Human Pose Estimation

Matteo Anedda, Christoph Schröter, Masoud Taghikhah

Abstract—later

I. Introduction

Human body estimation is a broad and increasingly relevant area of research that allows human movements to be detected or even predicted, and also their meshes to be created based on visual input. Due to the widespread distribution and availability of relatively inexpensive but powerful video recording hardware and computers, the researched technique of pose estimation can be used by many people in various fields of application. One such application may be in surveillance areas where installed cameras record information to facilitate later investigations into possible crime, e.g. in the security software of petrol stations or car parks. Or in control-based domains such as the new paradigms of human-computer interaction known as virtual and augmented reality, where immersion is achieved by capturing human movements and executing them through the avatar. This can be applied not only to today's video games, but also to online business meetings, sprints or educational courses that can take place in a virtual environment. Moreover, the approaches are existential for the futuristic meta-project in which social life is also shifting into a virtual world and new forms of communication are being introduced. An further interesting aspect is also the use in autonomous maschines. Here autonomous cars need to make decisions considering the near future, therefore they need to be able to predict human motion, which is tightly coupled with their pose. The same applies to robotic systems working around or even with humans. Since human posing is an indicator for their emotions aswell, social or collaborative robots could use this information in the future. Even medical information can be obtained from a person's posture and used in diagnostics and therapy or for the correct posture when sitting or doing sports. These examples show that human pose estimation is an important research area today and will become even more so in the future.

However, there are challenges to overcome in order to deliver accurate estimations. Firstly, a vulnerable source of error is associated with the acquisition of the image data, which 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. With passive sensing, errors can occur due to different lighting conditions, backgrounds or clothing. Multiple or incomplete humans could also hamper the algorithm. Furthermore, 3d pose estimation needs to add an extra dimension of depth, which could lead to ambiguities. As neural networks, which are used a lot for pose estimation today,

get bigger and used more frequently, enough and sufficient training and testing data needs to be available in the future. These datasets also need to be representative for humans in their normal environment. Annotation, especially for 3d data, can be a very resource-intensive and time-consuming task too. This paper aims to summarize metrics methods for human pose estimation in two or three dimensions and show off their strengths and weaknesses. Finally, the current state-of-the-art can be evaluated and an outlook into the possible future can be provided.

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. [47] Further work by Qianli et al. has generated a synthetic dataset on specific *CAPE* subjects and published it as *ReSynth* for researchers [48].

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. [80] A similar approach

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

is followed by Yinghao et al. with 17 *IMU* units for 10 subjects in 64 sequences, resulting in 330000 time instances [33]. Human-environment interaction is mainly covered in the datasets of Mohamed et al. which consist of three parts in different scenes [28]. 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 [71]. In contrast to Grab, the Lea et al. datasets include all of a person's interactions with themselves [53].

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. [64] Yana et al. also published a synthetically generated hand dataset *obman* that focuses on the manipulation of grasped primitives [29].

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 [78]. David T. et al. also published their data set with pure synthetic and more realistic mixed material [32].

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. [49]

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. [42] [1]

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. [19]
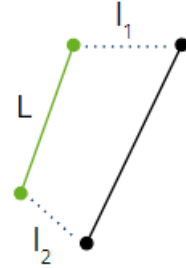

Fig. 1. Visualisation of estimated values for PCP calculation

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. [76]

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], [86]

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.

F. Mean Per Joint Position Error (MPJPE)

IV. 2D POSE ESTIMATION

The foremost step in Human Pose Estimation (HPE) is to model the human body entity representing its kinematic structure and shape information. In recent studies, benefiting from the intuitive nature of graph representations, the human body structure has been characterized by anatomical joints and their positions. This approach is known as the kinematic model and is widely employed to describe human poses. In 2D Human Pose Estimation, the goal is to localize the 2D position or spatial location of human body keypoints (kinematic joints) from the input data [60].

Recently, deep neural networks have achieved a significant breakthrough in 2D HPE and are mainly employed to extract robust features for keypoint recognition and localization directly from the input data (images and videos). Since the quality of feature representation is intimately tied to network architecture, the subject of network design was chosen to be thoroughly investigated.

A. HPE Network Architecture Design Challenges

Like many tasks in computer vision (e.g., object detection, image segmentation, etc.), the primary technological challenges in developing 2D HPE algorithms are the accuracy and efficiency of the proposed method. High precision pose estimation facilitates accurate human body information for subsequent tasks, including lifting from 2D to 3D HPE models.

In addition to the accuracy of an HPE algorithm, its efficiency (inference speed) is also desired for real-time applications. However, there is mostly a trade-off between accuracy and efficiency concerning algorithm development since the high accuracy methods tend to be more complex and demand powerful resources for computation. Consequently, lightweight models with comparable precision are more desirable for mobile or wearable devices.

Another crucial challenge in 2D HPE is estimating poses for multi-person scenarios. 2D single-person HPE localizes joint coordinates when only one human instance is in the input image, but for images containing multiple persons, the

---

HPE model needs to recognize each human in the input image before estimating its keypoints positions.

This process can be done by a two-stage process called the top-down approach that first detects individual persons, i.e., the input image is divided into sub-images (patches) enclosing just one person [68]. Then the deep learning keypoint detection algorithm can be applied to each patch.

On the other hand, there is an alternative group of approaches known as bottom-up, in which human joints are detected directly from the image without an explicit object detection stage [14]. The following sections will review current frameworks utilizing top-down and bottom-up approaches to their network architectures.

B. Top-Down Approaches

The top-down approaches employ human body detectors [51], [62] to obtain a set of bounding boxes (each corresponding to one person) from the input images and then perform pose estimation within each bounding box. The existing literature in 2D HPE based on top-down approaches can also be divided into fine-grained sub-categories discussed in this section: regression-based and heat map-based methods.

1) Regression-based methods: There are many works based on the regression technique [**?**], [9], [20], [22], [40], [59], [69], [70], [77], [81] that directly regress human body joint coordinates via an end-to-end trained network that maps the input image to the pre-defined graphical structures (kinematic keypoints). Accordingly, the regression hypothesis is defined as follows:

Given an image I, the goal of pose estimation is to predict a possibly empty set of human instances, $\{Pi\}_{i=1}^{N}$ where $\textbf{\textit{N}}$ is the number of persons in the image. For each person, we need to predict its bounding box information, $\{Bi\}$, as well as its keypoint coordinates, $Ki = \{(x_j, y_j)\}_{j=1}^{J}$, where $\textbf{\textit{J}}$ is the number of pre-defined joints in each dataset [39].

DeepPose [77] is a regressor that explicitly predicts human joint locations through fully connected layers of a deep neural network. AlexNet [38] as the backbone of DeepPose's cascaded architecture for feature extraction. Due to the remarkable performance of DeepPose in efficient keypoint detection by using convolutional neural networks (CNNs), human pose estimation techniques have gradually migrated from the conventional graphical models to deep learning approaches.

Carreira et al. [9] proposed a model based on GoogLeNet [220-2] that adopts an iterative feedback mechanism called Iterative Error Feedback (IEF) to enhance the efficiency of hierarchical feature extractors such as Convolutional Networks (ConvNets) in early layers. This self-correcting model progressively adjusts an initial estimate of joint coordinates by feeding back predictions error in the IEF process instead of directly predicting the coordinates in a feed-forward network.

Transformer models [79] based on the self-attention mechanism have significantly advanced the field of representation learning and promoted visual understanding tasks such as object detection frameworks that are free of region propos-

---

als, anchors, and post-processing (non-maximum suppression) procedures.

In a recent study [39], authors took full advantage of the tokenized representation in transformers with self-attention layers. They propose a regression-based human pose estimation method, called pose recognition Transformer (PRTR), using two types of cascade transformers based on an end-to-end object detection Transformer (DETR) [8]. In this proposed model, the second transformer is a regressor that predicts joints' coordinates within each image patch detected from the first transformer and performs multi-person pose estimation. PRTR reveals competitive results for pose recognition compared with other existing regression-based methods on the challenging COCO dataset.

2) Heatmap-based Methods: In the field of human pose estimation, heatmap-based approaches seek to train a deep neural network to approximate the locations of body parts and joints based on heatmap representations [12], [54], [83]. Accordingly, during training, the pose estimation model generates J heatmaps $\{Hj\}_{j=1}^{J}$, as a 2-dimensional Gaussian distribution centered at the ground-truth joint location, where J is the total number of joints. The pixel value $Hj\{(x_i, y_i)\}$ in each joint heatmap encodes the probability that the jth joint lies in the location (x, y) [74], [75].

The probability values provide richer supervision that facilitates the training of convolutional networks and significantly improves the performance of heatmap-based over regression-based approaches [4], [6], [24], [41], [54], [74]. Therefore, heatmaps preserve the spatial location information of joints and reduce the number of false-positive cases, which results in an increasing attraction to develop deep learning pipelines for HPE based on heatmap representations [9], [45], [77], [83].

Ramakrishna et. al [61] presented a sequential prediction framework based on convolutional networks named Convolutional Pose Machines (CPM) that learns rich implicit spatial structures by utilizing the inference machine framework to incrementally refine estimates of the human body part (e.g., head, leg, etc.) locations in multiple stages. In this method, a predictor is trained to predict the confidence of the body part locations in each stage. In the first stage, the predictors estimate the confidence of each part location from heatmaps computed on the input image patch. Since a heatmap produced only from image features is noisy and has multiple modes, in the subsequent stage, the network iteratively refines the estimated confidences using contextual information from outputs of the previous stage.

In a further expanding of Pose Machine architecture, [83] employed sequential convolutions to implicitly model long-range spatial relationships between different parts. It is worth mentioning that multiple stages increase the network depth and make it hard to train because of the vanishing gradient problem — as the gradient is back-propagated to the layers of earlier stages, repeated multiplication makes the gradient infinitely small. Thus, as the network goes deeper, its performance gets saturated or degrades rapidly.

Before ResNet [31], there were several methods to deal with the vanishing gradient issue; for instance, the authors of this work [83] developed an auxiliary loss in a middle layer as extra supervision. But this can not tackle the problem entirely since each stage will fail to extract robust semantic features from the input data and is prone to overfitting.

The core idea of ResNet is introducing "shortcut connections" that skip (jump over) one or more layers. These skip connections effectively simplify the network and reduce the impact of vanishing gradients as it allows the back-propagation of training errors at deeper levels (addressing the issue of iterative architectures with multiple stages). Residual networks dramatically improved the 2D HPE process, and numerous models [7], [13], [15], [36], [43], [54], [67], [68], [84], [85] have been developed due to their advantage. An encoder-decoder network called "stacked hourglass" (SHG) based on the sequential stages of pooling and upsampling layers was presented by Newell et al. [54] to capture spatial correlations between the human body keypoints by combining low and high-resolution feature representations.

Several more complicated variants have been introduced following the initial success of SHG architecture; specifically, Chu et al. [15] developed novel Hourglass Residual Units (HRUs) extended by a side branch of filters with larger receptive fields that learn features across various scales. Yang et al. [85] replaced the residual units in SHG with multi-branch Pyramid Residual Modules (PRMs) to enhance deep convolutional neural networks by constructing scale invariance features.

Cai et al. [7] recently proposed the Residual Steps Network (RSN), which efficiently maintains rich low-level spatial information by aggregating intra-level features with the same spatial size to localize the keypoints precisely using delicate local representations. In addition, they devised a novel attention mechanism, Pose Refine Machine (PRM), that refines the keypoint locations by finding a trade-off between the contribution of local and global representations in the resultant feature. Their approach accomplished state-of-the-art results on COCO and MPII benchmarks and won 1st place in the COCO Keypoint Challenge 2019.

C. Top-Down Approaches Summary

The primary components of top-down frameworks in human pose estimation consist of an object detector and a pose estimator to predict the joint positions of the human body. The object detector determines human proposal detection performance and influences pose estimation. On the other hand, the pose estimator component is the framework's heart that directly affects the pose estimation accuracy. The main reason for adopting heat map-based and regression-based methods for the pose estimator component is the speed-accuracy trade-off.

In regression-based human pose estimation, the problem is formulated as a regression one in which features extracted from convolution layers of a CNN are regressed to directly predict joint coordinates of the body parts. The regression-based networks can be trainable end-to-end and are efficient in

real-time applications since they have fewer intermediate non-differentiable steps [18], [21], [23], [43]; however, they typically perform less accurately than heat map-based approaches. Because when the body parts are not completely visible, they fail to accurately estimate the locations of human body joints. To address this shortcoming, probabilistic heatmaps are employed to learn a complex mapping from occluded part appearances to joint coordinates instead of directly regressing them.

Human pose estimation based on heatmaps comprises pre-processing and post-processing procedures to encode key-points' ground truth (GT) into heatmaps and decode heatmaps to predict joint locations. Consequently, they can be optimized easier and in comparison to regression-based methods, have a more robust generalization that delivers substantial performance, primarily suitable to be adopted when accuracy is the priority [58].

However, heatmap-based frameworks have various heuristic network designs that are mostly not end-to-end learnable and suffer from several shortcomings: Firstly, to generate the 2D heatmaps, computationally expensive upsampling operations (e.g., deconvolution layers in [88]) are required. Also, an extra post-processing step for reducing projection errors from heatmap to ground truth coordinates is unavoidable in further keypoint estimation refinements. This makes them sub-optimal, i.e., the speed efficiency (performance) of heatmap-based methods usually declines drastically as the input image resolution decreases.

D. Bottom-Up Framework

After analyzing the top-down frameworks in the network architecture design context for human pose estimation, it is noteworthy to consider the bottom-up approaches. These approaches (e.g., [15, 32, 56, 75, 76, 90, 109, 169, 174, 191, 226, all 2]) aim to perform two main tasks: firstly, human keypoint detection through extracting local features from the input image and proposing a set of human body joint candidates, and then grouping those candidates to build pose representations for each individual person. The main distinguishing characteristic of the bottom-up approach from the former approach is that the framework design does not rely on a human detection component to predict human bounding boxes separately. As a result, the computational overhead decreases substantially by directly estimating human poses from extracted features; however, the major challenge is an identification mechanism for estimated keypoints to associate different subsets of candidates to each individual human body.

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. [52]

As with 2D Pose Estimation, neural networks, particularly convolutional networks, have successfully been used to achieve more accurate results than earlier methods [3], [10], [11], [50], [73], [82], [87]. 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. Especially the lack of in-the-wild 3D training data, which is not created under laboratory conditions, can be partly overcome by using 2D pose estimation methods (which work way better in-the-wild because of more training data available) first before adding an extra dimension. Therefore [11] 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. [63] 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 [50], [73], matching the estimated 2D pose with a database [10] or triangulation using multiple viewpoints [17] useful. In particular [50] 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. [50] pioneered the research on lifting 2D poses to 3D space with a simple yet effective neural network. Other methods [26], [56], [72], [82], [89] focus on fusing 2D joint heat maps from the top-down 2D pose estimation methods with 3D image cues to reduce ambiguity. The uplifting approach also makes

it possible to project the calculated 3d pose to 2d again to make sure results are consistent. [82]

1) A simple yet effective baseline for 3d human pose estimation: While trying to investigate common errors in the uplifting process, [50] 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 [55], 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 [66] 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 [30]. 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 [34].

Testing results show that this approach outperformed previous methods like [57] 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.

2) SVMAC: Unsupervised 3D Human Pose Estimation from a Single Image with Single-view-multi-angle Consistency: [16] proposes a generative adversarial network (GAN) method which works unsupervised with just 2d joint positions as inputs. The generator gets trained to lift the input to a 3d pose estimation and extract the camera position, which enables rotating the model to reproject it into two dimensions. The reprojected 2d joint positions aswell as the ground truth ones are used to train the corresponding discriminator of the GAN. Single-view-multi-angle Consistency (SVMAC) denotes the ability to mix different rotations of an estimate pose with those of the estimated camera and ensuring that reprojections into 2d from the same angle are consistent. This is used in the loss function of the generator. The structure of the network is shown in Figure 4. SVMAC constraints are applied using the generators on the right. SVMAC resulted in heavily improved performance of the model when using just 2 angles. However using more than 2 did not yield a significantly better result but increased training time by a lot. Another experiment done was including 5% ground truth data (3d poses) for supervision (with 2 angles used for SVMAC), which led to impressive results being able to outperform many weakly- and even some fully supervised approaches in terms of P-MPJPE (mean per joint position error, see subsection III-F, when aligning with procrustes analysis transformation [25] for alignment first). When considering only unsupervised methods, this one outperformed any previous one.

B. 3d pose directly from 2d image

Methods for 3d human pose estimation directly from an imagine (or many) show the lack of 3d training data not captured by special indoor motion capture systems. This often leads to worse performance when comparing to uplifting methods in a real scenario test. With more in-the-wils datasets coming up in the future, generating a 3d pose directly from a 2d image could be outperforming the uplifting approach. For now, generating artificial training data or using unsupervised methods yields the best results in this area.

1) 2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning: In [46] a convolutional deep neural network structure is proposed, which is able to handle 2d and 3d pose estimation from still images plus action recognition from video, since those problems are coupled. While not achieving groundbreakingly new state-of-the-art performance results in any of the solved tasks individually, this approach goes to show that a single architecture is able to produce at least equally as good results as dedicated methods for each task separately. Since the solved problems overlap, the network can work very efficient computationally. The authors claim that merging the tasks and training them together also increases robustness. Datasets used include Human3.6M [34] and MPII Human Poses [2], which render comparison with the approach presented in subsubsection V-A1 possible. The multitask convolutional neural network outperforms [50] quite a lot, what in turn proves that complex network structures can produce even better results, as Martinez et al. stated.

C. Direct 3d pose estimation

Methods working directly on 3D data, such as [87] (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 [27] 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
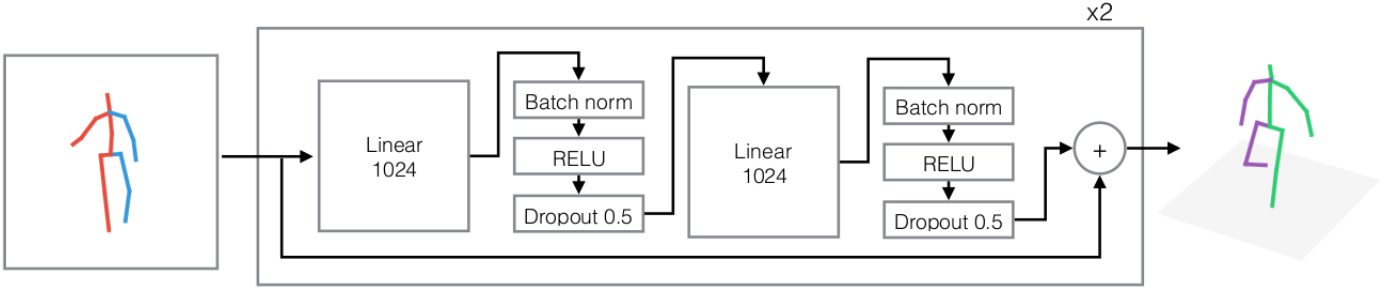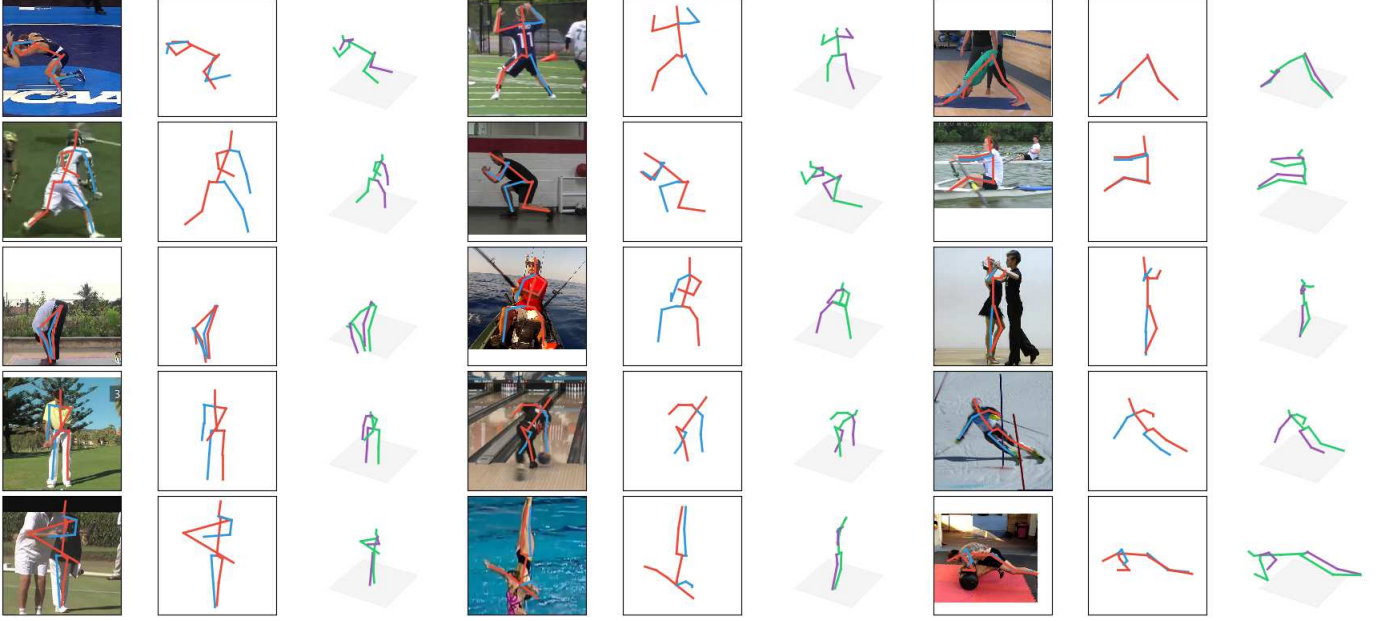
Fig. 2. Neural network structure from [50]

Fig. 3. Test cases from [50]

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 ϕ describes the displacement of soft tissues.

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

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. A skinning function $W(T, J, \vec{\theta}, \mathcal{W})$, on the other hand,

transforms the initial model T by the joint positions $J \in \mathbb{R}^{3K}$ and rotations $\vec{\theta}$ of the desired model. The simplest skinning approach is called *linear blend skinning*, where each resulting vertex $\bar{t}'_i$ depends on the sum of its weights $w_{k,i}$ per joint multiplied by its previous position $\bar{t}_i$ and a target destination, as documented in Equation 3.

$$\bar{t}'_i = \sum_{k=1}^{K} w_{k,i} \times G'_k(\vec{\theta}, J) \times \bar{t}_i \qquad (3)$$

Unfortunately, this approach is error-prone and requires additional refinement by the artists to reduce artifacts. To overcome these problems, the skinning method in *SMPL* introduces new functions $T_F(\vec{\beta}, \vec{\theta})$ and $J(\vec{\beta})$ so that Equation 2 and $W(T, J, \vec{\theta}, \mathcal{W})$ can be reformulated by Equation 4.

$$M(\vec{\beta}, \vec{\theta}) = W(T_F(\vec{\beta}, \vec{\theta}), J(\vec{\beta}), \vec{\theta}, \mathcal{W}) \mapsto \mathbb{R}^{3N} \qquad (4)$$

In $T_F(\vec{\beta}, \vec{\theta})$, the initial body T is considered as a known parameter, while the pose rearrangements $B_P(\vec{\theta})$ and the re-

Fig. 4. GAN structure from [16]

shapes $B_S(\vec{\beta})$ are added. Since the focus is on pose estimation, this section explains $B_P(\vec{\theta})$ in more detail. With respect to Equation 5, $B_P(\vec{\theta})$ can be viewed as linear combinations of pose-blend shapes P learned from the datasets mentioned in section II. Each joint orientation θ is varied by rotating it 9 times relatively, therefore the iteration space is extended to the cardinality of $R_n(\vec{\theta})$ which is $|R_n(\vec{\theta})| = K \times 9 = 207$. In Equation 5, a second rotation set of the initial T model is also initialised to define the vertex deviations and make them linear to P. Thus, the impact of an pose blend shape is neglected when a joint orientation is equal to the T orientation.

$$\begin{aligned}
T_F(\vec{\beta}, \vec{\theta}) &: \mathbb{R}^{|\vec{\theta}| \times |\vec{\beta}|} \mapsto \mathbb{R}^{3N} \\
T_F(\vec{\beta}, \vec{\theta}) &= T + B_S(\vec{\beta}) + B_P(\vec{\theta}) \\
&= T + B_S(\vec{\beta}) + \sum_{n=1}^{9K}(R_n(\vec{\theta}) - R_n(\vec{\theta}^*))P_n;
\end{aligned} \quad (5)$$

Finally, the *linear blend skinning* in Equation 3 can be rewritten by the formulations in Equation 4 and their application in Equation 5. This Equation 6 is formulated below, considering both B_P and B_S as matrices in which a vertex is indexed by i. [44]

$$\bar{t}_i' = \sum_{k=1}^{K} w_{k,i} \times G_k'(\vec{\theta}, J(\vec{\beta})) \times (\bar{t}_i + b_{S,i}(\vec{\beta}) + b_{P,i}(\vec{\theta})) \quad (6)$$

D. combining models with pose estimation

The publication by Bogo et al. [5] in 2016 is the first application of SMPL in the field of human pose estimation. The *SMPLify* algorithm is based on optimization and gradually tries to align the mean initial SMPL human model to 2d keypoints of the initial image. A more recent approach is presented by Rong et al. [65] which starts from the concept of splitting the estimation of human pose into hand, face and total body, which are then merged into a final model. Each aspect utilises its own regressors, which are trained on datasets selected for the respective purpose. In further sections, the different network architectures are explained.

Fig. 5. encoder-decoder architectur from [65]

1) hand estimation: In their paper, Kanazawa et al. [35] describe an encoder-decoder structure in which the encoder recognises features from given images while the decoder regresses the feature parameters. This structure also has validity for the *FrankMocap* approach to estimating human posture

by Rong et al. shown in Figure 5. For the incorporation into *SMPL-X*, a successor of the mentioned human body model from subsubsection V-C1, the parameters $[\phi_h, \theta_h, \beta_h, c_h]$, i.e. the global orientation, pose and shape data of the hands, as well as the camera parameters are required. The annotations in the trained dataset are either axis-angle representations to indicate 3D positions and 3D or 2D vertices as common positions, and therefore each annotation has its unique loss function. To increase the robustness of the model, random scaling, rotation, etc. and a blur filter are applied to the dataset to be learned.
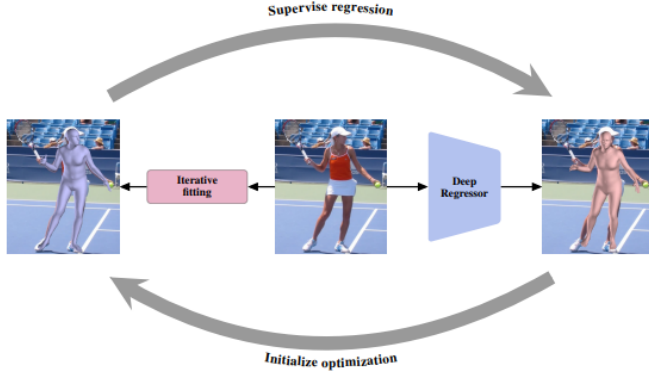
Fig. 6. visualisation of the SPIN loop [37]

2) full-body pose estimation: Similar to hand estimation, shape β_{reg}, pose θ_{reg} and camera parameters c_{reg} can also be regressed in whole body pose estimation approaches. The algorithm *SPIN* goes further by combining the advantages of optimization and regression-based implementations by using *SMPLify* and neural networks in loops. Substituting the parameters into Equation 4 yields a human output mesh. Furthermore, the joint positions can be projected using the information of c_{reg} defining the loss function for supervision. The self-impruving nature of *SPIN* can be observet in a small example. // In the first iteration, the parameters $[\theta_{reg}, \beta_{reg}, c_{reg}]$ are regressed through a neural network, *SMPLify* uses $M(\theta_{reg}, \beta_{reg})$ as the initial human model, simplifying the optimization-based aspect. After the final step of *SMPLifys*, both pose and shape parameters are optimal, leading to better loss function results through more precise joint positions and thus better supervision in the regression-based aspect. The loop is also showed in Figure 6. [37]

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.

[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] Bruno Artacho and Andreas Savakis. UniPose: Unified Human Pose Estimation in Single Images and Videos. Technical Report arXiv:2001.08095, arXiv, January 2020. arXiv:2001.08095 [cs] type: article.

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 561–578. Springer International Publishing, October 2016.

[6] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via Convolutional Part Heatmap Regression. volume 9911, pages 717–732. 2016. arXiv:1609.01743 [cs].

[7] Yuanhao Cai, Zhicheng Wang, Zhengxiong Luo, Binyi Yin, Angang Du, Haoqian Wang, Xiangyu Zhang, Xinyu Zhou, Erjin Zhou, and Jian Sun. Learning Delicate Local Representations for Multi-Person Pose Estimation. Technical Report arXiv:2003.04030, arXiv, July 2020. arXiv:2003.04030 [cs] type: article.

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. Technical Report arXiv:2005.12872, arXiv, May 2020. arXiv:2005.12872 [cs] type: article.

[9] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human Pose Estimation with Iterative Error Feedback. Technical Report arXiv:1507.06550, arXiv, June 2016. arXiv:1507.06550 [cs] type: article.

[10] 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.

[11] 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.

[12] Xianjie Chen and Alan Yuille. Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations. Technical Report arXiv:1407.3399, arXiv, November 2014. arXiv:1407.3399 [cs] type: article.

[13] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded Pyramid Network for Multi-Person Pose Estimation. Technical Report arXiv:1711.07319, arXiv, April 2018. arXiv:1711.07319 [cs] type: article.

[14] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. Technical Report arXiv:1908.10357, arXiv, March 2020. arXiv:1908.10357 [cs, eess] type: article.

[15] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L. Yuille, and Xiaogang Wang. Multi-Context Attention for Human Pose Estimation. Technical Report arXiv:1702.07432, arXiv, February 2017. arXiv:1702.07432 [cs] type: article.

[16] Yicheng Deng, Yongqi Sun, and Jiahui Zhu. SVMA: A gan-based model for monocular 3d human pose estimation. *CoRR*, abs/2106.05616, 2021.

[17] 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.

[18] M. Eichner and V. Ferrari. Human Pose Co-Estimation and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[19] 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.

[20] Xiaochuan Fan, Kang Zheng, Yuewei Lin, and Song Wang. Combining Local Appearance and Holistic View: Dual-Source Deep Neural Networks for Human Pose Estimation. Technical Report arXiv:1504.07159, arXiv, April 2015. arXiv:1504.07159 [cs] type: article.

[21] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial Structures for Object Recognition. *International Journal of Computer Vision*, 61(1):55–79, January 2005.

[22] Mihai Fieraru, Anna Khoreva, Leonid Pishchulin, and Bernt Schiele. Learning to Refine Human Pose Estimation. Technical Report arXiv:1804.07909, arXiv, April 2018. arXiv:1804.07909 [cs] type: article.

[23] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. Detect-and-Track: Efficient Pose Estimation in Videos. Technical Report arXiv:1712.09184, arXiv, May 2018. arXiv:1712.09184 [cs] type: article.

[24] Georgia Gkioxari, Alexander Toshev, and Navdeep Jaitly. Chained Predictions Using Convolutional Neural Networks. Technical Report arXiv:1605.02346, arXiv, October 2016. arXiv:1605.02346 [cs] type: article.

[25] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. ISSN: 1063-6919.

[32] 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.

[33] 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.

[34] 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.

[35] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131. IEEE Computer Society, 2018.

[36] Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. Multi-Scale Structure-Aware Network for Human Pose Estimation. Technical Report arXiv:1803.09894, arXiv, September 2018. arXiv:1803.09894 [cs] type: article.

[37] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[39] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose Recognition with Cascade Transformers. Technical Report arXiv:2104.06976, arXiv, April 2021. arXiv:2104.06976 [cs] type: article.

[40] Sijin Li, Zhi-Qiang Liu, and Antoni B. Chan. Heterogeneous Multi-task Learning for Human Pose Estimation with Deep Convolutional Neural Network. Technical Report arXiv:1406.3474, arXiv, June 2014. arXiv:1406.3474 [cs] type: article.

[41] Ita Lifshitz, Ethan Fetaya, and Shimon Ullman. Human Pose Estimation using Deep Consensus Voting. Technical Report arXiv:1603.08212, arXiv, March 2016. arXiv:1603.08212 [cs] type: article.

[42] 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.

[43] Wentao Liu, Jie Chen, Cheng Li, Chen Qian, Xiao Chu, and Xiaolin Hu. A Cascaded Inception of Inception Network With Attention Modulated Feature Fusion for Human Pose Estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. Number: 1.

[44] 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.

[45] Yue Luo, Jimmy Ren, Zhouxia Wang, Wenxiu Sun, Jinshan Pan, Jianbo Liu, Jiahao Pang, and Liang Lin. LSTM Pose Machines. pages 5207–5215, 2018.

[46] Diogo C. Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. *CoRR*, abs/1802.09232, 2018.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] Antonio S. Micilotta, Eng-Jon Ong, and Richard Bowden. Real-Time Upper Body Detection and 3D Pose Estimation in Monoscopic Images. In Aleš Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision – ECCV 2006*, Lecture Notes in Computer Science, pages 139–150, Berlin, Heidelberg, 2006. Springer.

[52] 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.

[53] 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.

[54] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked Hourglass Networks for Human Pose Estimation. Technical Report arXiv:1603.06937, arXiv, July 2016. arXiv:1603.06937 [cs] type: article.

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

[56] 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.

[57] 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.

[58] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing ConvNets for Human Pose Estimation in Videos. Technical Report arXiv:1506.02897, arXiv, November 2015. arXiv:1506.02897 [cs] type: article.

[59] Lingteng Qiu, Xuanye Zhang, Yanran Li, Guanbin Li, Xiaojun Wu, Zixiang Xiong, Xiaoguang Han, and Shuguang Cui. Peeking into occluded joints: A novel framework for crowd pose estimation. Technical Report arXiv:2003.10506, arXiv, March 2020. arXiv:2003.10506 [cs] type: article.

[60] Umer Rafi, Bastian Leibe, Juergen Gall, and Ilya Kostrikov. An Efficient Convolutional Network for Human Pose Estimation. In *Procedings of*

the British Machine Vision Conference 2016, pages 109.1–109.11, York, UK, 2016. British Machine Vision Association.

[61] Varun Ramakrishna, Daniel Munoz, Martial Hebert, James Andrew Bagnell, and Yaser Sheikh. Pose Machines: Articulated Pose Estimation via Inference Machines. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 33–47, Cham, 2014. Springer International Publishing.

[62] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Technical Report arXiv:1506.01497, arXiv, January 2016. arXiv:1506.01497 [cs] type: article.

[63] 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.

[64] 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.

[65] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021.

[66] 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.

[67] Kai Su, Dongdong Yu, Zhenqi Xu, Xin Geng, and Changhu Wang. Multi-Person Pose Estimation with Enhanced Channel-wise and Spatial Information. Technical Report arXiv:1905.03466, arXiv, May 2019. arXiv:1905.03466 [cs] type: article.

[68] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep High-Resolution Representation Learning for Human Pose Estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5686–5696, Long Beach, CA, USA, June 2019. IEEE.

[69] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional Human Pose Regression. Technical Report arXiv:1704.00159, arXiv, August 2017. arXiv:1704.00159 [cs] type: article.

[70] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral Human Pose Regression. Technical Report arXiv:1711.08229, arXiv, September 2018. arXiv:1711.08229 [cs] type: article.

[71] 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.

[72] 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.

[73] 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.

[74] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christopher Bregler. Efficient Object Localization Using Convolutional Networks. Technical Report arXiv:1411.4280, arXiv, June 2015. arXiv:1411.4280 [cs] type: article.

[75] Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. Technical Report arXiv:1406.2984, arXiv, September 2014. arXiv:1406.2984 [cs] type: article.

[76] 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.

[77] Alexander Toshev and Christian Szegedy. DeepPose: Human Pose Estimation via Deep Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, June 2014. arXiv:1312.4659 [cs].

[78] 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.

[79] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. Technical Report arXiv:1706.03762, arXiv, December 2017. arXiv:1706.03762 [cs] type: article.

[80] 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.

[81] Jian Wang, Xiang Long, Yuan Gao, Errui Ding, and Shilei Wen. Graph-PCNN: Two Stage Human Pose Estimation with Graph Pose Refinement. Technical Report arXiv:2007.10599, arXiv, July 2020. arXiv:2007.10599 [cs] type: article.

[82] 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.

[83] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional Pose Machines. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, Las Vegas, NV, USA, June 2016. IEEE.

[84] Bin Xiao, Haiping Wu, and Yichen Wei. Simple Baselines for Human Pose Estimation and Tracking. Technical Report arXiv:1804.06208, arXiv, August 2018. arXiv:1804.06208 [cs] type: article.

[85] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning Feature Pyramids for Human Pose Estimation. Technical Report arXiv:1708.01101, arXiv, August 2017. arXiv:1708.01101 [cs] type: article.

[86] 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.

[87] 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.

[88] Yixing Gao, Hyung Jin Chang, and Yiannis Demiris. User modelling for personalised dressing assistance by humanoid robots. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1840–1845, Hamburg, Germany, September 2015. IEEE.

[89] 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.