Hand Image Understanding via Deep Multi-Task Learning Supplementary Materials

Xiong Zhang[1]*, Hongsheng Huang[2], Jianchao Tan[3], Hongmin Xu[4],
Cheng Yang[1], Guozhu Peng[1], Lei Wang[1], Ji Liu[3]
[1]YY Live, Baidu Inc., [2]Joyy Inc., [3]AI Platform, Kwai Inc., [4]OPPO Inc.,

August 3, 2021

1 Implementation Details

In this section, we mainly introduce the technical details of the **HIU-DMTL** framework.

1.1 Backbone

Stem Module. The stem module consists of two 7×7 convolutional layers with stride 2, and the channels are set to 64 and 128, respectively.

Encoder. We employ the main-body of ResNet-50 [5] to implement the encoder. Specifically, the beginning $\mathrm{conv1}$ together with the prediction head are removed, while the remaining $\mathrm{conv2\_x}$, $\mathrm{conv3\_x}$, $\mathrm{conv4\_x}$, and $\mathrm{conv5\_x}$ are adopted to build the encoder module, and the number of repetitions are 3,4,5, and 6, respectively.

Heat-Map Decoder. The heat-map decoder estimates the feature maps $\mathrm{hms\_ft} \in \mathbb{R}^{256\times64\times64}$ to capture semantic features to encode the 2D hand pose. Similar with [10], the decoder also estimates the $\mathrm{hms} \in \mathbb{R}^{21\times128\times128}$, based on the $\mathrm{hms\_ft}$, to represent the locations of 21 hand key points, and the hms is used for intermediate-supervision. Skip connections between the encoder and the heat-map decoder are also adopted to favor the learning procedure.

Mask Decoder. The target of the mask decoder is to estimate the feature maps $\mathrm{mask\_ft} \in \mathbb{R}^{256\times64\times64}$ to capture semantic features that encode the hand segmentation mask. Similar to the heat-map decoder, the mask decoder also estimates the hand segmentation $\mathrm{mask} \in \mathbb{R}^{1\times256\times256}$ based on the $\mathrm{mask\_ft}$, and the mask is used for intermediate-supervision. Skip connections between the encoder and the mask decoder are also adopted to favor the learning.

POF Decoder. The POF decoder aims to estimate the feature maps $\mathrm{pof\_ft} \in \mathbb{R}^{256\times64\times64}$ to capture semantic features that encode 3D POF encoding. Similar to the

*corresponding author, zhangxiong@yy.com

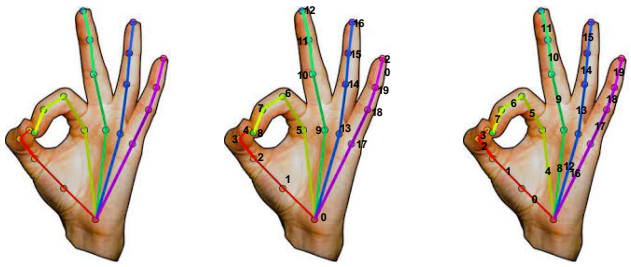


Figure 1: **Hand Hierarchy.** The figure presents a hand image sample, the corresponding 2D hand key points, and the bone skeleton hierarchy $\mathbb{S}$, respectively.

heat-map decoder, the POF decoder also estimates the $\mathrm{pof} \in \mathbb{R}^{20\times3\times128\times128}$ based on the $\mathrm{pof\_ft}$, and the pof is used for intermediate-supervision. Skip connections between the encoder and the POF decoder are also adopted to favor the learning procedure. As Figure 1 demonstrates one typical hand image, each hand image consists of 21 hand key points and 20 hand skeleton bones.

Task Attention Module. The task attention module (**TAM**) aims to bring together semantic features across individual tasks. Concretely, the TAM aggregates $\mathrm{hms\_ft}$, $\mathrm{mask\_ft}$, and $\mathrm{pof\_ft}$ to build the high-level task agnostic semantic features $\mathrm{tam\_ft} \in \mathbb{R}^{256\times64\times64}$. In practice, the TAM consists of three average pooling operations and certain point-wise convolutional layers.

Regressor Head. The purpose of the regressor head is to regress the hand shape $\beta \in \mathbb{R}^{10}$, hand pose $\theta \in \mathbb{R}^{15\times3}$, global rotation $\mathrm{R} \in \mathbb{R}^3$, and global translation $\mathrm{T} \in \mathbb{R}^3$. We shall point out a subtle but important detail here. In the blend skinning procedure of MANO [11], the correct way to obtain the 3D hand joints is rotating the rest hand joints $J(\beta)$ with the pose parameters θ. However, [18] firstly proposed a misleading way to obtain the 3D hand joints by linear blending the vertices in the hand mesh with blending weights $\mathcal{J}$. In fact, the blending weight $\mathcal{J}$ can only be used to estimate the hand joints from hand mesh in rest pose (see [7] for more technical details). *Since then, certain works adopt similar misleading strategy to infer 3D hand joints* [6, 8] *from the recovered hand mesh representation.* For instance, [8] proposes to regress of the 778 vertices that constitute the hand mesh, then obtains the hand joints by blending the vertices with $\mathcal{J}$. *We sincerely expect that such misleading usage will not happen since we point out it in this work.*

1.2 Training Objective

Achieving Self-Supervised Learning. To conduct the self-supervised learning (SSL), we employ two inherent constraints maintained among the reasonable predictions from each task. Specifically, 1) we use a differentiable renderer to obtain the re-projected hand mask that is also differentiable, and adopt the L_1 norm to penalize the misalignment between the re-projected hand mask and the estimated segmentation mask from the segmentation branch; 2) Equation 1 (see main paper) can be adopted to achieve a differentiable hand pose from the heat-maps. One may also obtain another hand pose

by projecting the 3D hand joints inferred from the hand mesh and employing the L_2 norm to make the two intermediate hand poses close to each other.

2 Experiment Details

In this section, we mainly introduce the configurations of all experiments.

2.1 Datasets

We mainly involve the CMU Panoptic Dataset (CMU) [12], the Rendered Hand dataset (RHD) [20], the Stereo Hand Pose Tracking Benchmark (STB) [17], the FreiHAND [21], the Dexter Object [13], and the HIU-Data.

CMU Panoptic Dataset (CMU). The CMU dataset [12] is an accurate large-scale dataset which contains hand images in various poses observed from multiple views in the Panoptic studio. The MPII+NZSL dataset is also contained in [12], and we *only exploit the MPII+NZSL part* .

Rendered Hand Dataset (RHD). The RHD dataset [20] is a synthetic dataset that contains $41,258$ training samples and $2,728$ testing samples. Each sample contains an RGB image, a depth image, a segmentation mask image, and both 2D/3D hand pose of the 21 standard key points.

Stereo Hand Pose Tracking Benchmark (STB). The STB [17] contains sequences with 6 different backgrounds. In this paper, the similar strategy [2, 18] is adopted to align the root joint of STB to make it consistent with the standard hand hierarchy (see Figure 1), and we adopt the guidance given by [20] to split STB into training and testing part.

Dexter Object Dataset. The Dexter Object [13] consists of 6 video sequences with 2 actors, which shows interactions of an actor's hand with a cuboid object from a third person view. For each sample, fingertip positions and cuboid corners are manually annotated for all sequence.

FreiHAND Dataset. The FreiHAND [21] contains $32,650$ training samples and $3,960$ testing images with MANO pose and shape parameters. In this paper, only the 3D hand joints, hand segmentation masks, and the 2D hand pose annotations are exploited for training. *Note that, the annotation quality in FreiHAND is not good enough*.

Hand Image Understanding Dataset. The collected **HIU-Data** fills the void that no accessible large-scale datasets contain high-quality hand masks and various challenging pose gestures. Specifically, the HIU-Data consists of $30,000$ training samples and $3,000$ testing samples. For each sample, both the 2D hand pose and hand mask are manually annotated rather than generating an approximate label automatically. [1]

[1]The dataset will be publicly available, and we sincerely expect that the HIU-Data will be beneficial to the community.

Dataset	Components					SSL
	2D Pose	3D Pose	Mask	Mesh	MANO	
STB [17]	✗	✗	-	-	-	✓
RHD [20]	✓	✓	✓	-	-	✓
Dexter [13]	✗	✗	-	-	-	✓
FreiHAND [21]	✗	✓	✗	✗	✗	✓
CMU [12]	✗	-	-	-	-	✓
HIU-Data	✓	-	✓	-	-	✓

Table 1: **Experiment Configuration.** The table presents the dataset configuration for *qualitative comparisons*. In which, MANO refers the 'ground-truth' MANO [11] parameters, Mask refers the annotated hand segmentation mask, and SSL refers the self-supervised learning strategy. ✓ for 2D Pose indicates exploiting 2D hand pose annotation for supervised learning; ✗ for 2D Pose denotes does not employ the 2D pose annotation for supervised learning; so to the 3D Pose, Mask, Mesh, and MANO.

Dataset	Components					SSL
	2D Pose	3D Pose	Mask	Mesh	MANO	
STB [17]	✗	✓	-	-	-	✓
RHD [20]	✗	✗	✗	-	-	✗
Dexter [13]	✗	✗	-	-	-	✓
FreiHAND [21]	✗	✓	✓	✗	✗	✗
CMU [12]	✗	-	-	-	-	✓
HIU-Data	✓	-	✓	-	-	✗

Table 2: **Experiment Configuration.** The table presents the data configuration for *quantitive evaluating* on the STB [17], Dexter Object [13], FreiHAND [21], CMU [12], and the HIU-Data. In which, each table element shares the same definition as Table 1.

2.2 Dataset Configuration

We have evaluated our HIU-DMTL framework on STB, RHD, Dexter Object, FreiHAND, CMU, and the HIU-Data datasets in the main paper. Existing approaches adopt various dataset configurations to benchmark the above datasets, For instance, [3] exploits additional $315,000$ training samples, [1] adopts MANO to generate extra training datas, [19] turns to training on datasets [9, 12, 20] jointly, and [4] synthesizes a new large-scale dataset. **Further, [1, 3, 4, 8] fully exploits labels that are hard to obtain in real-world situations, such as the whole hand mesh annotations and synchronized data from depth-sensor.** By contrast, HIU-DMTL entirely exploits easily annotated labels for supervised training, such as 2D hand pose, hand segmentation mask, and sometimes 3D hand joints. Besides, only a portion of data/labels are used for supervised training, and most images are exploited to achieve self-supervised learning only.

In this work, we try our best to find suitable data configurations for fair comparisons with SOTAs on the above datasets. Table 1 presents the data configuration for qualitative evaluation experiments. Table 2 reports the data configurations in the training procedure to benchmark on STB [17], Dexter Object [13], FreiHAND [21], CMU [12], and the HIU-Data, while a different data configuration (Table 3) was adopted

Dataset	Components					SSL
	2D Pose	3D Pose	Mask	Mesh	MANO	
STB [17]	✗	✗	-	-	-	✗
RHD [20]	✓	✓	✓	-	-	✓
Dexter [13]	✗	✗	-	-	-	✗
FreiHAND [21]	✗	✗	✗	✗	✗	✗
CMU [12]	✗	-	-	-	-	✗
HIU-Data	✓	-	✓	-	-	✓

Table 3: **Experiment Configuration.** The table presents the dataset configuration for *quantitive evaluating* on the RHD [20]. In which, each table element shares the same definition as Table 1.

to benchmark on RHD [20] due to the domain gap [9, 18] between RHD and other real-world datasets. **In all experiments, the ground-truth MANO parameters and hand mesh representations are never exploited for training and evaluation, though using those additional annotations may further improve the performance** ([3, 8] *et.al*, have exploited those labels to achieve SOTA results).

We shall point out a exceptional case, in the *ablation study of the Multi-Task Learning Setup*, to make a fair ablation study, we split out 20% training samples for testing the performance of each task, and only the other 80% training samples are used for training the HIU-DMTL framework.

2.3 Network Configuration.

In the absence of explicit instructions, a 4-stack HIU-DMTL framework is employed to conduct experiments on the above datasets, and the exceptions are as follow,

1. In the ablation study of the Multi-Task Learning Setup, we conduct the ablation experiments with a 1-stack HIU-DMTL framework to eliminate the interferences of other components.

2. In the ablation study of the Task Attention Module (TAM), we conduct the ablation experiments with a 4-stack and 1-stack HIU-DMTL framework that have similar FLOPs to better investigating the effect of the TAM under various configurations.

3. In the ablation study of the Cascaded Design (CD) Paradigm, we firstly design a 8-stack HIU-DMTL framework to explore the performance of each intermediate stack. Secondly, we design three models, which contain 1, 2, and 4 stacks (with similar FLOPs) to investigate the impact of the number of stacks.

3 Additional Evaluation.

2D Hand Pose Estimation. We present the quantitative compare comparison results with Stacked Hourglass [10], Convolutional Pose Machines [15], and HRNet [**?**]. As Table 4 reports the quantitative results on HIU-Data, our HIU-DMTL framework

	Hourglass [10]	CPM [15]	HRNet [14]	HIU-DMTL
AUC	0.814	0.821	0.852	**0.867**

Table 4: **Performance of 2D Hand Pose.** The table presents the quantitative results of 2D hand pose of [10, 15, 14] and Ours (HIU-DMTL), respectively. Approaches [10, 15, 14] are trained on the HIU-Data with the public accessible code.

	2D Pose	2D Pose†	Hand Mask	Hand Mask†	3D Pose	Hand Mesh
Ours	0.866	**0.704**	**0.974**	**0.770**	**0.860**	**0.856**
[16]	**0.869**	0.695	0.970	0.761	0.854	0.851

Table 5: **Ablation Study.** The table presents the ablation results across different POF encodings under various evaluation metrics, where † indicates inferring the pose/mask by projecting the 3D pose/mesh with proper camera parameters. The 3D hand pose/mesh are quantified on FreiHAND benchmark, while the 2D pose/mask are evaluated on the HIU-Data, since the quality of masks in FreiHAND benchmark is not good enough.

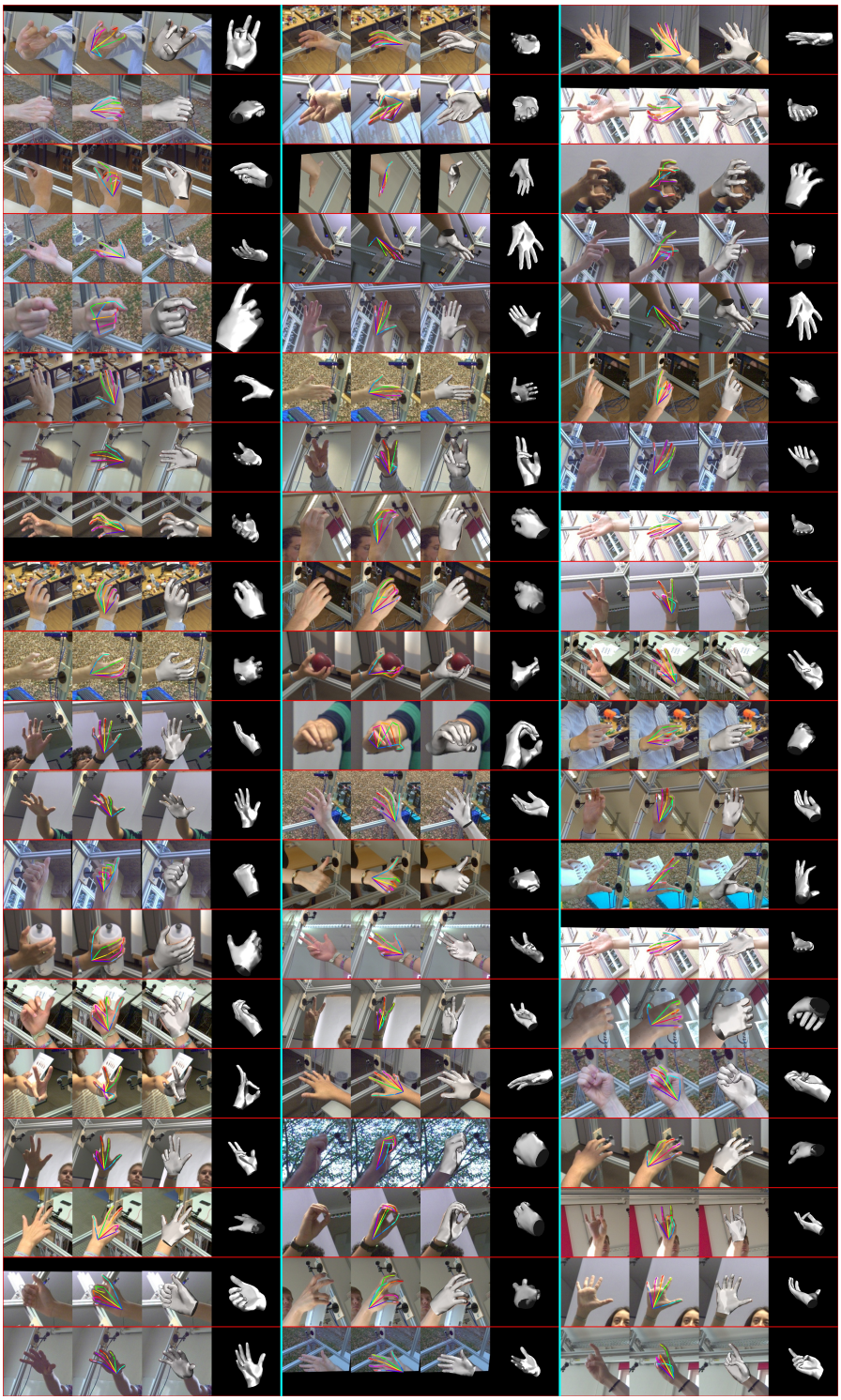
largely outperforms the classical methods [10, 15] and achieves better performance than [**?**].

Ablation of POF Encoding. We compare our POF encoding with that proposed by [16], and the comparison results are reported in Table 5. One may observe that, in terms of most evaluation metrics, our POF encoding obtains better performance than that exploited in [16].

4 More Qualitative Results.

The main paper presents several qualitative comparing results over the state of the art approaches. To better inspect the performance of HIU-DMTL in challenging situations, we randomly draw some representative samples from the FreiHAND [21] dataset and visualize the results below. One may observe that harnessing the advantages of the multi-task learning setup and the self-supervised learning strategy, our method can produce reasonable estimation in most typical challenging situations.

We kindly suggest the reviewers to review the demo videos in the supplementary for more detailed presentations, where the segmentation mask, 2D hand pose, and recovered hand mesh are well visualized on in-the-wild videos.



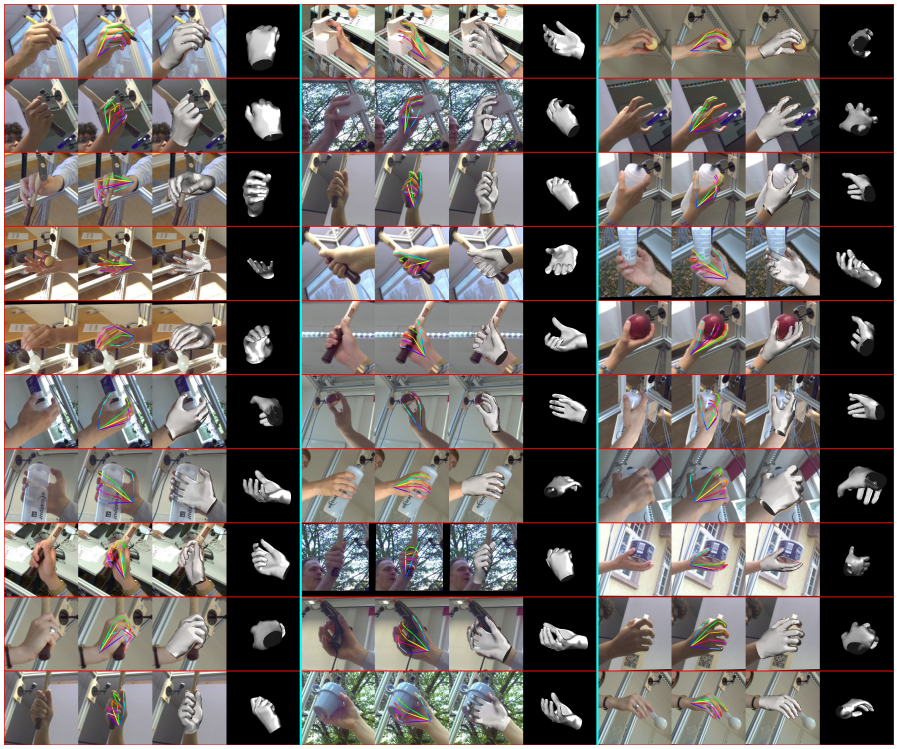


Figure 2: **More Qualitative Results.** The figure presents more qualitative results on the challenging FreiHAND dataset [21], which covers diverse situations, *e.g.,* hard hand pose, exaggerated camera view, hand object interaction, and heavy occlusion.

References

[1] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019. 4

[2] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 666–682, 2018. 3

[3] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10833–10842, 2019. 4, 5

[4] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11807–11816, 2019. 4

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[6] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4990–5000, 2020. 2

[7] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248, 2015. 2

[8] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. *arXiv preprint arXiv:2008.03713*, 2020. 2, 4, 5

[9] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated 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. 4, 5

[10] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016. 1, 5, 6

[11] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):245, 2017. 2, 4

[12] 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. 3, 4, 5

[13] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *European Conference on Computer Vision*, pages 294–310. Springer, 2016. 3, 4, 5

[14] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 6

[15] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016. 5, 6

[16] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10965–10974, 2019. 6

[17] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. A hand pose tracking benchmark from stereo matching. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 982–986. IEEE, 2017. 3, 4, 5

[18] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2354–2364, 2019. 2, 3, 5

[19] 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. 4

[20] 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. 3, 4, 5

[21] 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. 3, 4, 5, 6, 8