# Human Pose Estimation

Matteo Anedda, Christoph Schröter, Masoud Taghikhah

***Abstract*—later**

## I. Introduction

In computer vision, image data acquisition is divided into the following categories. *passive* or *active* sensing. Passive sensing uses visible light or other electromagnetic wavelengths to acquire position data, while active sensing requires special devices attached to the person's body to obtain parametric data.

## II. Benchmarks

Huge data sources are part of the benchmarking process for human posture estimation approaches. The datasets used specifically for this purpose contain images showing one or more individuals in different poses, as well as other information about joint and limb positions. This information is acquired by motion capture markers on the body or using *IMU* units as passive sensing devices, attached to the body to determine the position regardless of obstacles. Existing video footage can also be manually annotated. Some datasets focus on different features within its content to ensure the quality of a model in relation to that aspect. Commonly used datasets are explained in more detail in this section.

### A. Max Planck Institute datasets

*Perceiving Systems*[1] is a department of the *Max Planck Institute for Intelligent Systems*[2] that is specialized in computer vision and, in addition to scientific publications, also provides datasets[3] for e.g. pose estimation approaches. The listed datasets can be subdivided according to the following aspects:

*1) Clothing extension:* The individuality of a person is expressed by his clothing. To account for this feature, a model called *CAPE* is built from 4d posture sequences of 8 men and 3 women. This dataset consists of about 80000 frames. [18] Further work by Qianli et al. has generated a synthetic dataset on specific *CAPE* subjects and published it as *ReSynth* for researchers [19].

*2) Full-body scans:* Acquiring 3d scans and data from multiple people in an outdoor environment is challenging because the markers are difficult to track. Timo et al. have shown in their publication that capturing sufficient data in a scene is possible with 6-17 *IMU* units attached to each person, combined with a single hand-held camera. The recorded 51000 images are available for research. [35] A similar approach is followed by Yinghao et al. with 17 *IMU* units for 10 subjects in 64 sequences, resulting in 330000 time instances [14]. Human-environment interaction is mainly covered in the datasets of Mohamed et al. which consist of three parts in different scenes [10]. The *GRAB* dataset, on the other hand, targets the relationship between full-body models and object manipulation. It contains motion data of 10 individuals interacting with 51 objects in 4 different contexts, e.g., lifting, transferring, hand-to-hand transfer, and using [30]. In contrast to Grab, the Lea et al. datasets include all of a person's interactions with themselves [23].

*3) Hand scans:* The hand contributes to communication, e.g., the hand gesture is used to confirm a statement in conversation. To incorporate this expressiveness into existing full-body models, Javier et al. developed the *MANO* model from approximately 1000 3d scans of 31 subjects in 51 poses. These scans showed female and male hands, both left and right, interacting with primitives. [28] Yana et al. also published a synthetically generated hand dataset *obman* that focuses on the manipulation of grasped primitives [11].

*4) synthetic data:* A much more cost-effective approach is to create realistic body data from existing motion capture sources. An prominent example is SURREAL by Gül et al. which consists of 6 million frames [34]. David T. et al. also published their data set with pure synthetic and more realistic mixed material [13].

*5) Generalization of datasets:* Many different 3d scans are based on markers and motion capture software. Unfortunately, the number of markers varies from dataset to dataset, so their use as a data source for a body model leads to inaccuracies and further adjustments. A common solution to this problem is provided by the *MoSh++* algorithm, a descendant of the earlier motion capture software, and its resulting *AMASS* dataset. It consists of 11265 motions from 344 subjects with 40 hours of content. [20]

### B. COCO dataset

According to Tsung-Yi et al. the context of a given scene has an impact on the quality of the estimation. Therefore, to encode this contextual data into the dataset and the learned models, images with many classifiable objects, such as animals, people, etc., are the content of the dataset. In addition, mainly non-iconic images of no centered objects are included. On the other hand, useful labels were also used in a hierarchical modality, simply describing the subject and differentiating by body parts. The dimension of *COCO* is 328000 images, divided into 91 object categories with a total of 2500000 label instances. [16] [1]

[1] https://ps.is.mpg.de/

[2] https://is.mpg.de/

[3] https://ps.is.mpg.de/research_fields/datasets-and-code

## III. CRITERIA

The quality of the estimated poses and thus of the applied algorithm is evaluated with the help of metrics. The common approach is to calculate the body parts or joint positions and compare them with particular values from the ground truth data, e.g. the lengths between key-points or frames in a specific region. For this analysis, the previously mentioned datasets from section II are used as reference. The accuracy of the results can be adjusted by a threshold value for the respective metric.

### A. Percentage of Correct Parts PCP

This criterion encompasses a comparison of the recognized and recognizable body parts. The definition of a correctly recognized limb includes both the distances $l_1, l_2$ of its endpoints from those contained in the dataset and its total length $L$. Figure 1 illustrates the values explained earlier in an intuitive way, with the left (green) line indicating information from the dataset that is compared to the estimated right line. Another factor $p$ multiplied by $L$ defines the threshold value to which $l_1$ and $l_2$ are compared. If $l_1$ or $l_2$ exceed the threshold, the body part is not detected correctly, resulting in a lower $PCP$ score. The smaller p, the stricter the evaluation and thus higher the accuracy. [7]

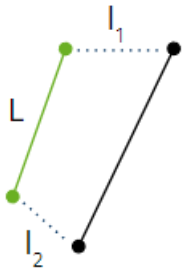


Fig. 1. Visualisation of estimated values for PCP calculation

### B. Percentage of Detected Joints PDJ

The metric *PCJ* addressed in this section is similar to the already established *PCP* methode. While *PCP* depends on individual limb lengths, *PCJ* uses the torso length $T$ as a global reference, so that a body part is correctly detected if both $l_1$ and $l_2$ do not surpass the threshold given by $T$ and a factor $p$. [33]

### C. Percentage of Correct Key-points PCK

For this particular criteria, the maximum bounding box length $B$ must be calculated from the existing dataset information. The *PCK* score is calculated analogous to the previously mentioned algorithms in subsection III-A and subsection III-B from $l_1, l_2$ and a further threshold defined by $B \times p$. [1], [37]

*1) Head-normalized Probability of Correct Key-points PCKh: PCKh* is a variant of *PCK* in which the threshold is set to half the length of the head frame $H \times 0.5$. Since the head frame is independent of the viewpoints and position of other body parts, *PCKh* is not affected by the subjects' articulation. [1]

### D. Area Under the Curve AUC

The measurement of the *PCK* under variation of the threshold value results in a curve. This analysis provides information on how the model is able to distinguish the individual joints of the body. A large curve area defines a qualitative model

### E. Object Key-point Similarity OKS

Equation 1 illustrates the simplified OKS score[4], which is a sum of $n \in \mathbb{N}$ detected and ground truth key-points. The parameters $d \in \mathbb{R}$ are the Euclidean distance between the corresponding ground truth value and the detected key-point, while $s \in \mathbb{R}$ denotes the object segment area and $k \in \mathbb{R}$ is a constant describing a falloff.

$$OKS = \sum_{i=1}^{n} e^{-\frac{d_i^2}{2 \times s^2 \times k_i^2}} \tag{1}$$

Optimal predictions have a high OKS value, while low values indicate poor predictions.

## IV. 2D POSE ESTIMATION

## V. 3D POSE ESTIMATION

From a historical perspective, a 3d motion capture algorithm consists of 4 sequential processes: initialisation, tracking, pose estimation and recognition. Initialization involves both camera and model initialization, i.e. setting the camera calibration and finding a model that represents the subject and assigning its initial pose manually or automatically. Model-based approaches can be viewed iteratively, with each frame of the data source representing an iteration in which the initial pose is refined. Tracking is concerned with the relationship between the parts of the subject's body. This leads to segmentation of the subject from the background, representation changes, and establishing tracking in further images. The next phase, which is mainly covered in this section, is the estimation of the pose. A distinction is made between model-based and non-model-based methods, with the former requiring *a priori* a model. In that approaches, especially human pose estimation, a human model is used to benefit from its encoded information. This model can either be used indirectly, considering e.g. only general aspects such as size and structure, or it is a direct used model. Directly used models are both more detailed and offer broader benefits in regards to occlusion handling and embedded kinematic constraints. In an application, the observed object is approximated by the model, which is continuously refined with further images. [22]
As with 2D Pose Estimation, neural networks, particularly convolutional networks, have successfully been used to

[4]https://cocodataset.org/#keypoints-eval

achieve more accurate results than earlier methods [3]–[5], [21], [32], [38]. Since neural networks can be very resource intensive to compute and the architecture can be extremely complex when working with 3D data, many of the presented approaches use 2D Pose Estimation followed by an uplifting process to 3 dimensions. Also, a lot of 2D training data is available in comparison to 3D data which could be used to train a neural network, because the annotation process is way harder in higher dimensions. Therefore [5] presents a process to synthesize training images and shows that neural networks training with data generated by their method are even more effective than neural networks which were trained using real images. [27] presents a similar approach.

### A. Lifting from 2D to 3D pose

For uplifting, recent work has proven statistical models such as (deep) neural networks themselves [21], [32], matching the estimated 2D pose with a database [4] or triangulation using multiple viewpoints [6] useful. In particular [21] shows that even very simple deep neural networks can be extraordinarily effective for uplifting 2D to 3D pose estimations, considering both computational resources and failure rate.
Inspired by various 2D human pose estimation algorithms, many studies have employed the outputs of 2D pose estimate methods for 3D human pose estimation to improve in-the-wild generalization performance. For example, Martinez et al. [21] pioneered the research on lifting 2D poses to 3D space with a simple yet effective neural network. Other methods [8], [25], [31], [36], [39] focus on fusing 2D joint heat maps from the top-down 2D pose estimation methods with 3D image cues to reduce ambiguity.

*1) A simple yet effective baseline for 3d human pose estimation:* While trying to investigate common errors in the uplifting process, [21] created a method with state-of-the-art results for 3d pose estimation using a very basic neural network with recently proposed optimization methods, whose structure is shown in Figure 2. Linear layers changing the input and output dimensions are not shown. 2d joint positions are used as input data determined by the so called stacked hourglass network as described in [24], while the output consists of 3d joint positions.
Since the 2d joint positions are low-dimensional and therefore no highly complex computation is necessary, a simple linear layer with a RELU activation function is used first, followed by batch normalization and dropout as presented in [29] to prevent overfitting and improve result quality at the cost of a slight increase in computation time during training and testing. This structure is then repeated and a residual connection to the output added. These connections are proposed to help improve performance, reduce training time and lower error rates in [12]. The whole network established so far is doubled to complete the architecture which in sum consists of 4 to 5 million trainable parameters. It was then trained on Human3.6M, a dataset with 3.6 million 3d poses of humans during normal activities such as eating or walking [15].
Testing results show that this approach outperformed previous methods like [26] in most cases, despite the simple architecture that was used. However, testing on the MPII Human Pose Dataset [2] also revealed limitations of the network shown especially in the bottom row of Figure 3. Firstly (left and right picture), the 2d joint position must be detected properly. The middle picture also rendered a problem, which according to the original paper comes down to poses being not included in the Human3.6M dataset, such as upside-down poses or examples not showing a full human body.
The authors conclude that basic neural networks today are already able to produce very good results in terms of accuracy for uplifting 2d to 3d human poses. Therefore one of the main error sources of this process remains 2d pose estimation and more complex models should be able to perform the task of uplifting even better.

### B. Direct 3d pose estimation

Methods working directly on 3D data, such as [38] (working with depth maps)can avoid potential sources of error such as projections or lighting conditions. This leads to more robust and accurate results, however more complex (neural network) architectures and computational resources are required. Traditional methods like the least-squares-estimation presented in [9] work without training data needed and are computationally inexpensive, but yield rather high error rates compared to newer methods.

*1) Skinned Multi-Person Linear model SMPL:* In *SMPL* a model $M(\vec{\beta}, \vec{\theta}, \phi)$ is learned from the 3d scans explained in section II, that returns a mesh from the input. This formulation is also included in Equation 2, where $\mathbb{R}^{3N}$ is a vector of *N = 6890* vertices sculpturing the mesh. In this formula, $\vec{\beta}$ is a vector of blend shapes, while $\vec{\theta}$ are poses and $\phi$ describes the displacement of soft tissues.

$$M(\vec{\beta}, \vec{\theta}, \phi) : \mathbb{R}^{|\vec{\theta}| \times |\vec{\beta}|} \mapsto \mathbb{R}^{3N} \tag{2}$$

*SMPL* is based on vertex skinning and blend shapes. A vertex changes its position depending on the motion of the associated joint. This displacement is controlled by assigned blend weights. A vector $T \in \mathbb{R}^{3N}$ of vertex positions describes a gender neutral initial human model, while a matrix $W \in \mathbb{R}^{N \times K}$ represents the blend weights per vertices and *K = 23* joints. The joints that describe the human structure and form the skeleton are represented by rotation vectors. Moreover, T can be rearranged by the pose-blending function $B_P(\vec{\theta})$ according to the given poses, leaving T unaffected, while $B_S(\vec{\beta})$ reshapes the identity model by its given shape blends.

[17]

## VI. CONCLUSION

## REFERENCES

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

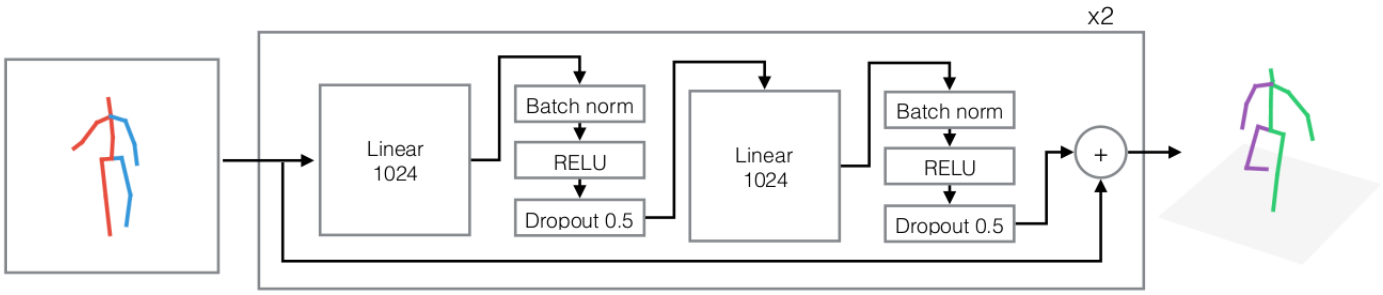


Fig. 2. Neural network structure from [21]

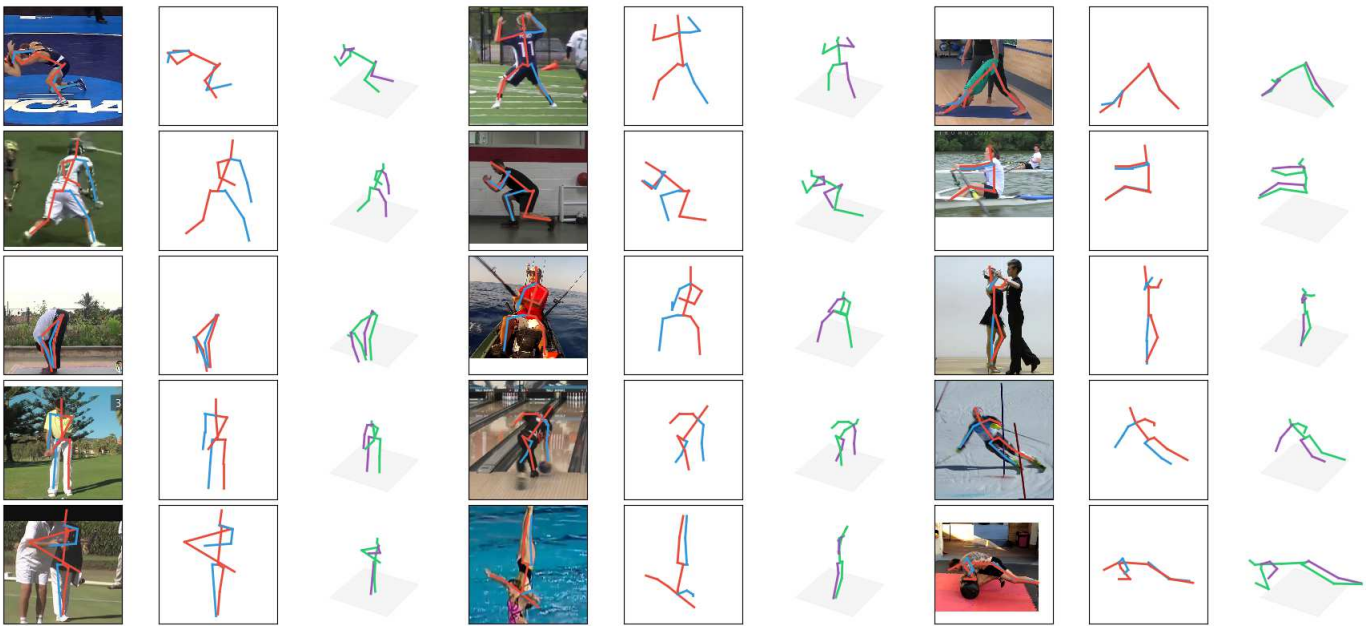

Fig. 3. Test cases from [21]


[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[3] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Monocular 3d pose estimation and tracking by detection. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 623–630, 2010.

[4] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[5] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3d pose estimation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 479–488, 2016.

[6] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[7] Marcin Eichner, Manuel Marin-Jimenez, Andrew Zisserman, and Vittorio Ferrari. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *International journal of computer vision*, 99(2):190–214, 2012.

[8] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the Wild Human Pose Estimation Using Explicit 2D Features and Intermediate 3D Representations. pages 10905–10914, 2019.

[9] R.M. Haralick, H. Joo, C. Lee, X. Zhuang, V.G. Vaidya, and M.B. Kim. Pose estimation from corresponding point data. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(6):1426–1446, 1989.

[10] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *Proceedings International Conference on Computer Vision*, pages 2282–2292. IEEE, October 2019.

[11] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816, June 2019.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[13] David T. Hoffmann, Dimitrios Tzionas, Michael J. Black, and Siyu Tang. Learning to train with synthetic humans. In *German Conference on Pattern Recognition (GCPR)*, pages 609–623. Springer International Publishing, September 2019.

[14] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time.

*ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 37:185:1–185:15, November 2018. Two first authors contributed equally.

[15] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014.

[17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[18] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3D people in generative clothing. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6468–6477. IEEE, June 2020.

[19] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *Proc. International Conference on Computer Vision (ICCV)*, pages 10974–10984, October 2021.

[20] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings International Conference on Computer Vision*, pages 5442–5451. IEEE, October 2019.

[21] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[22] Thomas B. Moeslund and Erik Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001.

[23] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 9990–9999, June 2021.

[24] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation, 2016.

[25] Sungheon Park, Jihye Hwang, and Nojun Kwak. 3D Human Pose Estimation Using Convolutional Neural Networks with 2D Pose Information. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, Lecture Notes in Computer Science, pages 156–169, Cham, 2016. Springer International Publishing.

[26] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. *CoRR*, abs/1611.07828, 2016.

[27] Gregory Rogez and Cordelia Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[28] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):245:1–245:17, November 2017.

[29] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[30] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision – ECCV 2020*, volume LNCS 12355, pages 581–600, Cham, August 2020. Springer International Publishing.

[31] Bugra Tekin, Pablo Marquez-Neila, Mathieu Salzmann, and Pascal Fua. Learning to Fuse 2D and 3D Image Cues for Monocular Body Pose Estimation. pages 3941–3950, 2017.

[32] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[33] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[34] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, pages 4627–4635, Piscataway, NJ, USA, July 2017. IEEE.

[35] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, volume Lecture Notes in Computer Science, vol 11214, pages 614–631. Springer, Cham, September 2018.

[36] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. Deep 3D human pose estimation: A review. *Computer Vision and Image Understanding*, 210:103225, September 2021.

[37] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2878–2890, 2013.

[38] Mao Ye, Xianwang Wang, Ruigang Yang, Liu Ren, and Marc Pollefeys. Accurate 3d pose estimation from a single depth image. In *2011 International Conference on Computer Vision*, pages 731–738, 2011.

[39] Kun Zhou, Xiaoguang Han, Nianjuan Jiang, Kui Jia, and Jiangbo Lu. HEMlets Pose: Learning Part-Centric Heatmap Triplets for Accurate 3D Human Pose Estimation. pages 2344–2353, 2019.