Toward 3D Hand Pose Estimation

Ruouyu He, Sherry Huang, Cheng Lu, Yanxun Li, Yijun Cheng
University of California, Los Angeles
{rhe9527, sherryhxr99, lucheng9, liyanxun1128, chengyj18}@ucla.edu

Abstract

Recent advancements in computer vision and deep learning have demonstrated many ways of achieving the goal of estimating hand pose in a fine-grained way. In this project, we focus on estimating 3D hand pose from a single RGB image. In particular, we propose to integrate an existing 3D hand pose estimation method with different loss functions and models to improve accuracy and efficiency. Additionally, our proposed method alleviates the image occlusion problem with hand pose estimation by utilizing a multi-view hands dataset that targets this issue. We demonstrated that our proposed method provides a solid estimation of hand pose while maintaining a relatively light-weighted model architecture.

1. Introduction

Hand pose estimation has been an ongoing research topics in computer vision and Human-Computer Interaction (HCI). With rising importance of technology in our daily life, hand pose estimation becomes more and more crucial in many applications, ranging from medical operations to virtual reality games. Hand pose estimation remains as a challenging topic due to occlusions, appearance diversity and deformable surfaces. Image occlusion is especially a fundamental challenge in hand pose estimation as an image would only contain limit amount of information. To overcome these challenges, numerous solutions have been proposed. In the past, hand motion capture systems performed by tracking retro-reflective markers [H. Woltring, 1973] and using surface meshes [Heap et al, 1996] have shown their ability to capture the motion with high accuracy. These traditional methods often require extensive human efforts. In recent years, marker-less deformable models, such as combining models with a hand key-point detector that can reconstruct 3D hand joints in multi-view systems [Joo et al, 2018], have demonstrated great advances in 3D hand estimation. However, these methods still often require specialized equipment and expensive models.

In search of a more efficient method, numerous works have proposed methods that aims to estimate hand pose from a single image. In particular, many works have demonstrated great performance in 3D hand pose estimation with depth cameras [Ge et al., 2018]. However, depth cameras are still not readily accessible to most people. Therefore, some recent works have addressed 3D hand pose estimation with RGB cameras. Among these works, Ge et al [Ge et al., 2019] proposed a graph convolutional neural network (graph CNN) based method to reconstruct a full 3D mesh of hand surface that contains richer information of both 3D hand shape and pose. While this method presents great performance without any specialized equipment, it still requires reconstructing 3D hand mesh and a heavy graph CNN. More importantly, this method suffers from image occlusion problem due to the limitation of their dataset.

In this project, we propose to improve the accuracy and efficiency of the graph CNN based method [Ge et al., 2019] by:

1. Replace the loss functions with cosine similarity loss and active shape model assisted loss.

2. Eliminate the reconstruction of 3D Mesh and replace the graph CNN with a more light-weighted multi-layered perceptron.

Additionally, we replace their choice of dataset with FreiHAND v2 [Zimmermann et al, 2019]–a wellstructured hand pose image dataset that contains extensive RGB images and masked images, including images with object-occlusion and self-occlusion. By utilizing this dataset, we hope to leverage the difficulties in hand pose estimation of occluded hands.

2. Related Works

In some previous methods, most models fit the deformable hand model to the input depth map, and after some optimization operations, such as iterative optimization, the estimation of the shape and posture of the three-dimensional hand can be performed. However, in the study of the estimation for 3D hand pose in-depth Images, it may be more effective to combine CNNs with depth images and try to recover the hand shape using the LSB model. But, due to the nature of CNNs training in an end-to-end manner (which includes the loss of pose and shape). This simple model limits the quality of hand mesh recovery and thus limits the performance of estimation. So, some research and work is still needed to be done in the future to solve this problem.

On the other hand, in the research of 3D hand pose estimation from RGB images, the early traditional method is to use the RGB sequence. Gradually methods are being explored to optimize and improve research in this area. For example, hand pose can be predicted by minimizing an objective function; Or improving the prediction results through multi-view RGB images and depth data. At present, many deep learning methods are used in hand pose estimation of RGB images. However, few similar studies have made headway in 3D Hand pose estimation. Many methods have been proposed but there are still many problems, such as the lack of flexibility to predict the changeable 3D hand poses. Therefore, relevant research in the future may also focus on optimizing the method of 3d hand pose prediction through RGB images.

In addition to hand pose estimation using machine learning and other methods, people also hope to try some methods that can predict hand pose and achieve 3D human body shape estimation. Many methods have been done to study a model to achieve body pose recognition and estimation. For example, some recent related works use an SMPL model-based approach. [M. Loper et al, 2015] In detail, it is worth studying to use the CNNs algorithm for regression analysis and prediction of some parameters in SMPL. Other methods such as trying to identify some 2D key points by fitting the SMPL model are also feasible. However, these methods are based on weak supervision of training, where depth maps are not used as weak 3D supervision when training on real-world datasets without 3D grids or 3D postures. Then, Graph CNNs is proposed, which is a method to evaluate 3D mesh points (vertices). In this way, nonlinear hand poses and shapes can be known. Furthermore, this approach has the advantage of being able to leverage vertices' interrelationships in the mesh topology.

3. Methods

3.1. Baseline Method

Instead of replicating Ge's 3D hand pose estimation method as our baseline method, we adapted Kuo and Lee's proposed method [Lee and Kuo]. Kuo and Lee uses a similar overall idea as Ge et al; unlike Ge's proposed method, Kuo and Lee used FreiHand dataset [Zimmermann et al, 2019] and constructed 3D hand pose directly without constructing 3D hand mesh first. They chose such approach as Ge et al demonstrated that replacing their proposed graph CNNs with an MLP network would achieve similar results, which inspires us to use similar practices. Thus, we decided to reproduce their work rather than Ge's proposed method for a fairer comparison. Figure 1 demonstrates their proposed model that we would be using as our baseline method. First, a single input RGB image is passed through a two stacked hourglass network that allows for repeated bottom-up, top-down inference [Newell et al., 2018]. The hourglass network would extract 2D heat-maps of 21 different hand joint positions that would be passed to a residual network to output an array of latent feature vector. The heat map loss is formulated as:

$$L_{hm} = \sum_{j=1}^{21} \|h_j - \hat{h_j}\|_2^2$$

After that, the latent feature vector is fed to a multi-layered perceptron network to estimate 3D hand pose with bone-constraint loss, a biologically inspired loss function. Specifically, the bone-constrained loss function can be formulated as:

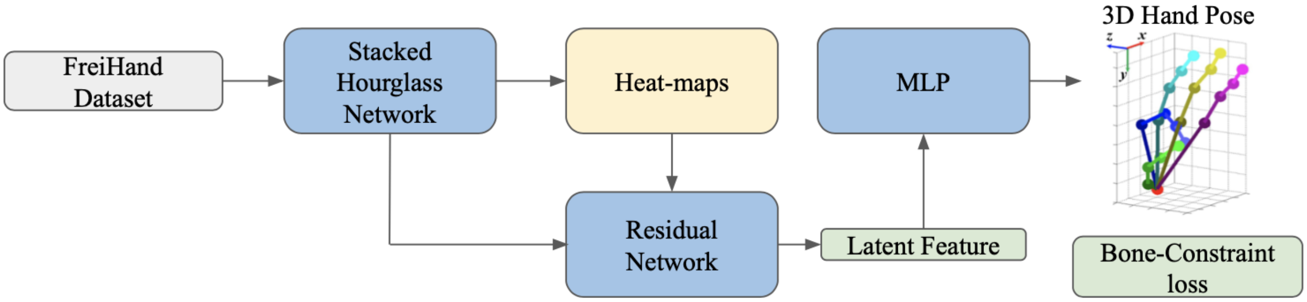

Figure 1. Basline structure

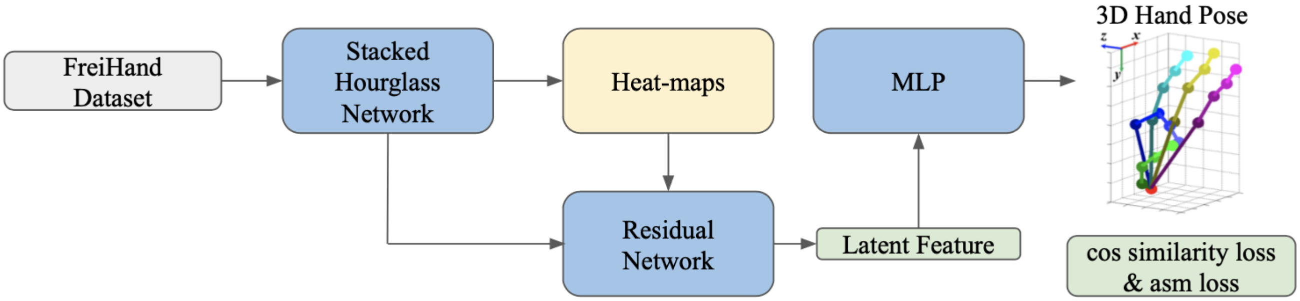

Figure 2. Model structure

$$L_{bone} = \lambda_{hm} L_{hm} + \lambda_{pose} L_{pose} + \lambda_{len} L_{len} + \lambda_{dir} L_{dir}$$

where λ_{hm}, λ_{pose}, λ_{len}, λ_{dir} are the hyperparameters and L_{pose}, L_{len}, L_{dir} are formulated as followings:

$$L_{pose} = \sum_{j=1}^{21} \| \phi_j - \hat{\phi_j} \|_2^2$$

where ϕ_j is defined as the ground truth joint locations and j and $\hat{\phi_j}$ is the prediction.

$$L_{len} = \sum_{i,j} | \, ||b_{i,j}||_2 - ||\hat{b_{i,j}}||_2 \, |$$

where $b_{i,j}$ is defined as the ground truth bone vector between joint and j and $\hat{b_{i,j}}$ is the prediction.

$$L_{dir} = \sum_{i,j} \| \frac{b_{i,j}}{\| b_{i,j} \|_2} - \frac{\hat{b_{i,j}}}{\| \hat{b_{i,j}} \|_2} \|$$

3.2. Loss Functions

3.2.1 Cosine Similarity Loss

Inspired by the baseline method [Lee and Kuo], we propose to use cosine similarity to measure the similarity between the bones vectors:

$$cos\_similarity(x_1, x_2) = \frac{x_1 \cdot x_2}{max(||x_1|| \cdot ||x_2||, \varepsilon)}$$

where $\varepsilon = 1e - 8$. Then the cosine similarity loss is defined as:

$$L_{cos} = 1 - mean(cos\_similarity(b_{pred}, b_{true}))$$

where b_{pred} is our estimation of the bone vectors and b_{true} is the ground truth bone vectors calculated from the coordinates of the joints. The total loss function is defined as followed:

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

where $\lambda_{hm}, \lambda_{pose}, \lambda_{len}, \lambda_{dir}, \lambda_{cos}$ are hyper-parameters that can be tuned to test the effectiveness of cosine similarity loss.

3.2.2 Active Shape Model Assisted Loss

Fard et al proposed to use the active shape model to guide deep neural networks to do 2d face alignment and pose estimation [Fard et al, 2021]. We extend this idea to the task of 3d hand pose estimation. Let Y_{true} and Y_{pred} be $batch\_size \times num\_joints \times 3$ matrices representing a batch of ground truth and estimated joints positions. After flattening Y_{true} and Y_{pred} to be $batch\_size \times (3num\_joints)$, Principal Component Analysis (PCA) is performed to simplify the problem and learn shape components. Let $V = [e_1, e_2, ..., e_k]$ be the matrix containing the top k principal components, which are the top k eigenvectors of the covariance matrix of the samples. Compute the matrix b as follows:

$$b = (Y_{true} - \overline{Y_{true}})V$$

where $\tilde{Y_{true}}$ is the sample mean of the ground truth label. Using the top k eigenvalues λ_i as threshold, we clamp value of the corresponding entry of b to $[-|3\sqrt{\lambda_i}|, |3\sqrt{\lambda_i}|]$. The resulting matrix is denoted as $\tilde{b}$. This ensures that the generated Active Shape Model is relatively close to the ground truth. Finally, we compute the ASM ground truth:

$$Y_{asm} = \overline{Y_{true}} + \tilde{b}V.T$$

Then the Active Shape Model assisted loss is defined as:

$$L_{asm} = sum((Y_{pred} - Y_{asm})^2)$$

The final loss is defined as:

$$L = \lambda_{hm}L_{hm} + \lambda_{pose}L_{pose} + \lambda_{asm}L_{asm}$$

The goal is to use ASM to guide the network to first learn a smoothed distribution of the joint locations. We want the network to first focus the generated ASM ground truth in the early stage of the training and then focus on the actual ground truth joints locations in the late stage the training. Therefore, we will give the ASM loss bigger weights at first and then gradually decrease λ_{asm} during training.

3.3. Multi-layered Perceptron

Inspired by Ge et al's evaluation [Ge et al., 2019] and Kuo and Lee's work [Lee and Kuo], we also replaced the graph CNNs in the original work with a simplified multiple-layered preceptron (MLP). The model architecture is shown as in figure 3. As mentioned in the previous section, Ge et al demonstrated that replacing the graph CNNs with a simple MLP does not hurt the performance in their tasks. Since we no longer need to reconstruct 3D mesh, we save great efforts in the training process.

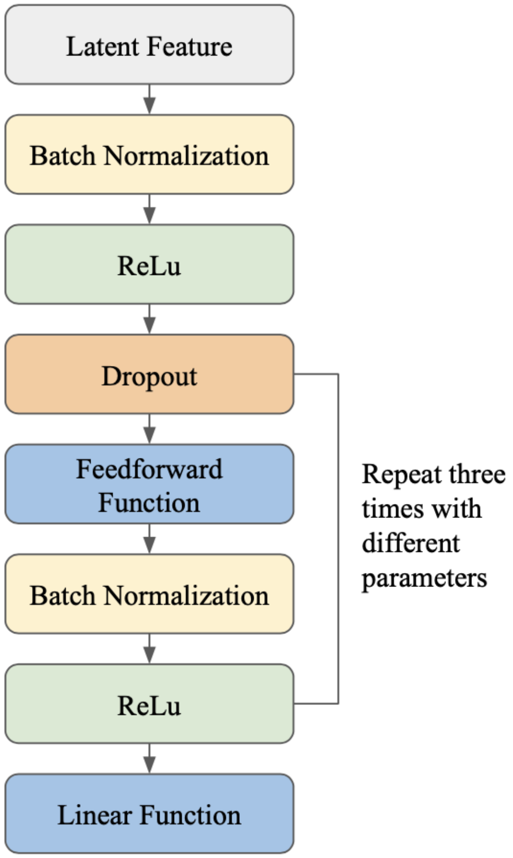

Figure 3. MLP structure

Through empirical experiments, we realized that our proposed methods might converge too quickly while failing to find correct optimal points. Therefore, we decided to use a deeper MLP with dropout and batch normalization at each layer before the last feed forward layer to help with this issue.

4. Experiments

We trained and evaluated the proposed cosine similarity loss model and ASM assisted loss model (M-asm) on the FreiHAND dataset. We tested two version of the cosine similarity loss model:

1. M-cos-1:
   $\lambda_{hm}L_{hm} = 0.1, \lambda_{pose}L_{pose} = 1, \lambda_{len}L_{len} = 0.01, \lambda_{dir}L_{dir} = 0.1, \lambda_{cos}L_{cos} = 10$

2. M-cos-2:
   $\lambda_{hm}L_{hm} = 0.1, \lambda_{pose}L_{pose} = 1, \lambda_{len}L_{len} = 0, \lambda_{dir}L_{dir} = 0, \lambda_{cos}L_{cos} = 10$

We trained M-cos-1, M-cos-2, M-asm and the baseline model on 3200 samples of the FreiHAND dataset for 100 epochs. Since the FreiHAND dataset did not provide 3d annotation for the evaluation set, we will take 200 samples from the training set as the test set (those 200 samples will not be used during training). We will use the 3d pose error as metric, which is defined as the Euclidean distance between the estimated joint positions and the ground truth joint position.

5. Results and Discussion

The 3d pose errors of our proposed models and based line are shown in Table 1. Some examples are shown in Figure 4 and Figure 5. As shown in the table, all of the above three models outperforms the baseline model. First let us focus on the two cosine similarity loss models. M-cos-2 outperforms both the baseline and M-cos-1 by achieving the 3d pose error of 18.31 mm. Compared to M-cos-1, M-cos-2 discards the bone length loss $\lambda_{len}L_{len}$ and the bone direction loss $\lambda_{dir}L_{dir}$, thus increasing the weight of cosine similarity loss. This may suggests that, in this case, compared to bone length loss and bone direction loss, cosine similarity is a better measurement for measuring the similarity between two bone vectors.

M-cos-1	M-cos-2	M-asm	baseline
26.00	18.31	15.54	37.53

Table 1. 3d pose error in mm

M-asm has a significantly lower 3d pose error than the baseline model. Moreover, M-asm also converge much faster in our experiment. While the baseline model takes nearly 100 epochs to converge, M-asm converges around 50 epochs. This suggests that the ASM assisted loss is working as we expected: first guide the model to learn a smoothed distribution of the samples, then the decreasing ASM weight allows the model to improve itself based on the learned distribution and the ground truth.

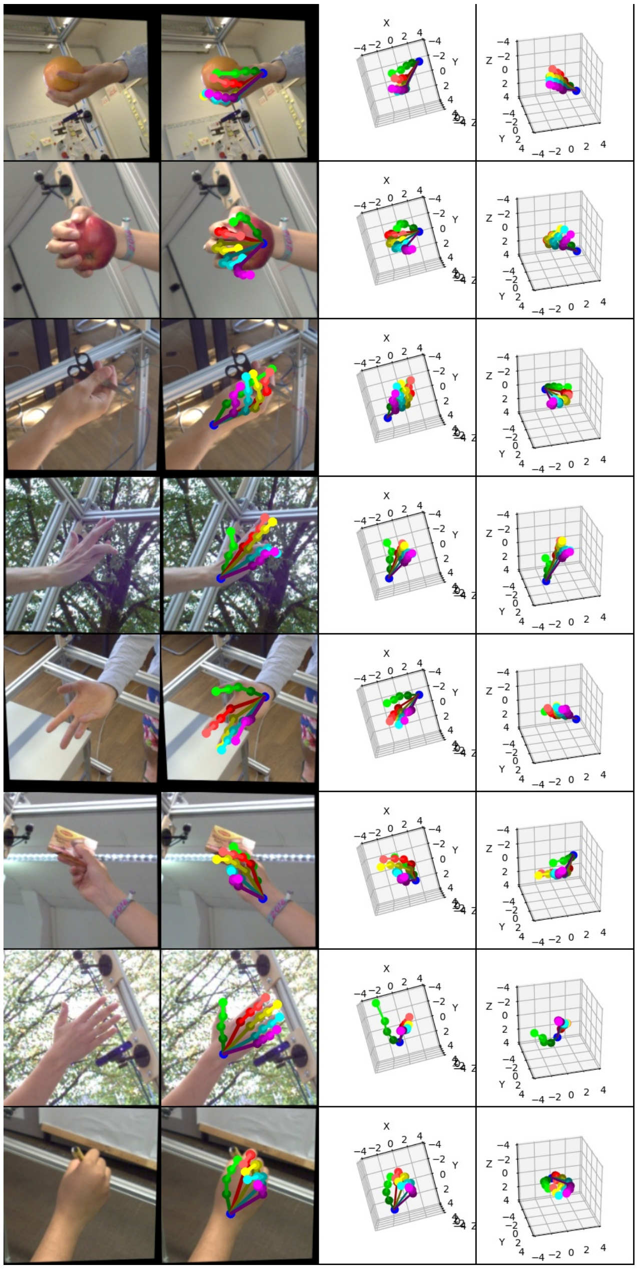

Figure 4. Predictions Made by M-asm

6. Conclusion

In conclusion, we explored 3D hand pose estimation from a single RGB image and present solid hand pose estimation results. Targeting at the challenge of image occlusion, we introduced FreiHAND dataset.

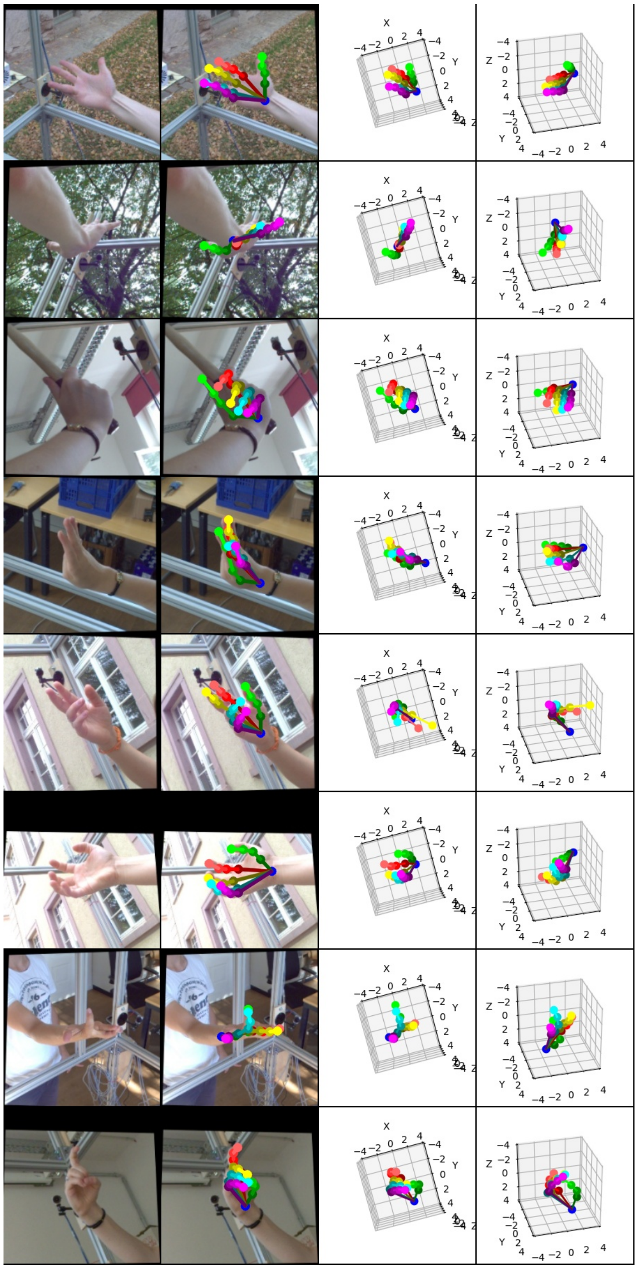

Figure 5. Predictions Made by M-asm

As illustrated in figure 5, our proposed method works well under occluded angels as well. In order to improve the efficiency of previous works, we propose to replace graph CNNs with a simplified MLP. To improve the accuracy of the original work, we propose to use cosine similarity loss and active shape model assisted loss instead of bone constraint loss. As discussed in the previous sections, our proposed methods successfully achieve both goals.

Throughout this project, one major limitation we faced is the limited computation power. While there are over 30k samples in FreiHAND dataset, we could only utilize about 10% of the whole dataset due to limited memory and computation power we have. Moreover, this limitation also made it hard for us to explore new possibilities and grid search for best hyperparameters.

References

[Ge et al., 2019] Liuhao Ge and Zhou Ren and Yuncheng Li and Zehao Xue and Yingying Wang and Jianfei Cai and Junsong Yuan 2019 3D Hand Shape and Pose Estimation from a Single RGB Image In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1, 4

[Ge et al., 2018] Liuhao Ge, Zhou Ren, and Junsong Yuan. 2018 Point-to-point regression pointnet for 3d hand pose estimation In *ECCV*, 2018. 1

[Newell et al., 2018] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016 Stacked hourglass networks for human pose estimation In *ECCV*, 2016. 2

[H. Woltring, 1973] H. Woltring 1973 New possibilities for human motion studies by real-time light spot position measurement In *Biotelemetry*, 1973. 1

[Heap et al, 1996] Heap, Tony and Hogg, David 1996 Towards 3D hand tracking using a deformable model In *Conference on Automatic Face and Gesture Recognition*, 1996. 1

[Joo et al, 2018] Hanbyul Joo and Tomas Simon and Yaser Sheikh 2018 Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies In *1801.01615 arXiv*, 2018. 1

[Lee and Kuo] Eng Hock Lee and Chih-Tien Kuo https://github.com/enghock1/Real-Time-2D-and-3D-Hand-Pose-Estimation 2, 3, 4

[Zimmermann et al, 2019] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russel, Max Argus and Thomas Brox 2019 FreiHAND: A Dataset for Markerless Capture of Hand Pose and Shape from Single RGB Images In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[M. Loper et al, 2015] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. 2015 FreiHAND: SMPL: A skinned multi-person linear model. In *ACM Transactions on Graphics (TOG)*, 2015. 2

[Fard et al, 2021] Fard, Ali Pourramezan, Hojjat Abdollahi, and Mohammad Mahoor. 2021 ASM-Net: A Lightweight Deep Neural Network for Face Alignment and Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2021. 4