

– Supplementary Material –

Zimeng Zhao    Xi Zhao    Yangang Wang\*

Southeast University, China

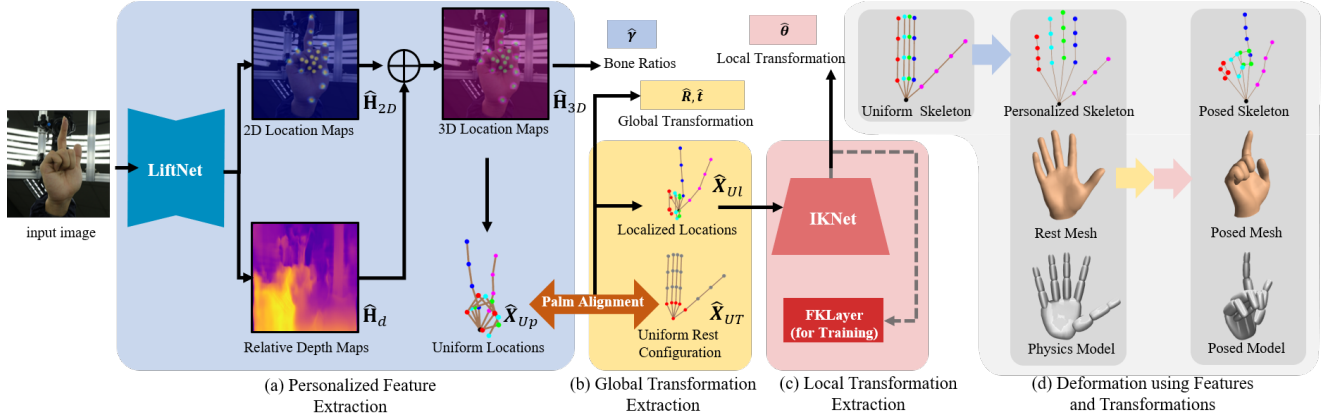


Figure 1. **Pipeline of hand physical pose estimation.** (a) From a single hand image, LiftNet estimates the corresponding joint 2D location maps and relative depth maps. The 3D location of each hand joint can be then estimated by those two branches. (b) We use [9] to optimize the global transformation used to align the palm joints of those two uniform joint locations; (c) IKNet learns to estimate local transformation from the localized locations. FKLayr is adopted to measure cycle consistency from angular space to Euclidean space only in the training phase; (d) Those estimated parameters can be used to deform the skeleton, mesh, and physical model.

## A. Overview

In this supplementary document, we first provide details about the hand pose estimator used in the main paper (Sec. B). Then, the limitations of the proposed method are discussed in Sec. C. They were not included in the paper due to the page limit.

## B. Physical Pose Estimator

### B.1. Formulation

As shown in Fig. 2, our articulated hand built in [1] follows the topology  $\mathcal{H}$ , rest configuration  $\mathbf{X}_T \in \mathbb{R}^{21 \times 3}$  and in-hand DoF axis  $\{\tilde{\alpha}_j\}_{j=1}^{21}$  similar to a human hand [2]. Since it is tedious for the physics engine to simulate various personalized meshes, we design a disentanglement to tackle personalization described by bone ratios vector  $\gamma \in \mathbb{R}^{20}$ . It

is defined to record the distances among adjacent joints and characterizes each bone length (unit: mm) of  $\mathbf{X}_p$  with the hand topology  $\mathcal{H}$ .

Our physical pose estimator aims to regress the corresponding physical pose state  $\theta \in \mathbb{R}^{21}$  from a single RGB image. Its entire pipeline is shown in Fig. 1. It first predicts personalized 3D joints location  $\mathbf{X}_p$  from given image, and then disentangles our canonical  $\theta$  from bone ratios  $\gamma$  and global transformations  $(\mathbf{R}, \mathbf{t})$  by steps.

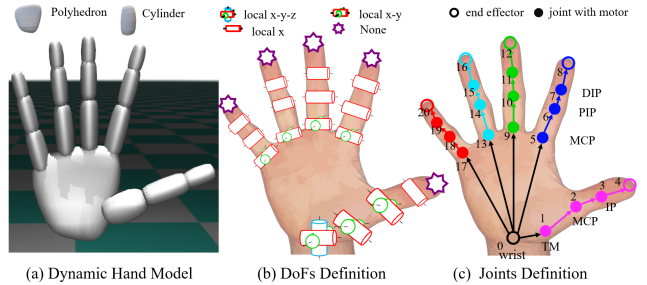


Figure 2. **Physical hand model.** Topology, DoFs and Joints of our hand model. The palm-related joint indices are: 0, 1, 5, 9, 13, 17.

\*Corresponding author. E-mail: yangangwang@seu.edu.cn. All the authors from Southeast University are affiliated with the Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing, China.

## B.2. Pipeline

**Personalized Feature.** The bone ratios vector  $\hat{\gamma} \in \mathbb{R}^{20}$  is extracted from the input image in the first step. LiftNet, a CNN-based architecture is designed to predict the joint 2D location maps and relative depth maps in a heatmap regression. Our loss design and map ground truth construction are similar to [4, 11].

Through maximum likelihood estimation on those maps, the corresponding joint information is combined to get personalized hand joints location  $\hat{\mathbf{X}}_p$ . After that,  $\hat{\mathbf{X}}_p$  is transformed to the uniform location  $\hat{\mathbf{X}}_{Up}$  by keeping the root joint fixed and storing each bone length to  $\hat{\gamma}$ . As a result, each bone counterpart length in  $\hat{\mathbf{X}}_{Up}$  is depersonalized to be  $l = 1\text{mm}$ . The rest configuration  $\mathbf{X}_T$  transforms to be  $\mathbf{X}_{UT}$  in the same way after being divided by its own bone ratios  $\gamma_T$ :

$$\hat{\mathbf{X}}_{Up} = \frac{\hat{\mathbf{X}}_p}{\hat{\gamma}}; \hat{\mathbf{X}}_{UT} = \frac{\hat{\mathbf{X}}_T}{\hat{\gamma}_T} \quad (1)$$

**Global Transformation.** Global transformation parameters  $(\mathbf{R}, \mathbf{t})$  from  $\mathbf{X}_{UT}$  to  $\hat{\mathbf{X}}_{Up}$  can be determined by the point pattern alignment [9]. Scale parameters in [9] can be fixed as 1.0 because they both have uniform configuration; This linear optimization problem have an exact solution because the spatial relationship of the palm-related joints (joint indices 0, 1, 5, 9, 13, 17 in Fig. 2(c).) is rigid. After that, we get the joint location  $\hat{\mathbf{X}}_{Ul}$  which only contains local transformation of different fingers:

$$\hat{\mathbf{X}}_{Ul} = \mathbf{R}^{-1} \cdot (\hat{\mathbf{X}}_{Up} - \mathbf{t}) \quad (2)$$

**Local Transformation.** The function of our IKNet (inverse kinematic network)  $\mathcal{G}_{IK}$  is to estimate physical in-hand pose  $\hat{\theta}$  from  $\hat{\mathbf{X}}_{Ul}$  produced in the previous step. Despite having the same name, the IKNet of this work is quite different from [16] in pose representation and input data configuration. Since Euler’s angle is difficult to regress [15], we also design a FKLayer (forward kinematic layer)  $f_{FK}$  as IKNet’s loopback inspection:

$$\mathbf{X}_{Ul} = f_{FK}(\theta) = f_{FK}(\theta; \{\vec{\alpha}_j\}_{j=1}^{21}, \mathcal{H}) \quad (3)$$

$f_{FK}$  is specific to our fixed topology  $\mathcal{H}$  and local DOF axis  $\{\vec{\alpha}_j\}_{j=1}^{21}$ , has no learnable parameters, and takes  $\theta$  as the input to output the deformed joint location  $\mathbf{X}_{Ul}$ . This makes the training of IKNet also have a self-supervised term:

$$\mathcal{L}_{IK} = \mathcal{L}_1(\mathcal{G}_{IK}(\hat{\mathbf{X}}_{Ul}), \theta) + \mathcal{L}_2(f_{FK}(\mathcal{G}_{IK}(\hat{\mathbf{X}}_{Ul})), \hat{\mathbf{X}}_{Ul}) \quad (4)$$

The first term  $\mathcal{L}_1$  performs the supervision between the prediction of  $\mathcal{G}_{IK}$  and the ground truth  $\theta$  in angular space. And the second term takes the prediction  $\hat{\theta}$  back to the Euclidean space to ensure the cycle consistency to the original input  $\hat{\mathbf{X}}_{Ul}$  of IKNet.

**Training and Inference.** According to this three-step design, the pose estimator can make full use of abundant multi-modal datasets: The RGB dataset with 2D annotations [8, 10, 5] can train the backbone of LiftNet; The depth [12] and synthetic datasets [8, 17, 7, 6] can be used for IKNet training; The real RGB dataset [13, 18, 3, 14] with 3D annotations can be used for the fine-tuning of the whole pipeline.

Our physical pose estimator assists TravelNet in both training and inference. We use it to construct part of our pose state archive for TravelNet training. In the inference of TravelNet, the physical pose estimator separates the bone ratio from images before TravelNet and reloads them again on TravelNet outputs to get a personalized motion.

## C. Limitations

Our approach has several limits, but it also opens up new possibilities for future research.

**Physical Modeling.** Our hand model is an articulated rigid body in a physics engine. It keeps the same skeleton and approximate shape as the MANO original template. These configurations are fixed during our motion solving because the physics engine is not supported to adjust these rigid parts dynamically. Theoretically, the TravelNet avoids the penetrations caused by the invalid pose states without shape variation. As for non-rigid part deformation, we directly use linear blend skinning defined on the MANO because of the consistency in the skeleton definition of the two models. A promising future direction is to describe the hand model as a soft body based on the finite element method. The motion analysis based on this model will have higher flexibility.

**Learning Paradigm.** In our pipeline, the pose estimator and the TravelNet have the obvious division of labor: the former is responsible for image feature extracting and the TravelNet for motion learning. Under this division, both kinds of training data are sufficient. However, the image quality is also helpful for noise filtering in the decoder. Improving the current learning paradigm according to this intuition may make the reconstructed motion more robust.

## References

- [1] Mujoco physics engine. <http://www.mujoco.org>. 1
- [2] F Dincer and G Samut. Hand function: A practical guide to assessment. 2014. 1
- [3] Francisco Gomez-Donoso, Sergio Orts-Escolano, and Miguel Cazorla. Large-scale multiview 3d hand pose dataset. *Image and Vision Computing*, 81:25–33, 2019. 2
- [4] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 118–134, 2018. 2
- [5] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and

- Cewu Lu. Pastanet: Toward human activity knowledge engine. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 382–391, 2020. 2
- [6] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Gnerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–59, 2018. 2
- [7] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1284–1293, 2017. 2
- [8] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017. 2
- [9] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (4):376–380, 1991. 1, 2
- [10] Yangang Wang, Cong Peng, and Yebin Liu. Mask-pose cascaded cnn for 2d hand pose estimation from single color image. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(11):3258–3268, 2018. 2
- [11] Yangang Wang, Baowen Zhang, and Cong Peng. Srhand-net: Real-time 2d hand pose estimation with simultaneous region localization. *IEEE Transactions on Image Processing*, 29:2977–2986, 2019. 2
- [12] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim. Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4866–4874, 2017. 2
- [13] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. 3d hand pose tracking and estimation using stereo matching. *arXiv preprint arXiv:1610.07214*, 2016. 2
- [14] Zhengyi Zhao, Tianyao Wang, Siyu Xia, and Yangang Wang. Hand-3d-studio: A new multi-view system for 3d hand reconstruction. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2478–2482. IEEE, 2020. 2
- [15] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 2
- [16] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5346–5355, 2020. 2
- [17] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017. 2
- [18] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 813–822, 2019. 2