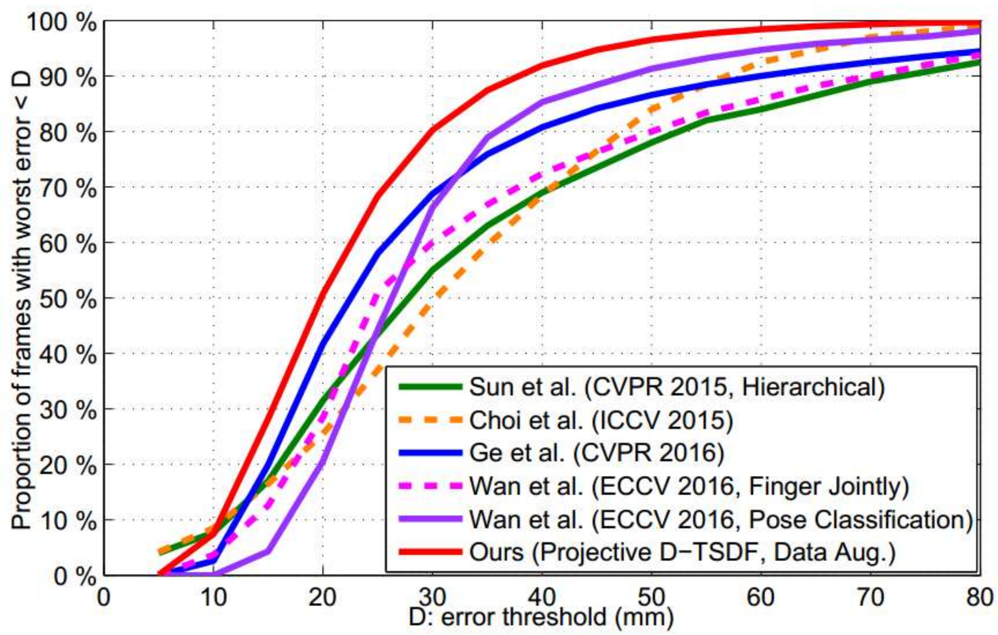
Introducing variations of training data

Data Augmentation Can Help Training



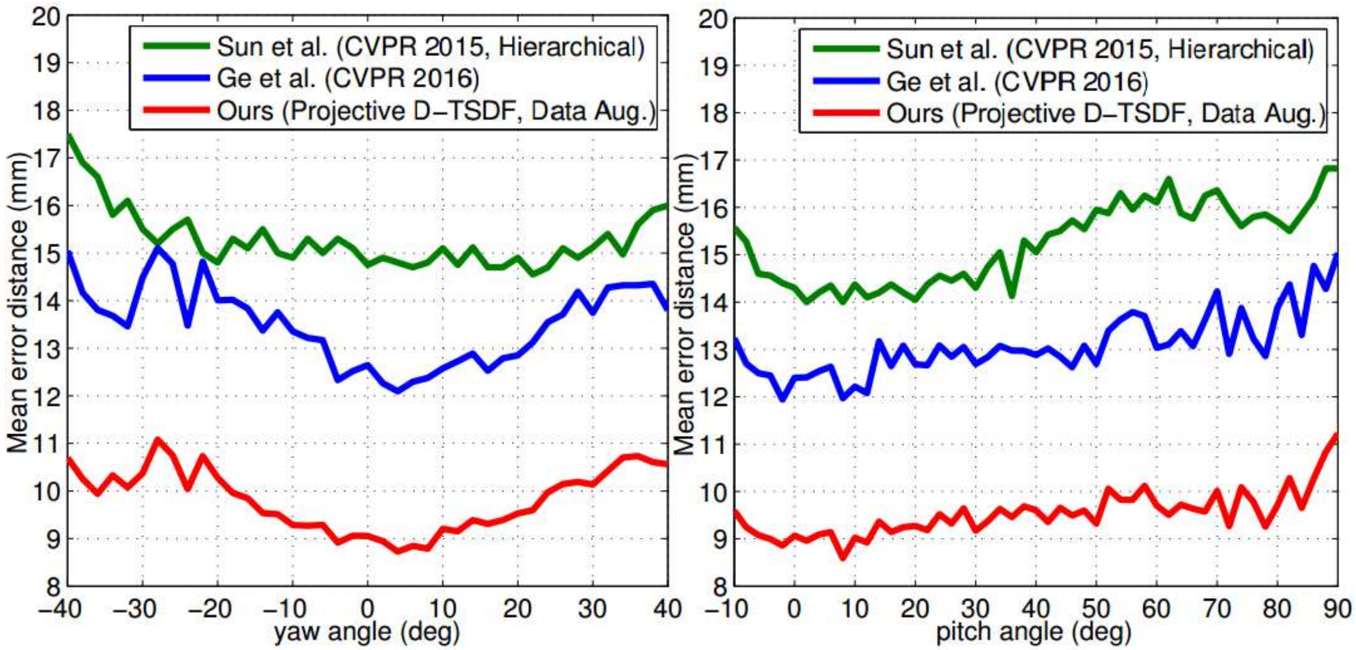


Test on MSRA hand pose dataset



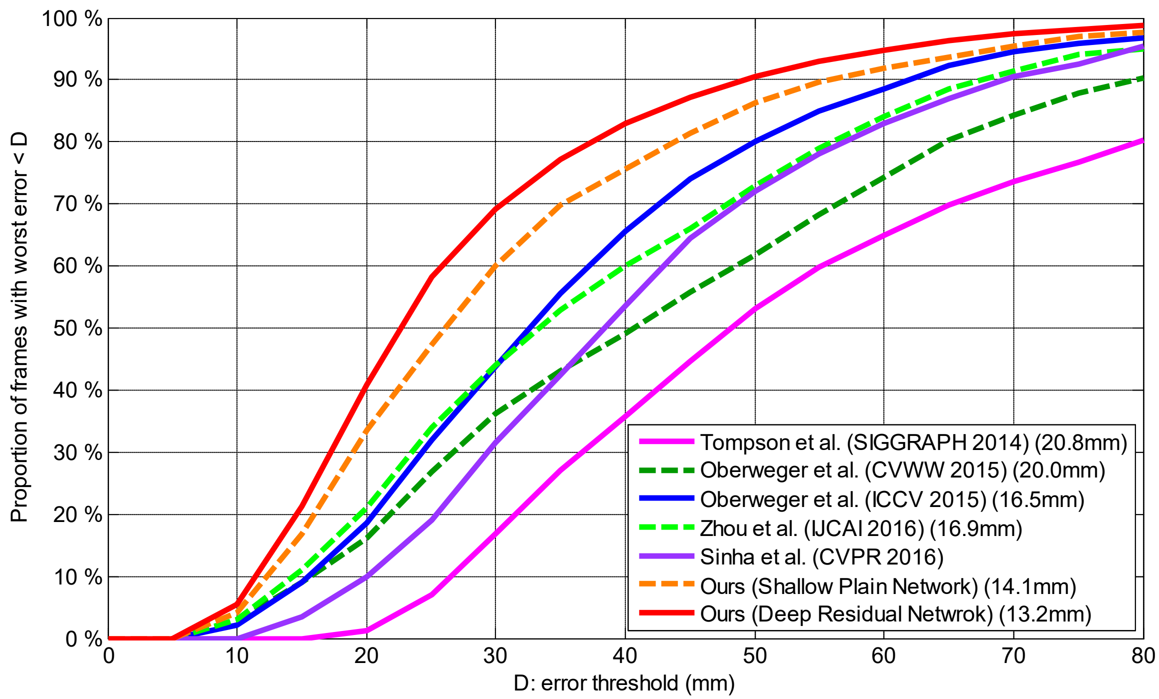
- X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun, "Cascaded hand pose regression," *CVPR*, 2015.
- C. Choi, A. Sinha, J. Hee Choi, S. Jang, and K. Ramani, "A collaborative filtering approach to real-time hand pose estimation," *ICCV*, 2015.
- L. Ge, H. Liang, J. Yuan, and D. Thalmann, "Robust 3D hand pose estimation in single depth images: from single-view CNN to multi-view CNNs," *CVPR*, 2016.
- C. Wan, A. Yao, and L. Van Gool, "Direction matters: hand pose estimation from local surface normals," *ECCV*, 2016.

Test on MSRA hand pose dataset



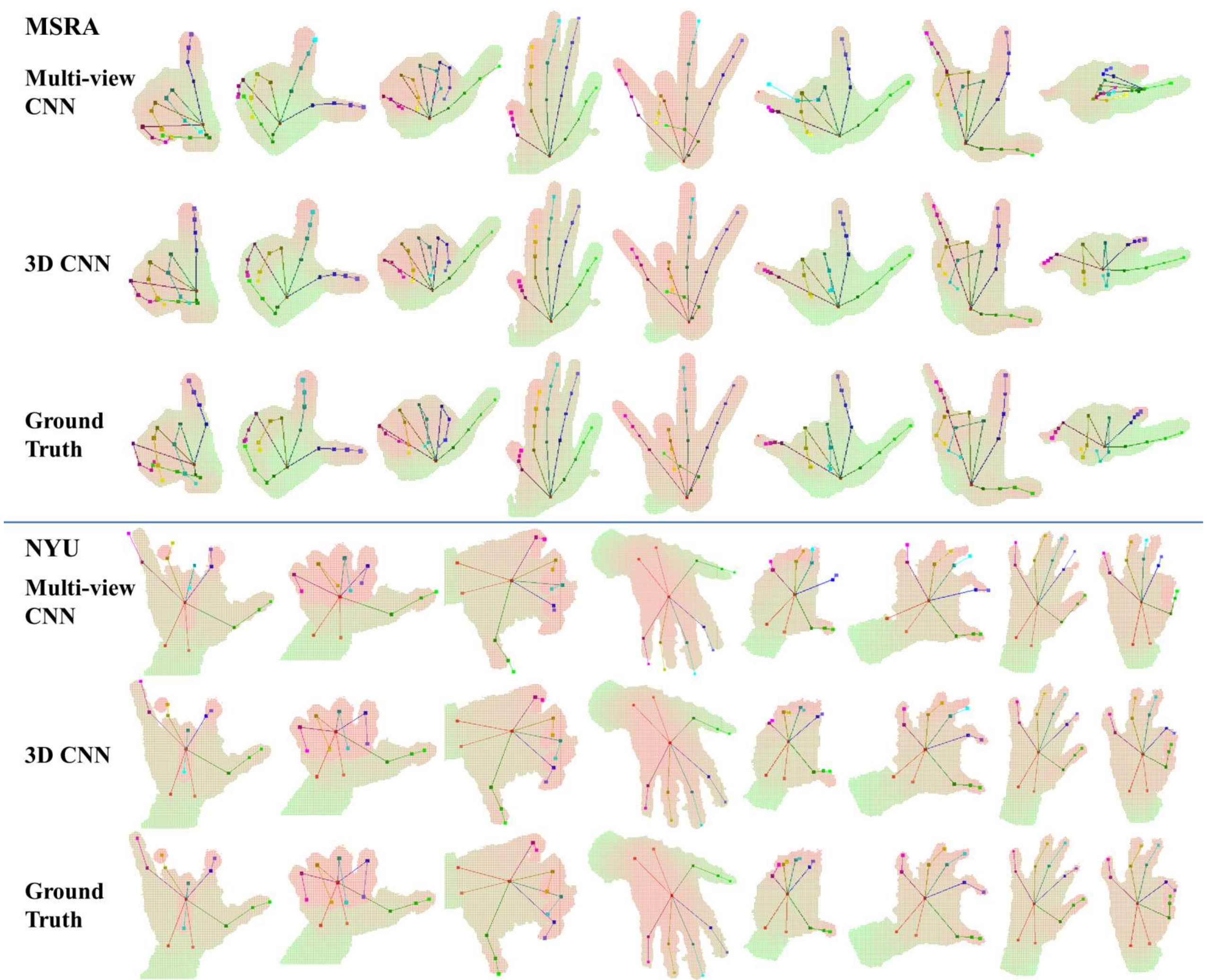
- X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun, "Cascaded hand pose regression," *CVPR*, 2015.
- C. Choi, A. Sinha, J. Hee Choi, S. Jang, and K. Ramani, "A collaborative filtering approach to real-time hand pose estimation," *ICCV*, 2015.
- L. Ge, H. Liang, J. Yuan, and D. Thalmann, "Robust 3D hand pose estimation in single depth images: from single-view CNN to multi-view CNNs," *CVPR*, 2016.
- C. Wan, A. Yao, and L. Van Gool, "Direction matters: hand pose estimation from local surface normals," *ECCV*, 2016.

Test on NYU hand pose dataset



- J. Tompson, M. Stein, Y. Lecun, and K. Perlin, “Real-time continuous pose recovery of human hands using convolutional networks,” *SIGGRAPH*. 2014.
- M. Oberweger, P. Wohlhart, and V. Lepetit, “Training a feedback loop for hand pose estimation,” *ICCV*, 2015.
- X. Zhou, Q. Wan, W. Zhang, X. Xue, and Y. Wei, “Model-based deep hand pose estimation,” *IJCAI*, 2016.
- A. Sinha, C. Choi, and K. Ramani, “Deephand: Robust hand pose estimation by completing a matrix with deep features,” *CVPR*, 2016.

Sample Results



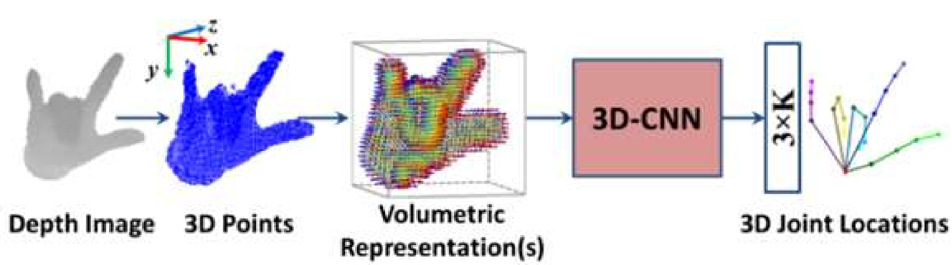
3D Convolutional Neural Networks for Efficient and Robust Hand Pose Estimation from Single Depth Images

Liuhao Ge, Hui Liang, Junsong Yuan, Daniel Thalmann

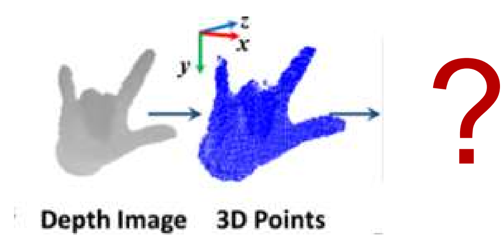
Institute for Media Innovation

Nanyang Technological University

Can we process 3D point cloud directly instead of 3D convolution?



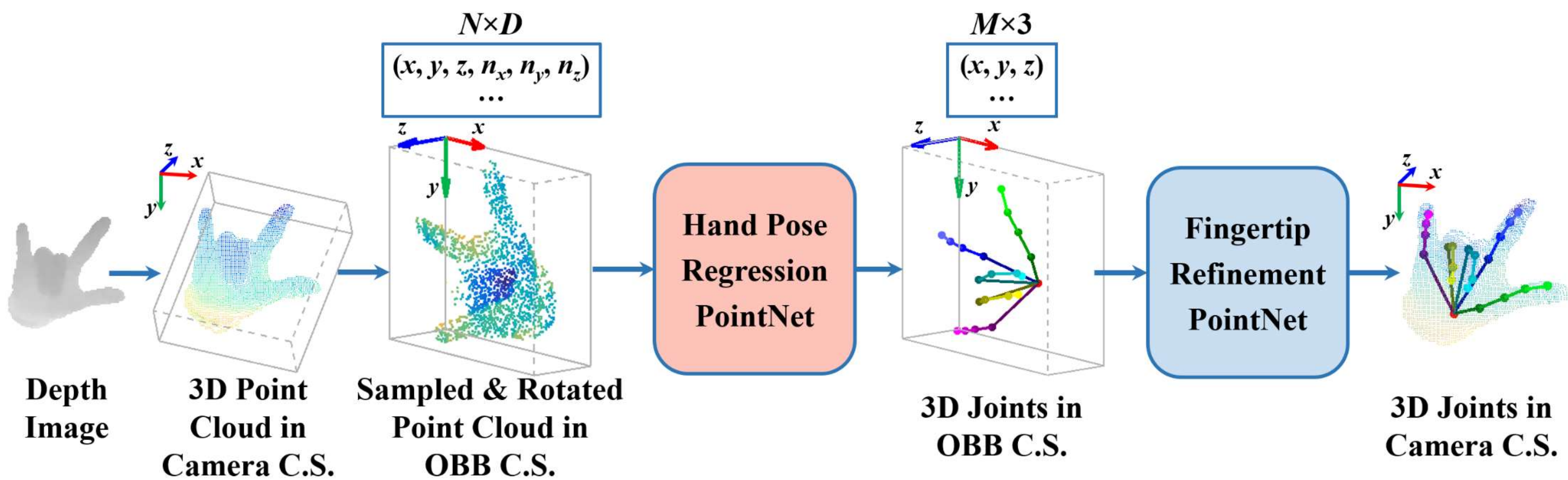
3D convolution for depth image



Process 3D point cloud directly?

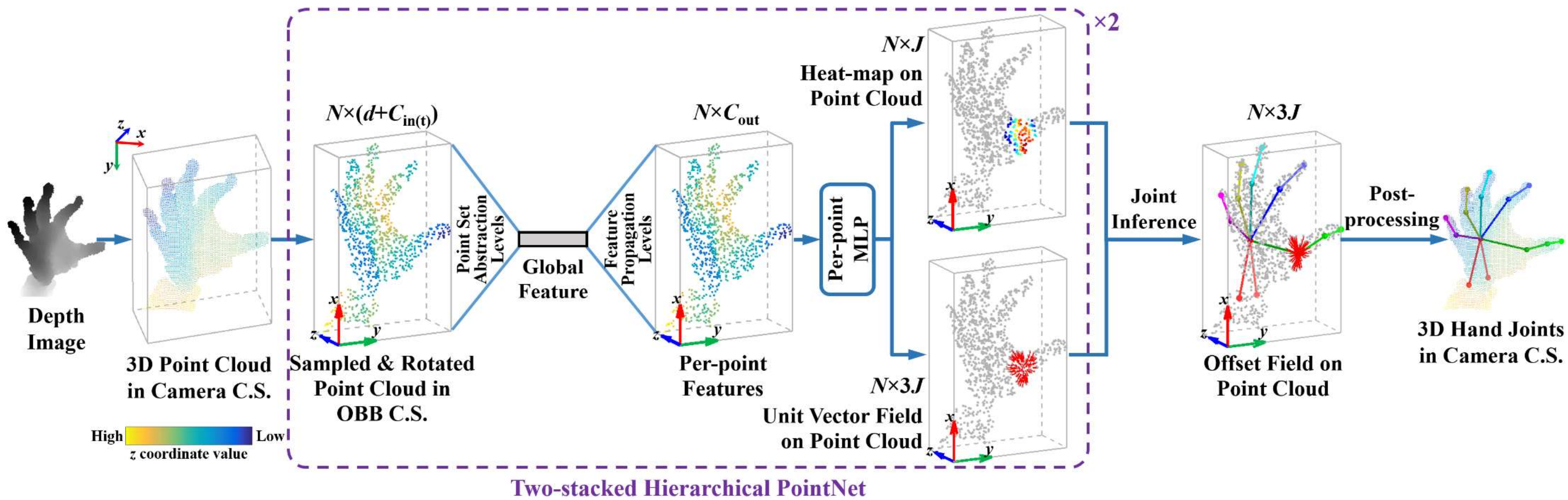
Hand PointNet

A <u>point cloud</u> based hand pose estimation approach by holistically regressing the 3D hand pose.



Further improvement

Estimate the point-wise closeness and offset directions to hand joints from the input point cloud using a stacked point-to-point regression PointNet, which is able to capture local evidence for estimating accurate 3D hand pose



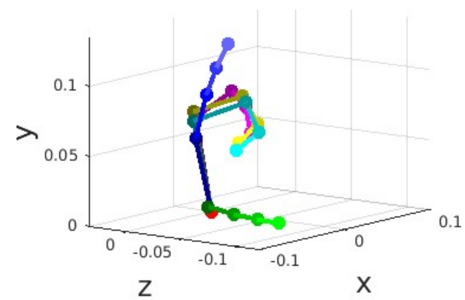
Hand PointNet: 3D Hand Pose Estimation using Point Sets

Liuhao Ge[1], Yujun Cai[1], Junwu Weng[1], Junsong Yuan[2]

[1]Nanyang Technological University
[2]State University of New York at Buffalo

3D hand pose estimation:
Can we use RGB camera instead of depth camera?

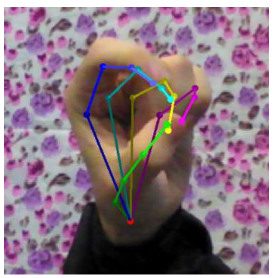
Monocular RGB-based Approach



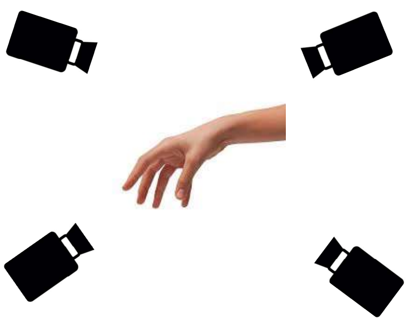
From 2D images to 3D skeleton results

Challenges: difficult to obtain 3D labeled data

For Real Dataset:



annotate accurate 3D hand pose is difficult

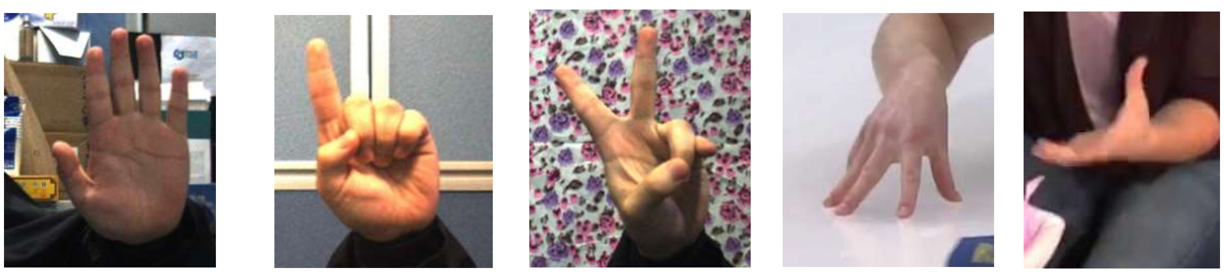


- **Multi-view annotation method is labor-costing**
- **Reconstructed 3D labels may not be perfect**

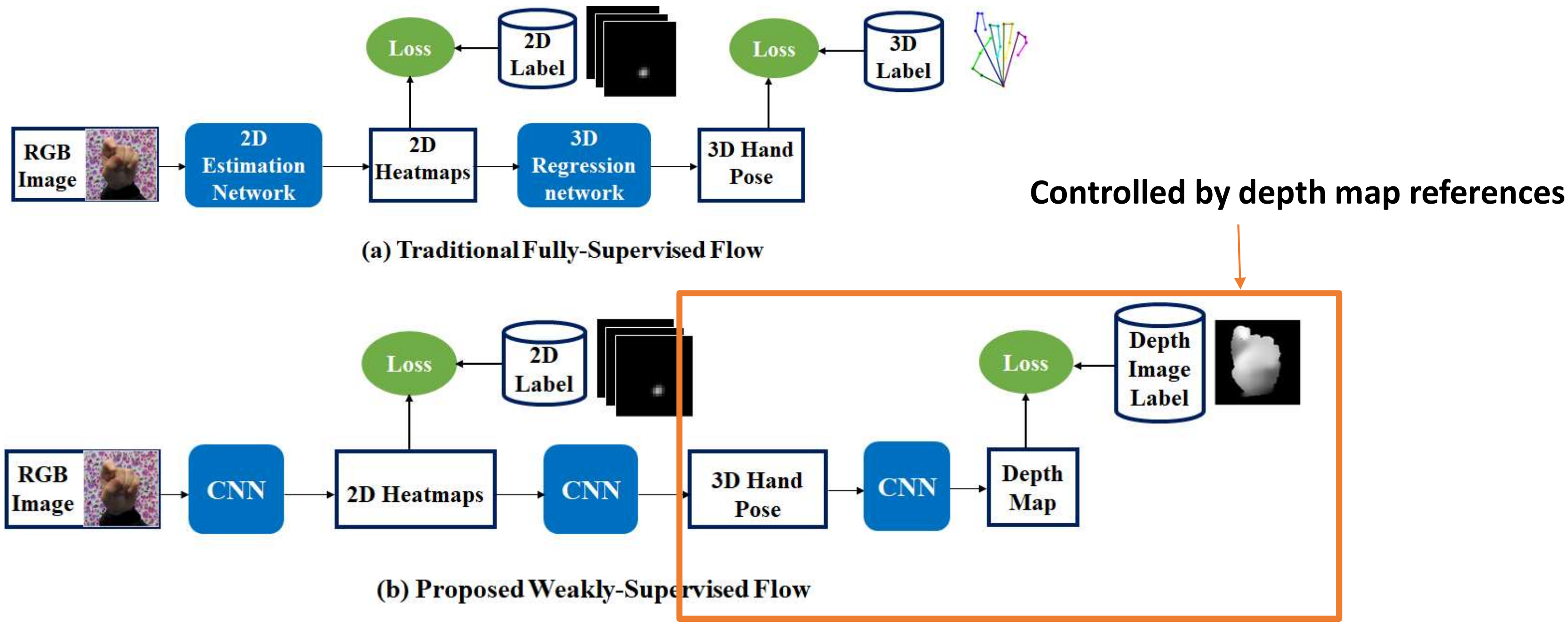
Using synthetic data for machine learning?

synthetic dataset for hand pose [Zimmermann et al. ICCV 2017]

Synthetic data can provide accurate 3D annotations while quite <u>different from real ones</u>



Weakly Supervised Learning

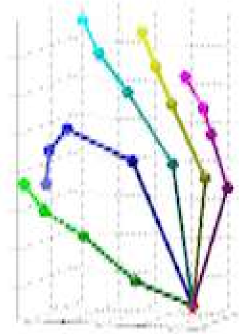


Y Cai, L Ge, J Cai, J Yuan, Weakly-supervised 3d hand pose estimation from monocular RGB images, ECCV'18

From 3D hand pose estimation to joint 3D hand pose and shape estimation

Joint hand pose and shape estimation

Input Image

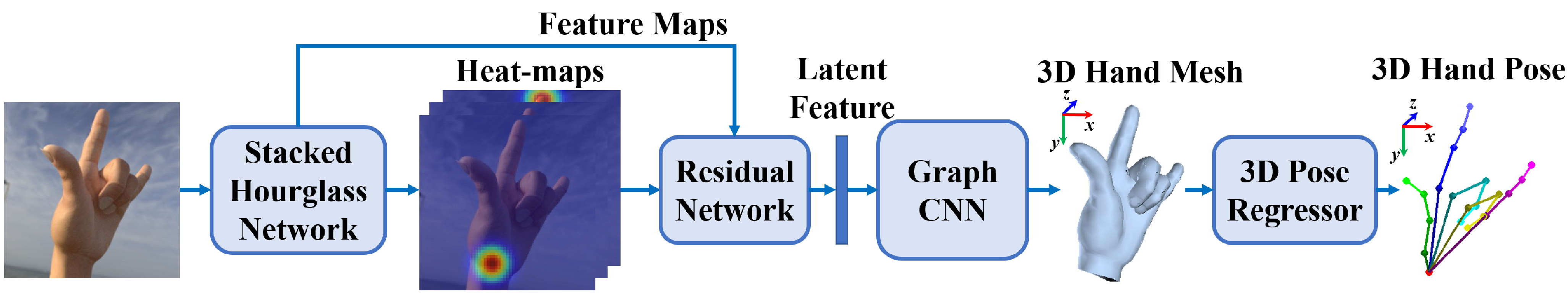


2D/3D Locations of Hand Joints

Challenges

- **High dimensionality of the output space (3D mesh)**
  - ➢We propose a novel Graph CNN-based approach to generate 3D hand mesh vertices in a graph
- **Lack of ground truth 3D hand mesh training data for real-world images**
  - ➢We propose a novel weakly-supervised method by leveraging depth map as a weak supervision for 3D mesh generation

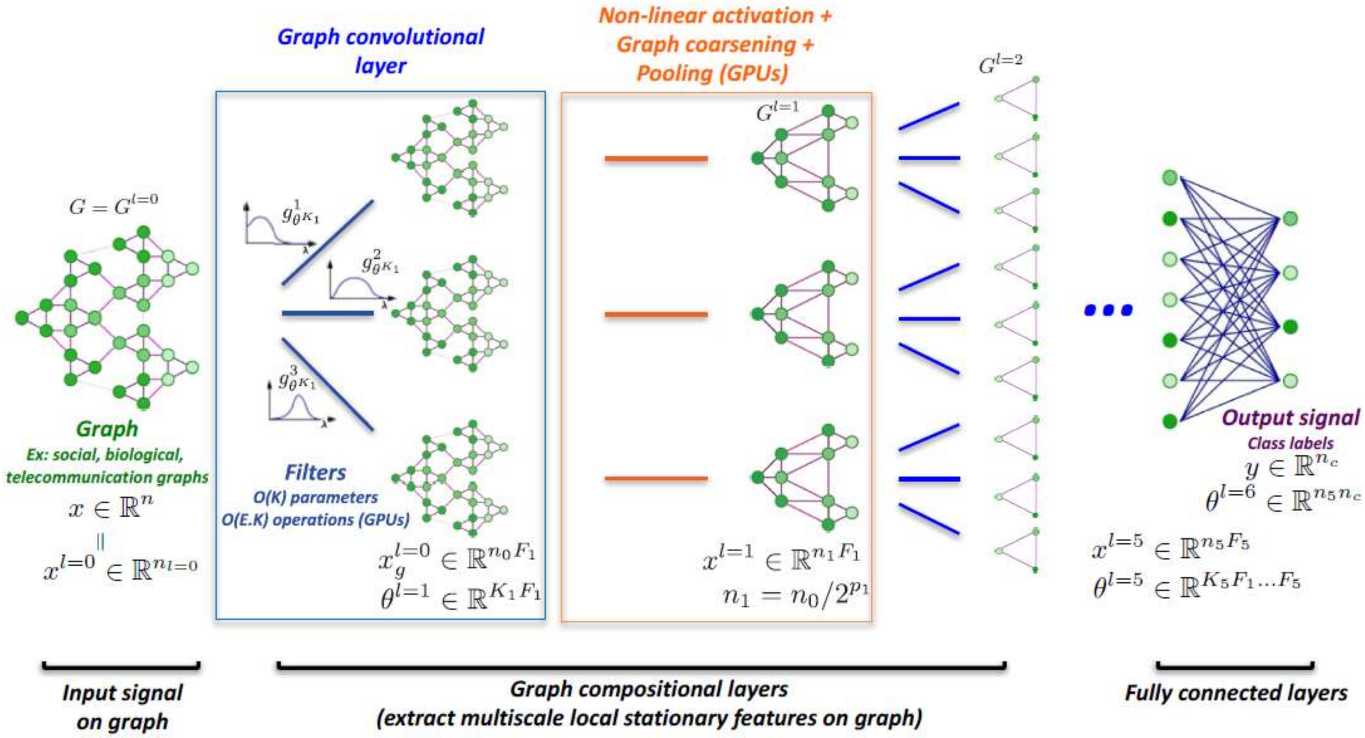
Method – Overview



If we cannot solve a simple problem, try a complex one

Method – Graph CNN

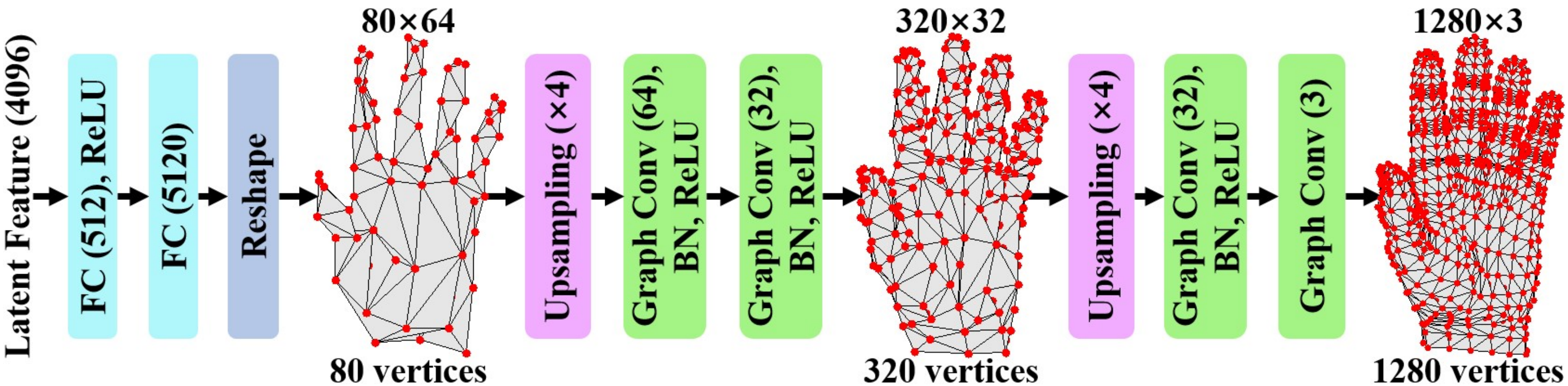
Chebyshev Spectral Graph CNN [1]



[1] Michael Defferrard, *et al*. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In *NeurIPS*, 2016.

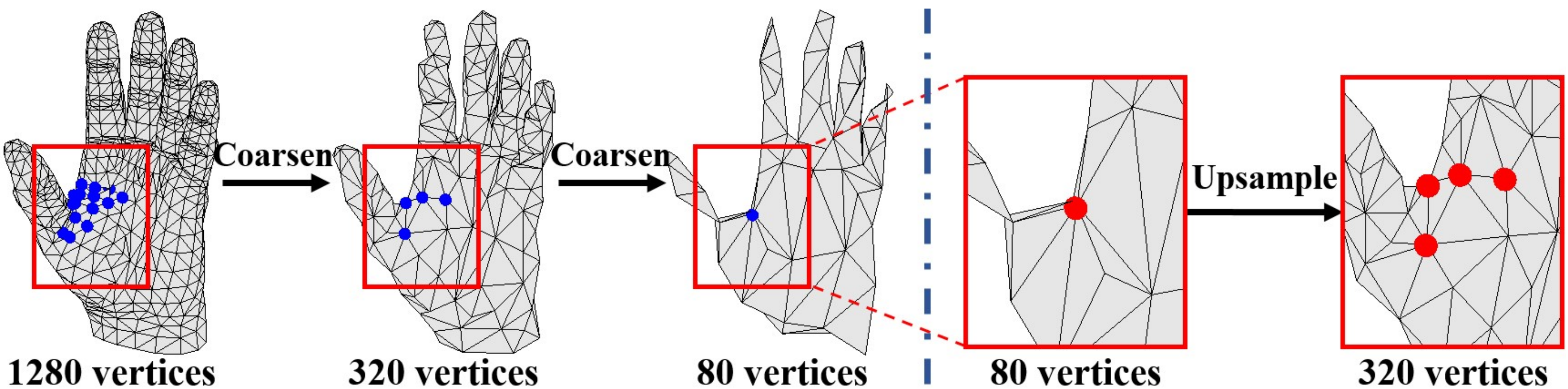
Method – Graph CNN

Graph CNN for Mesh Generation



Method – Graph CNN

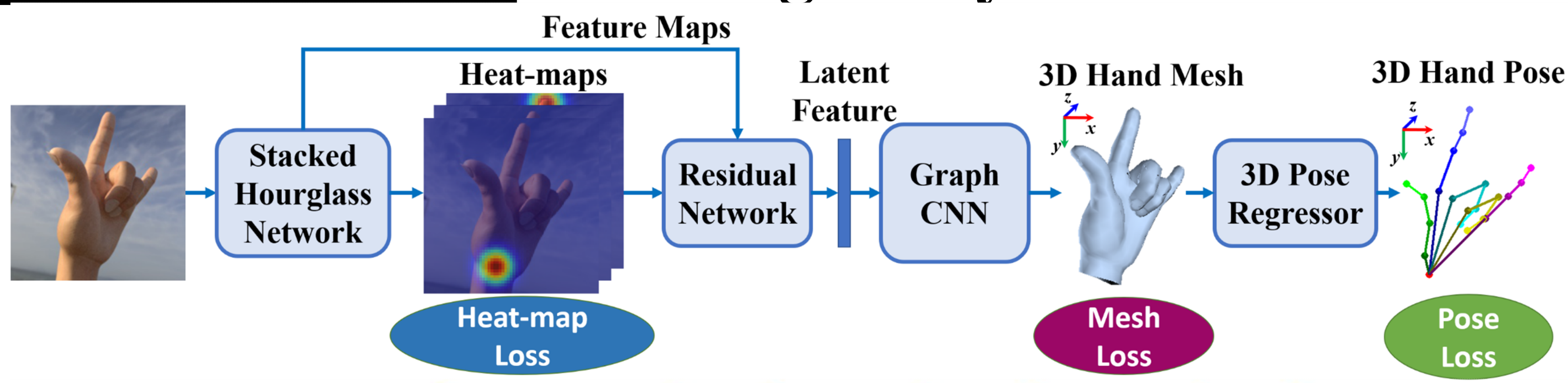
Graph CNN for Mesh Generation



(a) Graph Coarsening (b) Example of Feature Upsampling

Method – Training

Fully-supervised Training on Synthetic Dataset

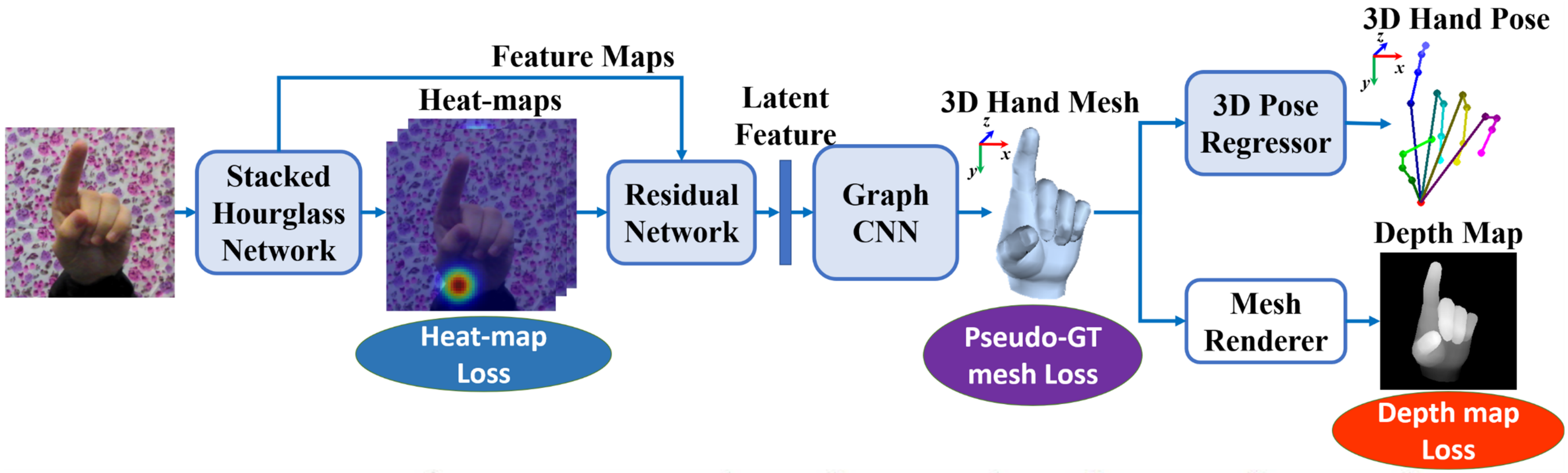


Loss Function $\mathcal{L}_{fully} = \lambda_{\mathcal{H}}\underbrace{\mathcal{L}_{\mathcal{H}}}_{\text{Heat-map Loss}} + \lambda_{\mathcal{M}}\underbrace{\mathcal{L}_{\mathcal{M}}}_{\text{Mesh Loss}} + \lambda_{\mathcal{J}}\underbrace{\mathcal{L}_{\mathcal{J}}}_{\text{3D Pose Loss}}$

Mesh Loss $\mathcal{L}_{\mathcal{M}} = \lambda_v\underbrace{\mathcal{L}_v}_{\text{Vertex Loss}} + \lambda_n\underbrace{\mathcal{L}_n}_{\text{Normal Loss}} + \lambda_e\underbrace{\mathcal{L}_e}_{\text{Edge Loss}} + \lambda_l\underbrace{\mathcal{L}_l}_{\text{Laplacian Loss}}$

Method – Training

Weakly-supervised Finetuning on Real-world Dataset



Loss Function $\mathcal{L}_{weakly} = \lambda_{\mathcal{H}} \underline{\mathcal{L}_{\mathcal{H}}} + \lambda_{\mathcal{D}} \underline{\mathcal{L}_{\mathcal{D}}} + \lambda_{p\mathcal{M}} \underline{\mathcal{L}_{p\mathcal{M}}}$

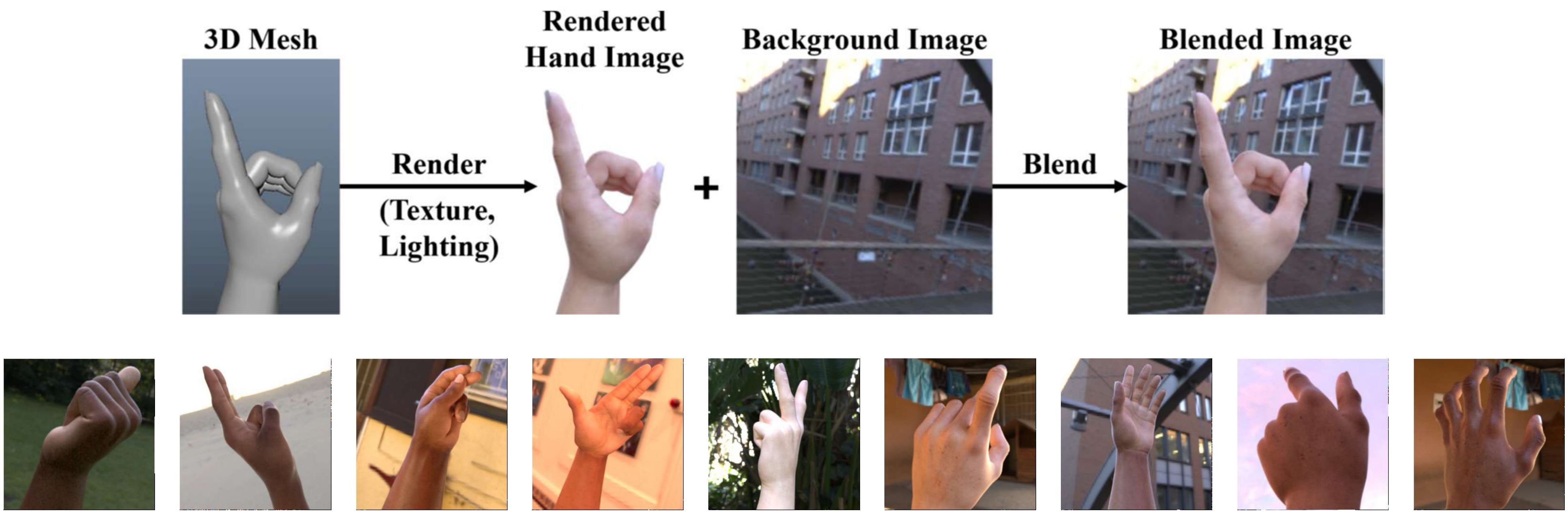
Heat-map Loss

Depth Map Loss

Pseudo-GT Mesh Loss

Synthetic Dataset Creation

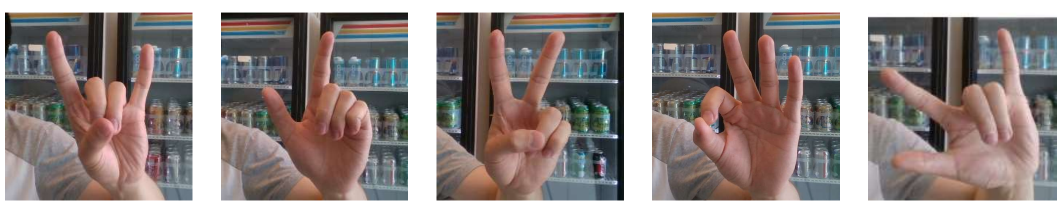
A Large Synthetic Dataset for Training and Validation (375,000 RGB images with hand mesh and pose annotations)



3D Hand Shape and Pose Dataset

- **A Large Synthetic Dataset for Training and Validation (375,000 hand RGB images)**

- **A Real-world Dataset for Testing (583 hand RGB images)**



Experiments

Evaluation of 3D Hand Mesh Reconstruction

Error (mm)	−Normal	−Edge	−Laplacian	−3D Pose	Full
Mesh error	8.34	9.09	8.63	9.04	**7.95**
Pose error	8.30	9.06	8.55	9.24	**8.03**

Ablation study by eliminating different loss terms from our fully-supervised training loss.

Experiments

Evaluation of 3D Hand Mesh Reconstruction

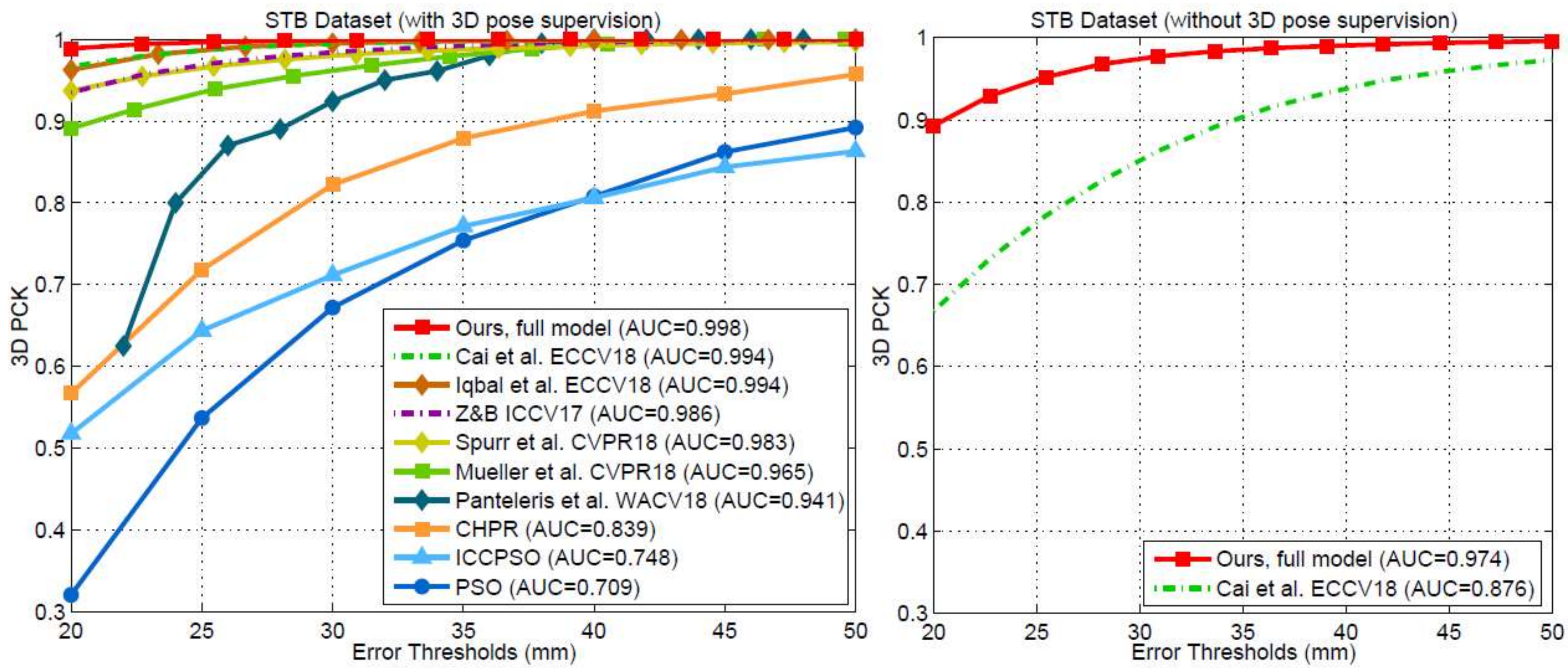
Mesh error (mm)	MANO-based	Direct LBS	Ours
Our synthetic dataset	12.12	10.32	**8.01**
Our real-world dataset	20.86	13.33	**12.72**

Comparison with direct Linear Blend Skinning (LBS) method and MANO-based method.

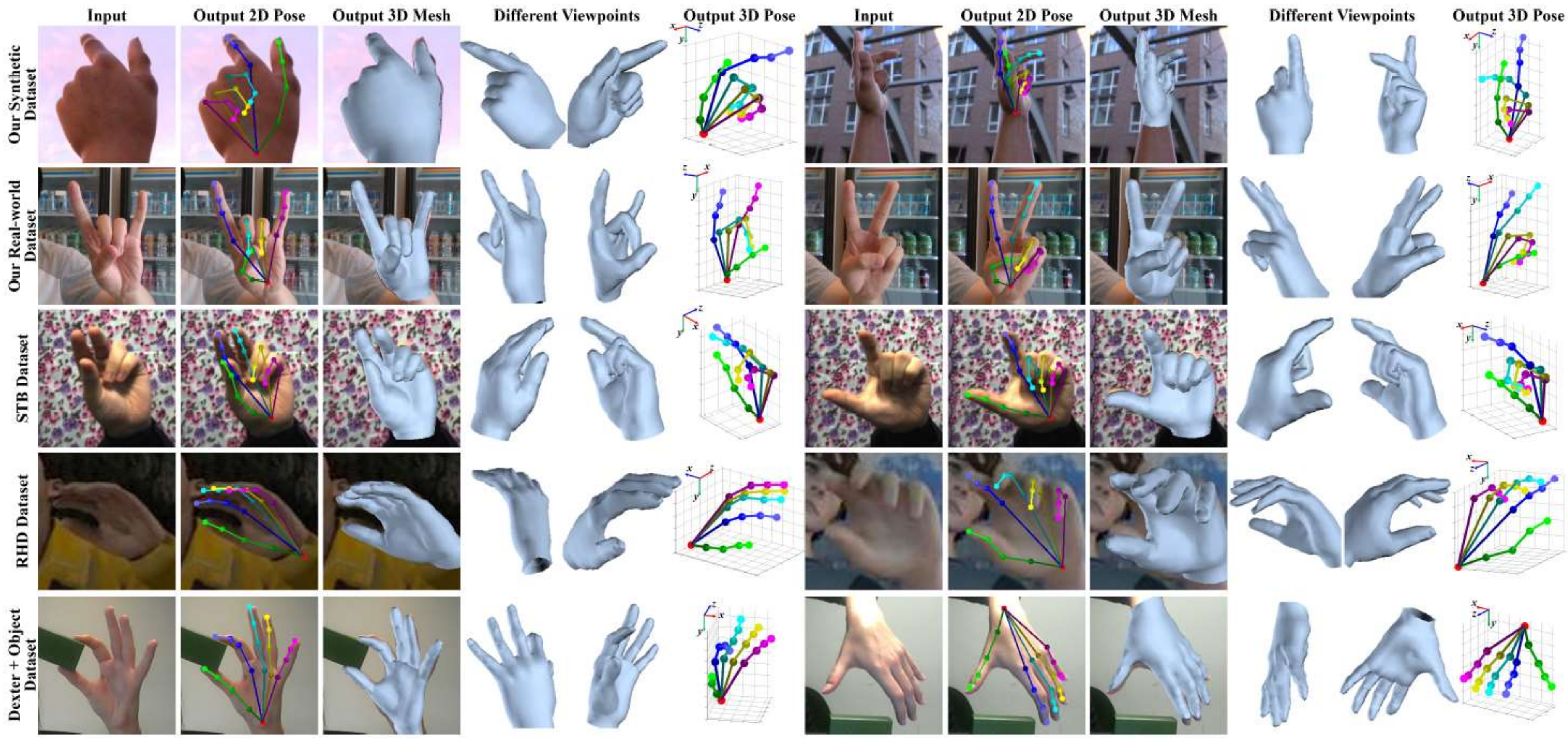
Experiments

Evaluation of 3D Hand Pose Estimation

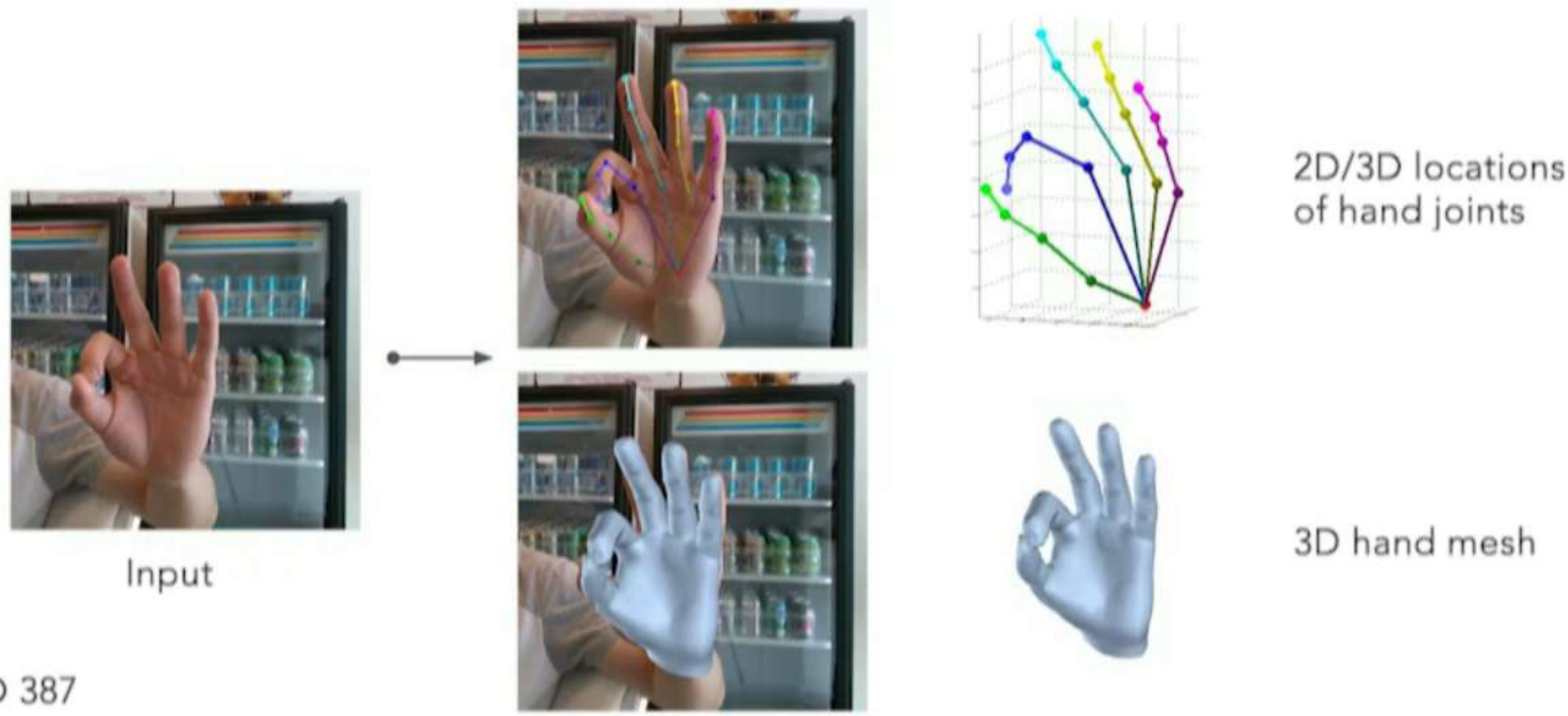
Comparisons with state-of-the-art methods on STB dataset



3D mesh + 3D pose estimation



3D Hand Shape and Pose Estimation from a Single RGB Image



PaperID 387

Summary

- Hand Sensing for Augmented interactions
  - Hands are important tools for interactions and communications
  - Hand sensing from depth camera and optical camera
  - If we cannot solve a simple problem, try a complex one

- Graphics is more than rendering
  - Graphics synthesised data play important role for AI
  - We want creations that look both real and smart

Thank you!


Ge Liuhao


Jingjing Meng


Nadia Thalmann


Daniel Thalmann


Jianfei Cai


Ying He


Xiao Yang


Yu Gang


Liang Hui


Ren Zhou


Yujun Cai