3D Hand Pose Estimation from RGB Image

Eng Hock Lee
Chih-Tien Kuo

Motivation

Why Tracking Hand?

- Effective and general purpose tool
- High DoF control for human-computer interaction

Application Examples:

Virtual Reality

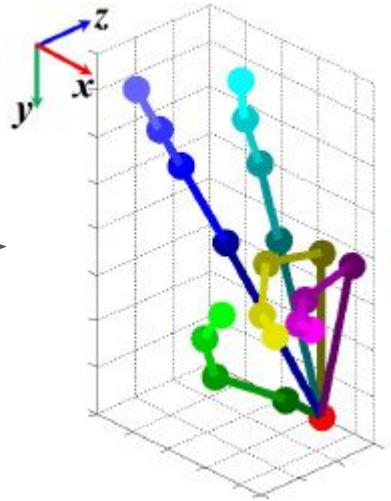
Augmented Reality

Problem Statement

- The goal of our project is to estimate 3D hand pose from RGB image

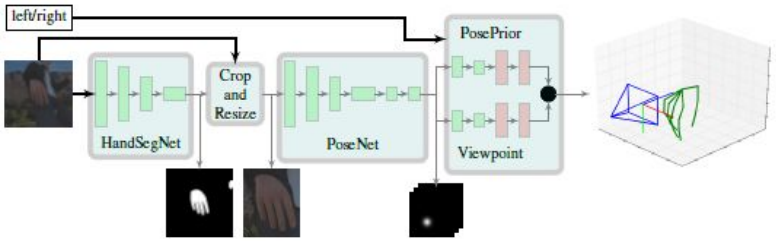
Input RGB Image

Output 3D Hand Pose

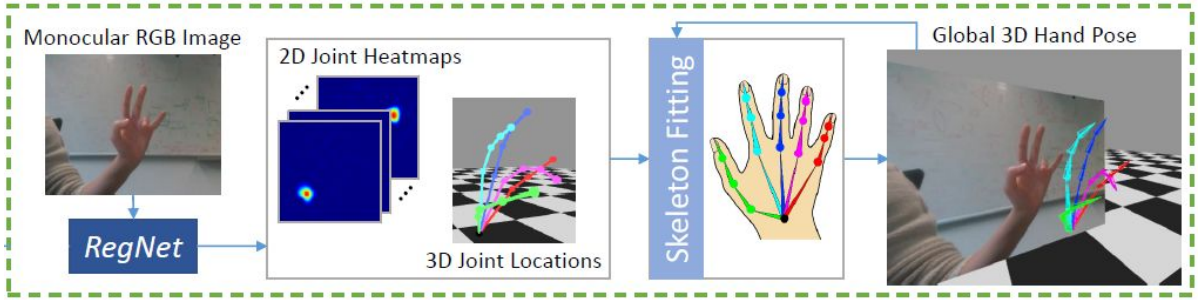


Related Works

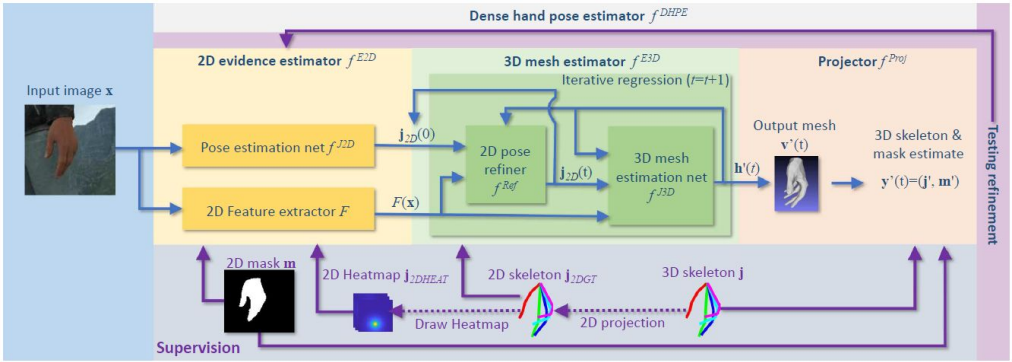
[1] Zimmermann, *et al*.



[2] Mueller, *et al*.



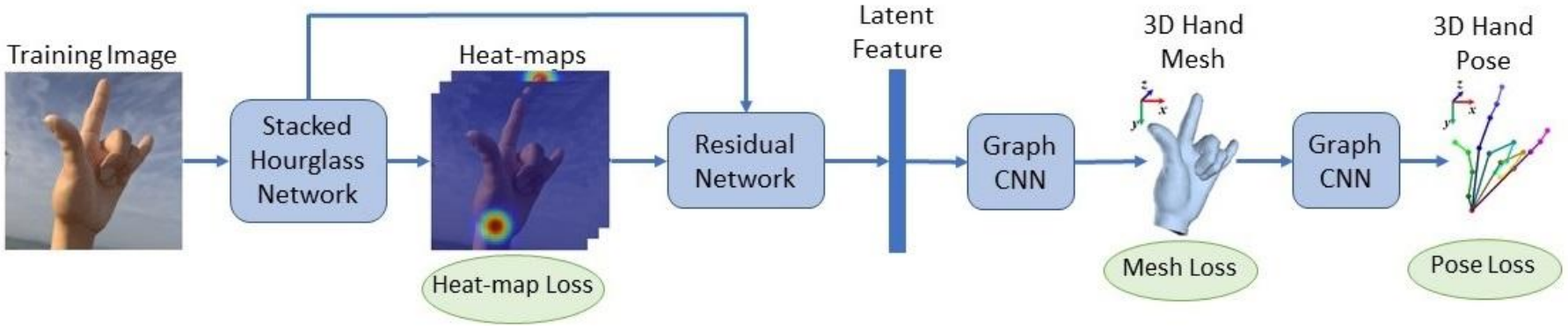
[3] Baek, *et al*.



[1] Zimmermann, Christian, and Thomas Brox. “Learning to estimate 3d hand pose from single rgb images.” Proceedings of the IEEE International Conference on Computer Vision. 2017.
[2] Mueller, Franziska, et al. “Real-time hand tracking under occlusion from an egocentric rgb-d sensor.” Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017.
[3] Baek, Seungryul, et al. “Kinematic-layout-aware random forests for depth-based action recognition.” arXiv preprint arXiv:1607.06972 (2016).

Baseline Method

- Following method from Ge, *et al*. [4]
- Using Stereo Hand Pose Benchmark (STB) Dataset for model training

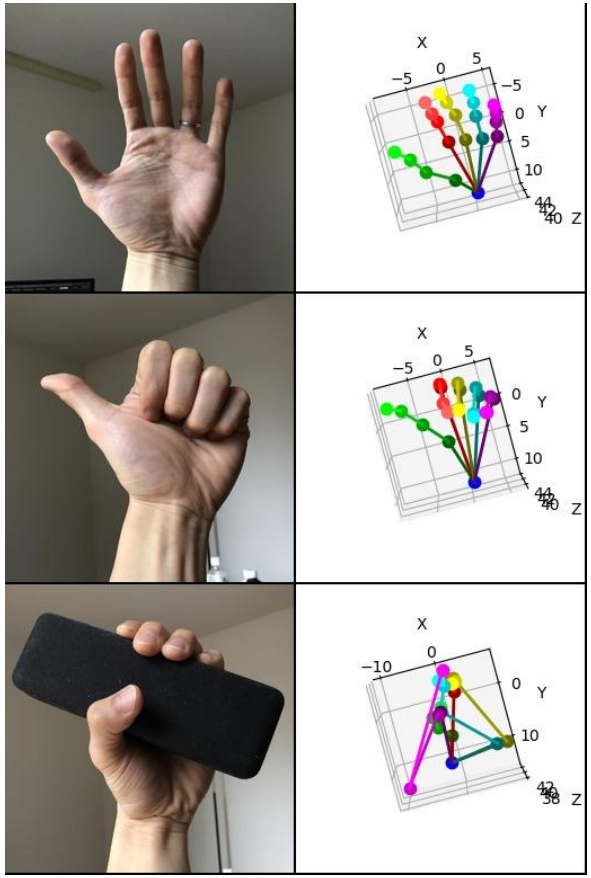


[4] Ge, Liuhao, et al. “3d hand shape and pose estimation from a single rgb image.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2019.

Baseline Method Result

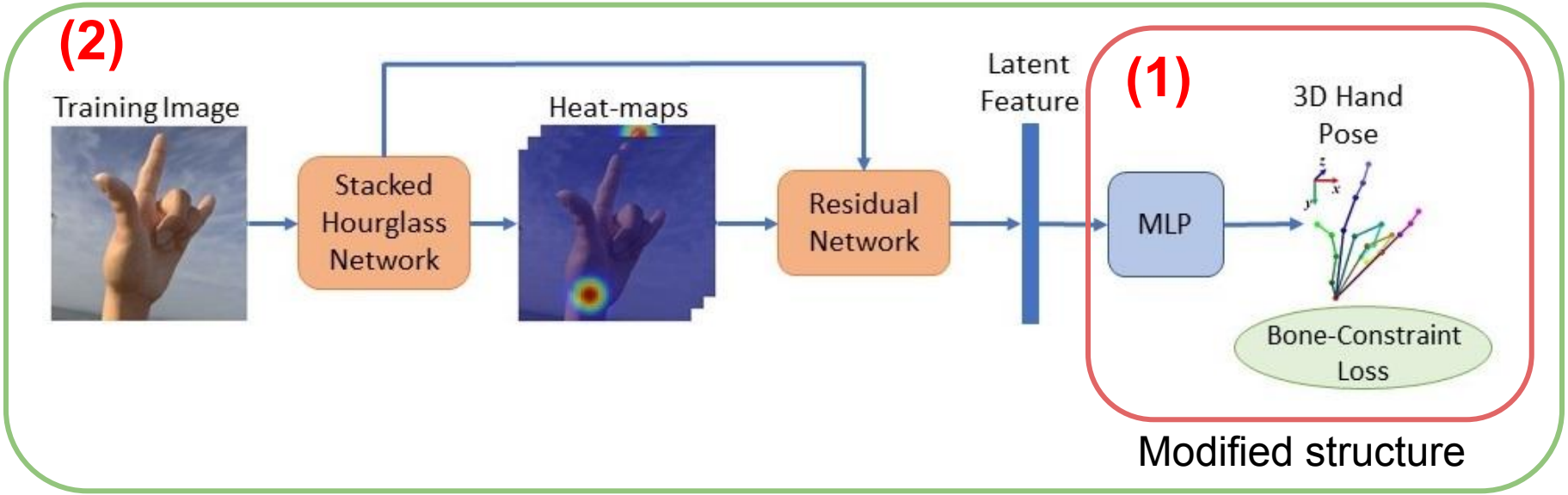
- Qualitative Result

Issue of Occlusion →



Proposed Method

- Goals of our proposed method:
  (1) Train the modified structure (MLP) with the proposed loss function
  (2) Resolve the issue of occlusion by using a new dataset



Model training using new dataset

Proposed Method

- Original Loss

$$L_{pose} = \sum_{j=1}^{21} \left\| \boldsymbol{\phi}_j - \hat{\boldsymbol{\phi}}_j \right\|_2^2$$

- Bone-Constraint Loss

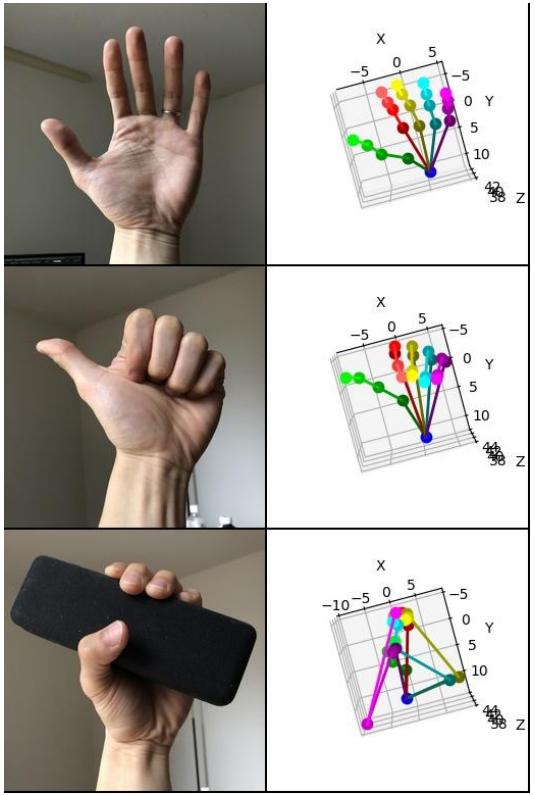
$$L_{len} = \sum_{i,j} \left| ||\boldsymbol{b}_{i,j}||_2 - ||\hat{\boldsymbol{b}}_{i,j}||_2 \right|$$

$$L_{dir} = \sum_{i,j} \left\| \frac{\boldsymbol{b}_{i,j}}{||\boldsymbol{b}_{i,j}||_2} - \frac{\hat{\boldsymbol{b}}_{i,j}}{||\hat{\boldsymbol{b}}_{i,j}||_2} \right\|$$

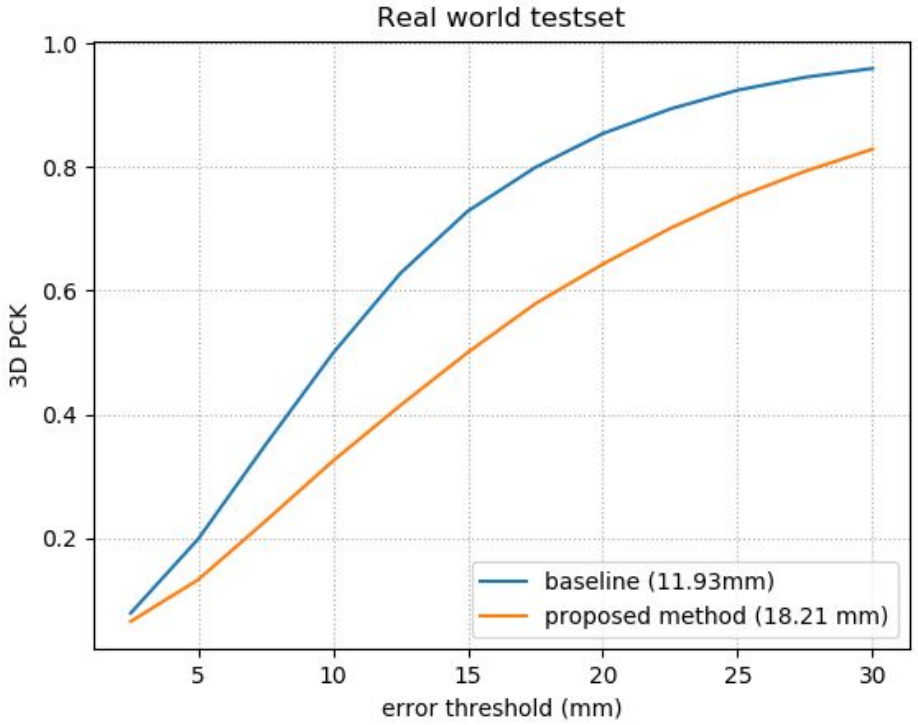
$$L = \lambda_{pose} L_{pose} + \lambda_{len} L_{len} + \lambda_{dir} L_{dir}$$

Preliminary Result

- Qualitative Result



- Quantitative Result



Next Step...

- Train the model with our proposed loss
- Train with GANerated Hands Dataset
  - Synthetic dataset generated using GAN
  - Interaction with objects



Thank you