The Journal of Supercomputing (2022) 78:7616–7654
https://doi.org/10.1007/s11227-021-04184-7

Human pose, hand and mesh estimation using deep learning: a survey

Mukhiddin Toshpulatov[1] · Wookey Lee[1] · Suan Lee[2] · Arousha Haghighian Roudsari[3]

Accepted: 19 October 2021 / Published online: 1 January 2022

Abstract
Human pose estimation is one of the issues that have gained many benefits from using state-of-the-art deep learning-based models. Human pose, hand and mesh estimation is a significant problem that has attracted the attention of the computer vision community for the past few decades. A wide variety of solutions have been proposed to tackle the problem. Deep Learning-based approaches have been extensively studied in recent years and used to address several computer vision problems. However, it is sometimes hard to compare these methods due to their intrinsic difference. This paper extensively summarizes the current deep learning-based 2*D* and 3*D* human pose, hand and mesh estimation methods with a single or multi-person, single or double-stage methodology-based taxonomy. The authors aim to make every step in the deep learning-based human pose, hand and mesh estimation techniques interpretable by providing readers with a readily understandable explanation. The presented taxonomy has clearly illustrated current research on deep learning-based 2*D* and 3*D* human pose, hand and mesh estimation. Moreover, it also provided dataset and evaluation metrics for both 2*D* and 3*D HPE* approaches.

Keywords 3*D* pose estimation · Generator · Discriminator · Loss function · Deep neural network · Deep learning · Mesh estimation · Evaluation metric · Dataset

1 Human pose estimation

Human pose estimation (*HPE*) recently has been significantly studied in the *AI* research community. *HPE* aims to obtain the posture of the human body from given sensor inputs. *HPE* is a crucial research study in the modern computer vision field and has been implemented into many applications, such as human–computer interaction (*HCI*) [14, 20], healthcare, motion analysis, virtual reality (VR) and augmented

✉ Wookey Lee
trinity@inha.ac.kr

Extended author information available on the last page of the article

reality (AR). Novel deep learning (DL) models outperformed classical methods in several research fields such as image classification, semantic segmentation, object and face detection. Subsequently, the *HPE* has also achieved outstanding achievement utilizing the *DL* methods. With the rapid advancement in the *HPE* research field, this work tracks recent progress and summarizes their achievements to provide readers with a clear understanding of current research on *DL*-based *HPE*.

Several review works are published on the 2*D* and 3*D HPE* topics. The authors of [20], in their survey study, presented only 2*D HPE* methods, including single- or multi-person approaches. In [14], the authors provide a review on 2*D* and 3*D HPE* methods with their related categories. Zheng et al. [124] also covers a similar context with [124]. However, all the mentioned studies include papers from the early DL implementation to the *HPE* field. Unlike the existing survey papers, this research work thoroughly summarizes the recently published *DL*-based 2*D HPE* and 3*D HPE*. One of the key points of the current research that previous ones did not includes 3*D* human hand, and mesh estimation approaches, which are important for a high rapidly growing *AI* applications. We have used the novel taxonomy, which differs from the previous ones, and it covers state-of-the-art 3*D* human hand with mesh estimation approaches. We have also discussed several recently published papers in great detail, which revolutionized in the related research field. This research addresses the weaknesses of the existing survey studies in the *HPE* field, covering the key points of *HPE* methods, advantages and drawbacks, extensive analysis of their experimental implementation details. We have reviewed the recent *HPE* milestone, which the previous survey studies did not cover. In addition, it includes recently published *DL*-based 3*D* human hand and mesh estimation approaches, which are rapidly growing and gaining a great attraction among the *AI* researchers.

In this survey study, we have reviewed and discussed the recently published research works in the related field in two main divisions: 2*D HPE* (Sect. 2) and 3*D* human pose estimation (Sect. 3). Each of them is also divided into subcategories based on their respective characteristics. Table 1 shows a taxonomy of all the reviewed papers during this research process by their related categories.

Our contribution through this survey study and advantages of the research work from the previous similar surveys are: Recently published novel *DL*-based 2*D HPE* and 3*D HPE* methods including 3*D* human hand and mesh estimation approaches are extensively reviewed; Provided a taxonomy of all reviewed approaches by a category corresponding to 2*D* single or multiple *HPE* and 3*D* single or multiple *HPE*, covering single or double stage, model-based or model-free subcategories. Provided extensive performance evaluation of 2*D HPE* and 3*D* human hand, pose, mesh estimation approaches; Described mainly used datasets and widely used evaluation metrics used in 2*D HPE* and 3*D* human hand, pose, mesh estimation; Reviewed various types of human pose, hand and mesh estimation applications, such as computer gaming, video surveillance, movies and animation, human and computer interaction, self-driving, *AR*/*VR*, medical assistance and healthcare; Presented vital points of the state-of-the-art 2*D HPE* and 3*D* human pose, hand and mesh estimation approaches extensively compared their pros and cons, input and output data, used dataset, backbone, and loss function experimental implementation details and evaluation measures. Also, it presented an insightful discussion of 2*D HPE* and 3*D* human pose,

Table 1 Sections of deep learning-based human pose estimation approaches [14]

Division	Sub-division	Section	Sub-section
2*D HPE*	2*D* Single	Regression-based	− Direct prediction − Multi-task
		Detection-based	− Network design
	2*D* Multiple	Top-down	− Coarse-to-fine − Bounding box refinement
		Botttom-up	− Two-stage − Single-stage − Multi-task
3*D HPE*	3*D* Single	Model-free	− Single-stage − Two-stage
		Model-based	− SMPL-based
		Depth-based	− Generative − Discriminative
	3*D* Multiple		− Bottom-up − Top-down − SMPL-based

hand and mesh estimation methods regarding open research issues with key challenges in the *HPE* field, including future research direction.

The paper is organized as follows. Section 1 Human Pose Estimation describes the briefly introduction to the research field with existing problems and includes Sect. 1.1 Recent advancements, Sect. 1.2 Impactful Past Papers and Sect. 1.3 Traditional Human body models. Section 2 extensively describes 2*D HPE* approaches, dividing them into related categories and given a detailed description of some state-of-the-art methods. Section 3 covers 3*D HPE* method including single and multi-view approaches. It also describes in detail several novel methods which made the revolution in the research field. Section 4 discusses 3*D* human hand and mesh estimation approaches in detail, as hand and mesh estimation is important and gaining tremendous interest among the researchers in the *AI* community. The Datasets and Evaluation metrics used in the related field are given and discussed with great details in Sect. 5. Section 6 presents Open issues and challenges in the *HPE* research community while Sect. 7 gives Summary and Conclusion of this survey study. And Sect. 8 gives the Future research directions in *HPE* research field.

1.1 Recent advancements

Despite the significant progress and remarkable performance of *HPE*, challenges such as occlusion, lacking training data and the depth ambiguity still cause difficulties that need to be overcome. Moreover, compared with 2*D HPE*, obtaining precise 3*D* pose annotations is much more complicated. For 3*D HPE* from 2*D* data, the main difficulty is depth ambiguities. Researchers have employed inertial measurement units (*IMU*s), depth sensors and radiofrequency devices as a solution.

However, these methods are usually not considered as being cost-effective and need special-purpose hardware [124]. A novel Bayesian formulation of Capsule networks [88, 93] was implemented for estimating the 3*D* human pose from a single *RGB* image. The obtained result was that the pose is given by the 3*D* coordinates of 17 joints in a human pose skeleton. 3*D* pose restoration from 2*D* input is an ill-posed optimization problem and should be regularized. They chose J capsules with size S as 512 and 8, respectively. Each capsule is cloned $K = 17$ times to predict 17 human pose joints. Another fascinating recent approach is Deep High-Resolution Representation Learning for *HPE*, where the authors proposed that the network maintains high-resolution representations through the whole process [44]. The proposed model starts from a high-resolution subnetwork as the initial stage. It then gradually adds high-to-low-resolution subnetworks one by one to form more stages and connects the multi-resolution subnetworks in parallel. They repeat these multi-scale unions. Each high-to-low resolution representation receives information from other parallel representations, leading to rich, high-resolution representations. The authors argue that the predicted key point heatmap is potentially more accurate and spatially more precise. Moreover, they have empirically demonstrated the effectiveness of the proposed approach through the superior pose estimation results over the *COCO* and the *MPII* dataset.

Multi-person *HPE* is an attractive and compared with the single-person *HPE* is a challenging task. Existing methods are mainly based on two-stage and generally suffer from low efficiency. A single-stage-based model, namely Single-stage multi-person Pose Machine(*SPM*) [71], was proposed to simplify the pipeline and enhance the efficiency for multi-person human pose estimation. The authors propose a Structured Pose Representation (*SPR*) to unify human body instance and joint positions. The developed *SPM* based on *SPR* directly predicts structured poses for multi-person in a single stage. Moreover, the proposed approach offers a more compact pipeline and an attractive efficiency advantage than previous state-of-the-art ones. Even though the single-stage paradigm aims to simplify the multi-person pose estimation and receives much attention, they still have low performance due to the difficulty of regressing various full-body poses from a single feature vector. Unlike previous solutions involving complex heuristic designs, Shi et al. [97] presented a simple and effective solution by employing instance-aware dynamic networks. Specifically, they propose an instance-aware module to adaptively adjust (part of) the network parameters for each instance. The authors argue that the proposed approach can significantly increase the capacity and adaptive-ability of the network for recognizing various human poses while maintaining a compact end-to-end trainable pipeline. The extensive experiments on the *MS-COCO* dataset significantly improve existing single-stage methods and make a better balance of accuracy and efficiency compared to the state-of-the-art two-stage *HPE* approaches.

1.2 Impactful past papers

HPE is defined as the problem of localization of human joints [122] in images or videos or searching for a specific pose in the space of all articulated poses, such

as illustrated in Fig. 1. It has already been widely exploited in *Action recognition*, *Animation*, *Gaming* applications, such as a very popular *Deep Learning* app *HomeCourt*, which uses Pose Estimation to analyze Basketball player movements. Realistic 3*D HPE* aims to localize semantic key points [3] of single or multiple human bodies in 3*D* space. It is an crucial element for human behavior understanding, activity recognition with various applications, such as augmented reality or human-computer interaction. Even though it has been studied for decades in the computer vision field, it has attracted significant research interest due to the introduction of low-cost depth cameras [55, 75, 117].

Convolutional neural networks (*CNNs*) [59, 68, 76, 89, 106, 108]-based methods outperform existing ones in *HPE* from a single depth map and achieved noticeable performance improvement. Even though they achieved significant advancement in 3*D HPE*, they still suffer from inaccurate analysis because of severe self-occlusions, highly articulated shapes of target objects, and low-quality depth images. To overcome these issues, Moon et al. [67] proposed the voxel-to-voxel prediction network for pose estimation (*V2V-PoseNet*). The proposed method takes a voxelized grid as input and estimates the per-voxel likelihood for each key point. By converting the 2*D* depth image into a 3*D* voxelized form as input, the network can see objects' actual appearance without perspective distortion. Moreover, estimating each key point's per-voxel likelihood enables the system to learn the desired task more quickly than the highly nonlinear mapping that estimates 3*D* coordinates directly from the input. *HPE* is a vital research field and can be applied to various applications such as action/activity detection, action recognition [41, 54, 120], human tracking [19,

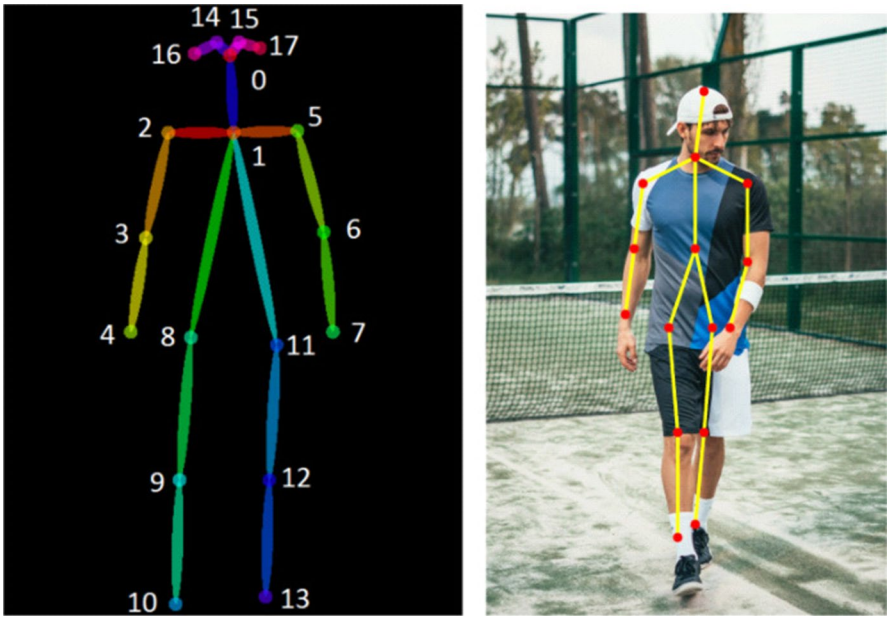

Fig. 1 Pose estimation [3]

39], virtual reality, video surveillance [2], movies and animation, human–computer interaction [85], self-driving, medical assistance [107] and sports motion analysis. Table 2 shows and describes a list of the applications that can be done with the pose estimations.

1.3 Human body models

Modeling of the human body is a key component of *HPE*. The human body is a flexible and complex non-rigid object. It has many specific characteristics like kinematic structure, surface texture, body shape, the position of body parts or body joints, etc. A mature model for the human body does not necessarily need to contain all human body attributes. When building the human pose model, only the requirements for a specific task need to be met. Based on various application scenarios and levels of representations required, three kinds of human body models are commonly used in *HPE*. As presented in Fig. 2, these three types include skeleton based, contour based and volume based.

Kinematic model The kinematic model includes a set of joint positions and limb orientations to represent the human body structure. The Pictorial Structure Model (*PSM*) is a widely used graph model, also known as the tree-structured model. This flexible and intuitive human body model is successfully utilized in 2*D HPE* and 3*D*

Table 2 A list of the applications that can be done with the pose estimations [14]

Application	Description
Virtual reality	Promising technology that can be applied in both education and entertainment. Estimation of human posture can further clarify the relationship between the social and virtual reality world and enhance the interactive experience
Video surveillance	One of the early applications to adopt *HPE* technology in tracking, action recognition, re-identification of people within a specific range
Movies and animation [42]	Generation of various vivid digital characters is inseparable from the capture of human movements. A cheap and accurate social motion capture system can better promote the digital entertainment development industry
Human–computer interaction	*HPE* is very important for computers and robots better to understand people's identification, location, and action. With humans' posture, computers and robots can efficiently execute instructions and be more intelligent
Self-driving	Advanced self-driving cars with *HPE* can respond more appropriately to pedestrians and offer more comprehensive interaction with traffic coordinators
Medical assistance	*HPE* can provide physicians with quantitative human motion information, especially for rehabilitation training and physical therapy
Sports motion analysis	Estimating players' posture in sports videos can further obtain the statistics of athletes' indicators. *HPE* can provide a quantitative analysis of action details. Instructors can make more objective evaluations of students with *HPE*

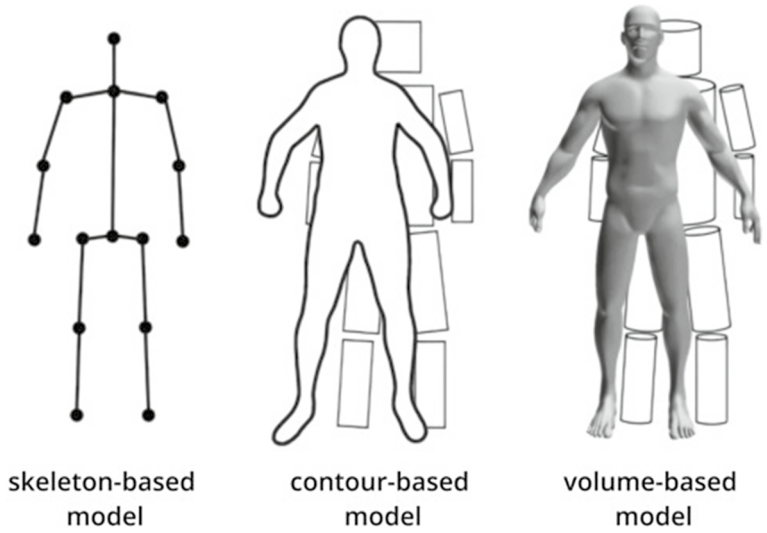

Fig. 2 Three types of models for human body modeling [124]

HPE. Moreover, the kinematic model has the advantage of flexible graph representation even though it is limited in representing texture and shape information.

Planar Model Besides the kinematic model, to capture the relations between different body parts, the planar model represents the shape and appearance of a human body. In the planar model, body parts are usually represented by rectangles approximating the human body contours.

Volumetric Model With the increasing interest in 3*D HPE*, many human body models have been proposed for a wide variety of human body shapes. It is commonly used in deep learning-based 3*D HPE* methods for recovering 3*D* human mesh.

SMPL The skinned Multi-Person Linear model (*SMPL*) is a skinned vertex-based model, which depicts a wide range of human body shapes. It can be modeled with natural pose-dependent deformations representing soft-tissue dynamics. There are 1786 high-resolution 3*D* scans of various poses, including template mesh in *SMPL*. These 3*D* scans are used to learn how people deform with various poses by optimizing the blend weights, the mean template shape, pose-dependent blend shapes, and the regressor from vertices to joint locations. The existing rendering engines are compatible with *SMPL* and easy to utilize; therefore, it is broadly adopted in 3*D HPE* methods.

DYNA Dynamic Human Shape in Motion (*DYNA*) model represents realistic soft-tissue motions for different human body shapes. A low-dimensional linear subspace approximates motion-related soft-tissue deformation. For predicting the low-dimensional linear coefficients of soft-tissue motion, the whole body's velocity and acceleration, the body part's angular velocities and accelerations, and the coefficients of the soft-tissue shape are used. *DYNA* utilizes the body mass index (*BMI*) to generate deformations for different shaped people.

2 2*D* human pose estimation approaches

Some solution methods of the 2*D HPE* task have been reviewed in this section. 2*D* Human pose estimation has received significant attention recently due to its various applications in the real world. It aims to locate the human body parts from images or videos automatically. All the human parts in the given input image or video should be detected, and the key points of the same person, even in a crowded scene, should be associated. *DL*-based 2*D HPE* methods are capable of extracting more sufficient features from metadata. Such methods have resulted in excellent performances and outperformed the non-deep state-of-the-art methods with a considerable margin. Although the deployment of *DL* in the *HPE* field is relatively new, several notable works on this issue have been proposed. This section aims to present a comprehensive overview of state-of-the-art *DL*-based 2*D* human pose estimation methodologies.

2.1 2*D* single-person pose estimation

2*D* Single-person *HPE* refers to the task of localizing human skeletal key points of a person from the given input image or video frame. There have been proposed several traditional research approaches where was exploited handcrafted feature extraction methods—developing the novel *DL* technique has been widely implementing into the 2*D HPE* research field also. *DeepPose* [105] was the first significant paper that applied Deep Learning to 2*D HPE*. It achieved *SOTA* performance and exceeded the existing models. *DeepPose* formulates the pose estimation as a CNN-based regression task toward body joints. Also, the pose estimates are refined by using a cascade of regressors to obtain a better result. The proposed approach does pose reasoning in a holistic manner. That is, even if certain joints are hidden, they can be estimated. The authors argue that *CNNs* naturally provide this sort of holistic reasoning and demonstrate strong results. The key point of the proposed model is implementing the refinement of the predictions using cascaded regressors. The initial coarse pose is refined, and a better estimate is achieved. Images are cropped around the predicted joint and fed to the next stage. In this way, the subsequent pose regressors see higher resolution images and thus learn finer scales, which ultimately leads to higher precision. In [105], the Cartesian coordinates of body joints are directly estimated using a multi-stage deep network and produced state-of-the-art achievement. Multi-stage *CNN* also progressively enlarges receptive fields and refines the pose estimation result. In addition, it is trainable with a graphical model. The *CNN* estimated 2*D* heatmaps [63] for each joint, and they were exploited as the unary term for the model. In [70], a stacked hourglass network was proposed, which repeats downsampling and upsampling to exploit multi-scale information effectively. Chu et al. [18] attempted to enhance the stacked hourglass network [70] by integrating it with a multi-context attention mechanism. In [13], an iterative error feedback-based *HPE* system was proposed. Ke et al. [43] presented a multi-scale structure-aware network to achieve

a leading position in the publicly available *HPE* benchmark. *Mask R-CNN* [17] was proposed to perform human detection and key point localization in a single model. The proposed model crops human features from a feature map via the differentiable *RoIAlign* layer. The schematic view of the proposed *DeepPose* system is represented in Fig. 3. It consists of an *AlexNet* backed (8 layers) with an extra final layer, which outputs $2n$ coordinates—$(x_i, y_i) * 2$, where $i \in \{1, 2, \ldots, n\}$, and n is total number of joints. For training, the model uses L_2 loss for regression and applies refinement of the predictions using cascaded regressors, resulting in achieving better estimates. Images are cropped around the predicted joint and fed to the next stage; in this way, the subsequent pose regressors see higher resolution images and thus learn features for finer scales which ultimately leads to higher precision.

2.1.1 Regression-based methods

There have been proposed several research works related to regression model predict human joint coordinates from the input image or video frame. *AlexNet* is one of the initial networks for deep learning-based 2*D HPE* approaches due to their simplistic architecture and remarkable performance. It was first trained to learn joint coordinates from full input images in a very straightforward manner. A cascade structure of multi-stage refining regressors was employed to refine the previous stage's cropped images and showed improved performance. It was also applied for predicting the human pose in the videos using a sequence of concatenated frames as an input. Networks handling multiple closely related tasks of the human body may learn various features to improve the prediction of joint coordinates of the human pose. The *AlexNet* multi-task framework was also employed to handle the joint coordinate prediction task from the given input images regression. We briefly describe the layers of the *AlexNet* network as it plays a significant role in our research domain.

Figure 4 shows that the *Alexnet* has eight layers with learnable parameters. It consists of five layers with a combination of max-pooling layers followed by three fully connected layers and use a *Relu* activation in each of these layers except the output one. It was found that using the *Relu* as an activation function speeds up the training by almost six times. It also uses the dropout layers that prevent overfitting. The model is trained on the *ImageNet* dataset, with 14 million images.

Fig. 3 Schematic view of the proposed *DeepPose* system [105]

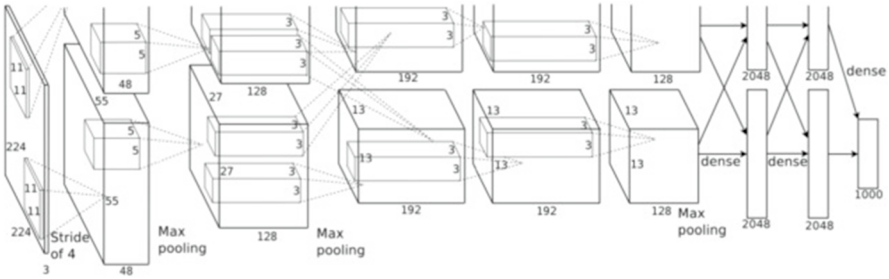

Fig. 4 *AlexNet* system [20]

2.2 2*D* multi-person pose estimation

The multi-person pose estimation has two main approaches: Top-down approach and Bottom-up approach (Fig. 5).

Top-down approach The top-down approach relies on a human detector that predicts bounding boxes of humans. The detected human image is cropped and fed to the pose estimation system. In other words, first is a person detection stage followed by the key point detection stage. These approaches still dominate the leaderboard of public benchmark datasets like MS *COCO*10 dataset [52, 78] and can be summarized based on the following aspects:

- Context modeling
- Effective training strategy
- Post-processing techniques

[15, 34, 80, 102, 112, 115] researches are based on the top-down approach. Chen et al. [15] introduced a cascaded pyramid network whose cascaded structure refines an initially estimated pose by focusing on hard key points. Xiao et al. [115] proposed a simple pose estimation network that consists of a deep backbone network and several upsampling layers. This model achieved state-of-the-art performance based on simple network architecture on the commonly used benchmark. Papandreou et al. [81] proposed 2*D* offset vectors and 2*D* heatmaps for each joint. They fused the estimated vectors and heatmaps to generate highly localized heatmaps.

Bottom-up approach The bottom-up approach localizes all human body key points in an input image and assembles them using proposed clustering algorithms in each work. In other words, this approach directly detects all key points from the picture and associates them with similar person occurrences. Bottom-up approaches are usually faster than top-down methods. [12, 36, 48, 69, 70] works are based on the bottom-up approach.

A novel method called *DeepCut* [84] formulated the assignment of the detected key points to each person in a given input image as an integer linear program. It improves the performance by introducing image-conditioned pair-wise terms. Part affinity fields (*PAFs*) [12] exposed the association between human body key points

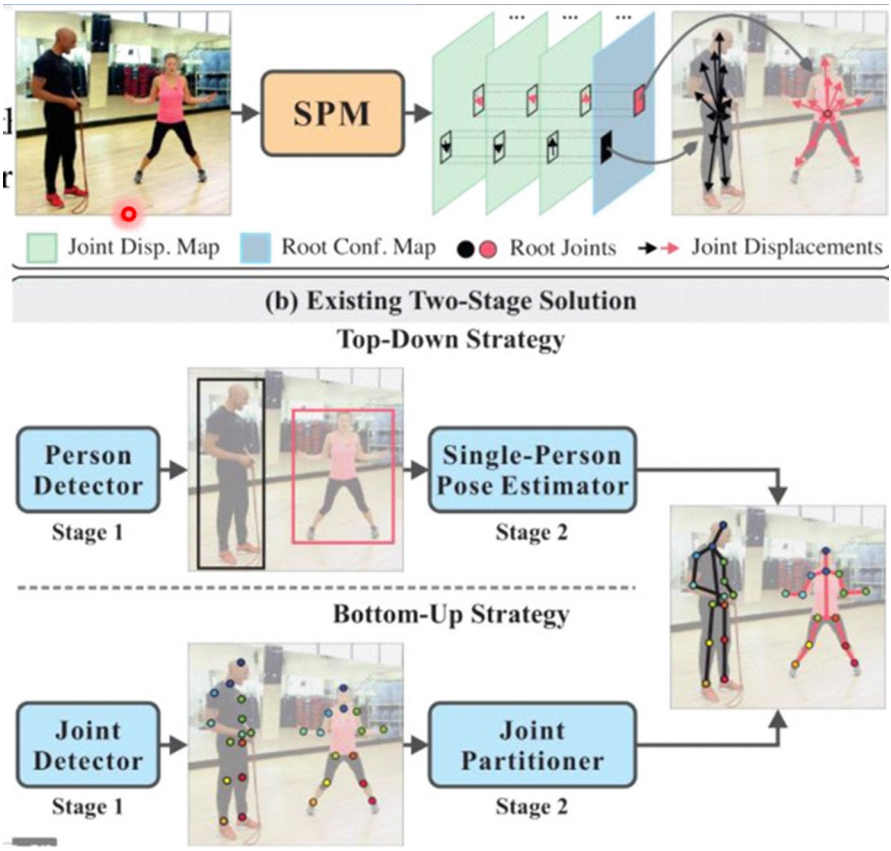

Fig. 5 Two-staged approaches of multi-person pose estimation [71]

directly. Authors assembled the localized key points of all the people in the input image by using the estimated *PAFs*. Newell et al. [69] proposed a pixel-wise tag value to assign localized key points to a certain human. Pose residual network (*PRN*) [48] is a pose estimation model that can assign detected key points to each person while having the ability to jointly perform key point detection, person detection and person segmentation. Chen et al. [15] proposed a cascaded pyramid network (*CPN*) which consists of two structures: *GlobalNet*; *RefineNet*.

GlobalNet GlobalNet is based on a deep backbone network and upsampling layers with skip connections (Fig. 5).

RefineNet RefineNet is built to refine the estimation results from the *GlobalNet* by focusing on hard key points.

2.3 2*D* human pose refinement

Many methods endeavored to refine the approximated key point for more realistic representation. Chen et al. [15], Newell et al. [70], Bulat and Tzimiropoulos [9–11] exploited an end-to-end trainable multi-stage architecture-based network. The utilized model at each stage tries to refine the pose estimation results of the previous stage via end-to-end learning. The model proposed in [13] iteratively estimated error feedback from a shared weight model. The previous iteration's output error feedback is transformed into the input pose of the next iteration, which is repeated several times for progressive pose refinement. These methods combine pose estimation and refinement into a single model. The refinement module is dependent on the estimation, and models have a refinement module with a different structure. Hence, they are not guaranteed to work appropriately, combining with other estimation methods. Moon et al. [65] proposed a pose refinement method independent of the estimation, where the results can be consistently improved regardless of the prior pose estimation method. Fieraru et al. [25] proposed a post-processing network to refine the pose estimation results of other methods, which is conceptually similar to [65]. The proposed model synthesizes pose for training and uses simple network structure that estimates refined heatmaps and offset vectors for each joint. It follows ad hoc rules [8, 96] to generate input pose, while the previous [65] approach is based on actual error statistics obtained through empirical analysis.

In [65], a refinement network *PoseFix* was proposed to estimate a refined pose from a tuple of an input image and a human pose. It takes pose estimation results of any other method with an input image and outputs an elegant pose. Multi-stage architectures have mainly performed pose refinement. However, this approach is positively related to the pose estimation model and requires careful refinement design. The authors proposed a model-agnostic pose refinement method that does not depend on the pose estimation model. The proposed model takes the input pose in a coarse form and estimates the refined pose in a finer form. The coarse input pose enables the model to focus not only on an exact location of the input pose but also around it. Besides, the finer form of the output pose enables to localize the location of the pose. *PoseFix* can be applied to the pose estimation results of any single- or multi-person pose estimation method. Figure 6 shows a pose refinement pipeline of the *PoseFix*.

The *PoseFix* model refines the input 2*D* coordinates of all the persons' human body key points in an input image. It is built based on the top-down pipeline, which processes a cropped human image's tuple and a given pose estimation result of that person. In the training stage, the input pose is synthesized on the ground-truth pose realistically and diversely. In the testing stage, the pose estimation results of any other methods can be the input pose to the system. The overall pipeline of the *PoseFix* is described in Fig. 7.

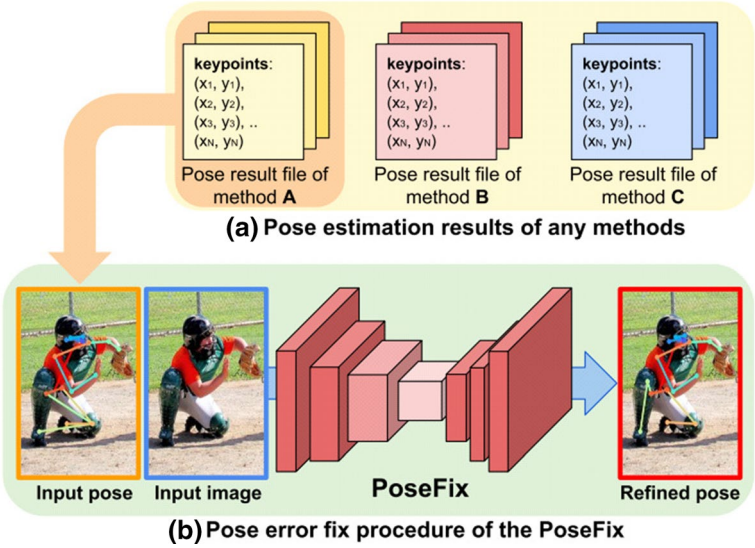

Fig. 6 Testing pipeline of the *PoseFix* [65]

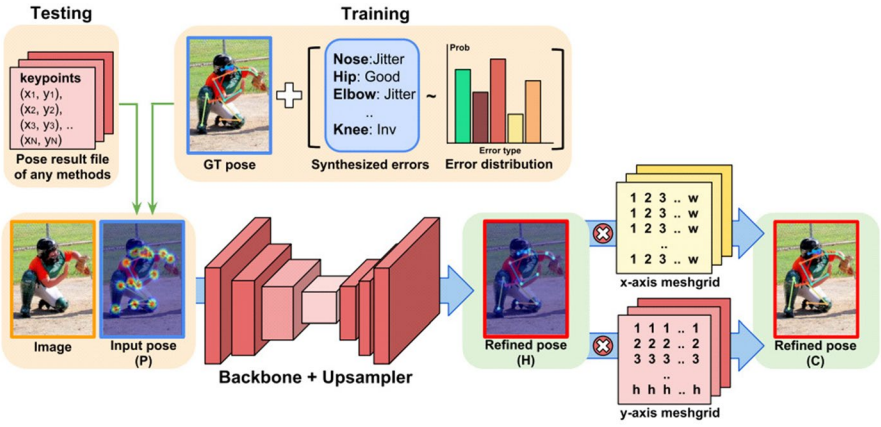

Fig. 7 Overall pipeline of the *PoseFix* [65]

3 3D human pose estimation approaches

3*D HPE* predicts locations of the human body joints in 3*D* space from the given input image sources. The progress of 3*D HPE* from the input data can further improve the multi-view 3*D HPE* in constrained environments. This section will focus on the *DL*-based methods that estimate 3*D* human pose from the 2*D RGB* images and videos, including 3*D* single-person and 3*D* multi-person pose estimation approaches. Most of the previous 3*D HPE* methods are designed for a single-person

case. They crop the human area in an input image with a ground-truth bounding box that is predicted from a human detection model. The cropped patch of a human body is fed into the 3*D* pose estimation module, which then estimates each key point's 3*D* location. Estimating the absolute camera-centered coordinate of each key point is difficult because the models take the cropped image as input. Many approaches estimate the relative 3*D* pose to reference point in the body to manage this problem. The final 3*D* pose is obtained by adding the 3*D* coordinates of the root to the estimated root-relative 3*D* pose. Prior information on the bone length [83] or the ground-truth [102] has been commonly used to localize the root.

3.1 3*D* single-person pose estimation

3*D HPE* is challenging as it needs to predict the depth information of human body joints. Preparation of the training data is also not easy. The existing datasets for 3*D HPE* are obtained under constrained environments. In 3*D* single person pose estimation cases, the bounding box of the person is usually provided. 3*D* single person pose estimation is divided mainly into two: model-free and model-based categories.

Model-free methods The model-free 3*D* single-person pose estimation methods do not employ human body models as the predicted target. According to the input type, Current 3*D* single-person pose estimation methods can be categorized into three approaches: single-stage approach; two-stage approach; depth-based approach.

Single-stage approach The single-stage approach takes an *RGB* image as an input for the 3*D* human pose estimation model and directly localizes the 3*D* body key points from the input data. The compositional loss was proposed [100] to consider the joint connection structure. Soft-argmax operation was exploited [102] to obtain the 3*D* coordinates of body joints in a differentiable manner. Moreover, they introduced a multi-task framework that jointly trains both the pose regression and body part detectors. Tekin et al. [104] modeled high-dimensional joint dependencies by adopting an auto-encoder structure. Pavlakos et al. [83] extended the *U-net* shaped network to estimate a 3*D* heatmap for each joint. They used a coarse-to-fine approach to boost the performance. Martinez et al. [58] proposed a simple network that consists of consecutive fully-connected layers, which lifts the 2*D* human pose to the 3*D* space. Sharma et al. [95] combined a generative model and depth ordering of joints to predict the most reliable 3*D* pose corresponding to the estimated 2*D* pose. The 2*D* pose-based approach lifts the 2*D* human pose to the 3*D* space. Zhao et al. [123] generated a semantic *GraphCNN* to use spatial relationships between joint coordinates. Choi et al. [16] follow the 2*D* pose-based approach to make the *Pose2Mesh* more robust to the domain difference between the training set's controlled environment and in-the-wild environment of the testing set.

Two-stage approach The two-stage methods utilize the high accuracy of 2*D HPE*. They localize body key points in a 2*D* space and lift them to a 3*D* area. Motivated by the recent success of 2*D HPE*, 2*D* pose-based 3*D HPE* estimation approaches that infer 3*D* human pose from the intermediately estimated 2*D* human pose have become a popular 3*D HPE* solution. Benefiting from the excellent performance of state-of-the-art 2*D* pose detectors, 2*D* pose-based 3*D HPE* approaches generally

outperform direct image-based 3*D HPE* approaches. In the beginning stage, off-the-shelf 2*D HPE* models are applied to estimate 2*D* pose from the input data, and then in the next stage, 2*D* pose-based 3*D HPE* model is used to obtain the 3*D* pose. Martinez et al. [58] proposed a simple network that directly regresses the 3*D* coordinates of body joints from 2*D* coordinates. Yang et al. [118] utilized adversarial loss to handle the wild's 3*D HPE*. Park et al. [82] estimated the initial 2*D* pose and utilized it to regress the 3*D* pose. Zhou et al. [125] introduced a geometric loss to facilitate weakly supervised learning of the depth regression module.

Depth-based 3*D human pose estimation* With the recent success of networks in the image generation process has been demonstrated the use of generative networks to guide the 3*D HPE* during the training process. Depth-based 3*D HPE* exploits depth maps from estimated skeletons of the human body. Depth-based 3*D HPE* methods also rely on: generative models and discriminative models.

Generative models The generative models estimate the posture by finding the similarities between the pre-defined body and input 3*D* point clouds. The *ICP* [64] algorithm is usually used for 3*D* body tracking problems. Template fitting with Gaussian mixture models also was proposed.

Discriminative models The discriminative models directly estimate the positions of body joints without requiring body templates. Conventional discriminative methods are mostly based on random forests. Haque et al. [30] proposed the viewpoint-invariant pose estimation method using *CNN* and multiple recurrent neural network rounds. The proposed approach learns viewpoint-invariant features, which makes the model robust to viewpoint variations.

3.2 3*D* multi-person pose estimation

3*D* multi-person *HPE* for crowded scenes is essential in many computer vision applications such as autonomous driving, surveillance, and robotics. However, estimating the 3*D* human pose from a crowded real-world setting is still challenging. A three-step process is commonly used in the multi-person 3*D HPE* problem: (1) detecting human body key points; (2) matching people across different views; (3) reconstructing 3*D* human pose. Unfortunately, the critical second step of matching people across different views is non-trivial. Pose estimation in group pictures with severe occlusions attracts much attention. The 3*D* multi-person pose estimation from multiview images aims to estimate each key point's 3*D* coordinate rather than the 2*D* coordinate on the group image. Some joints may be more relevant to specific actions than others. Attention mechanism has been used to discover informative joints.

Estimation of a human pose can be very useful in many real-world *AIoT* scenarios, such as rehabilitation exercises monitoring and assessment, dangerous behavior monitoring and human–machine interaction. Some researches have been done on 3*D* multi-person pose estimation from a single *RGB* image. Mehta et al. [62] presented a bottom-up approach system. They proposed an occlusion-robust pose-map formulation that supports pose inference for more than one person through *PAFs*. In [91]

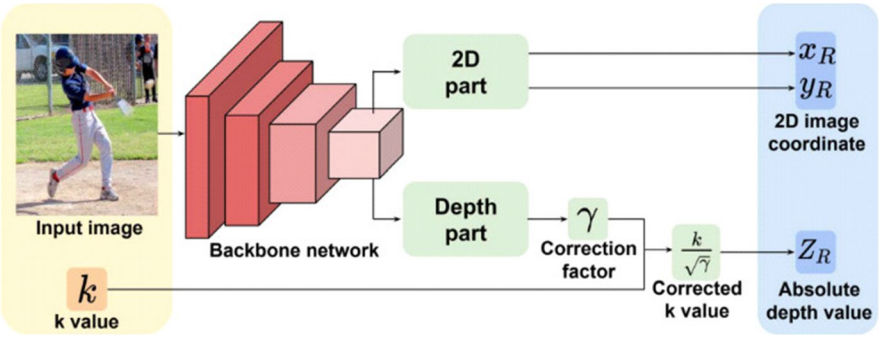

Fig. 8 Network architecture of the *RootNet* [64]

was introduced a top-down approach called *LCR-Net*. The proposed system consists of: localization part; classification part; regression parts.

Localization part The first localization part detects a human from an input image.

Classification part And the next classification part classifies the detected human into several anchor-poses. The anchor-pose is defined as a pair of 2*D* and root-relative 3*D* pose. It is generated by clustering poses in the training set.

Regression part Then, the last regression part refines the anchor-poses.

A novel and general framework was proposed by Moon et al. [64] for 3*D* multi-person pose estimation from a single *RGB* image. The presented framework consists of three Networks: human detection *DetectNet*, 3*D* human root localization *RootNet* and root-relative 3*D* single-person pose estimation *PoseNet* models. The authors declared that existing human detection and 3*D* single-person pose estimation models could be plugged into their proposed framework as it is very flexible and easy to use.

DetectNet. *Mask R-CNN* was exploited as the framework of *Detect-Net*. *Mask R-CNN* consists of three parts: backbone; region proposal network; classification head network.

- *Backbone*. It extracts useful local and global features from the input image by using a deep residual network (*ResNet*) [23, 31] and feature pyramid network [22, 24, 98].
- *Region proposal network*. It proposes human bounding box candidates based on the extracted features.
- *Classification head network*. The *RoIAlign* layer extracts each proposal's features and passes them to the third part, which is the classification head network. The head network determines whether the given proposal is a human or not and estimates the bounding box refinement offsets.

RootNet. The *RootNet* localizes the human's root $R = (x_R, y_R, Z_R)$ from a cropped human image, where x_R and y_R are pixel coordinates, Z_R is absolute depth value. *RootNet* estimates the 2*D* image coordinates (x_R, y_R) and the human root's depth value separately. The 2*D* image coordinates are back-projected to the camera-centered coordinate space using the estimated depth value. The image provides sufficient information on

where the human root is located in the image space. The 2*D* estimation part can learn to localize it easily. By contrast, estimating the depth only from a cropped human image is difficult because the input does not provide information on the camera and human's relative position. The network architecture of the *RootNet* is visualized in Fig. 8.

RootNet is trained by minimizing the L_1 distance between the estimated and ground-truth coordinates. The loss function L_{root} is defined as follows:

$$L_{root} = \|R - R^*\|_1 \tag{1}$$

where $*$ indicates the ground-truth.

PoseNet. The *PoseNet* estimates the root-relative 3*D* pose from a cropped human image as follows:

$$P_j^{rel} = \left(x_j, y_j, Z_j^{rel}\right) \tag{2}$$

where j is the number of human joints. It was exploited by Sun et al. [102], as a current state-of-the-art method, and it consists of two parts: Backbone; Pose estimation.

- *Backbone*. The first part, which is the backbone, extracts a useful global feature from the cropped human image using *ResNet*.
- *Pose estimation*. It takes a feature map from the backbone part and upsamples it using three consecutive deconvolutional layers with batch normalization layers [5, 37, 74] and *ReLU* [35, 53] activation function. A 1-by-1 convolution is applied to the upsampled feature map to produce the 3*D* heatmaps for each joint. The soft-argmax operation is used to extract the 2*D* image coordinates (x_j, y_j) and the root-relative depth values Z_j^{rel}

PoseNet is trained by minimizing the L_1 distance between the estimated and ground-truth coordinates. The loss function L_{pose} is defined as follows:

$$L_{pose} = \frac{1}{J} \sum_{j=1}^{J} \left\| P_j^{rel} - P_j^{rel*} \right\| \tag{3}$$

where $*$ indicates the ground truth, and J is the total number of coordinates.

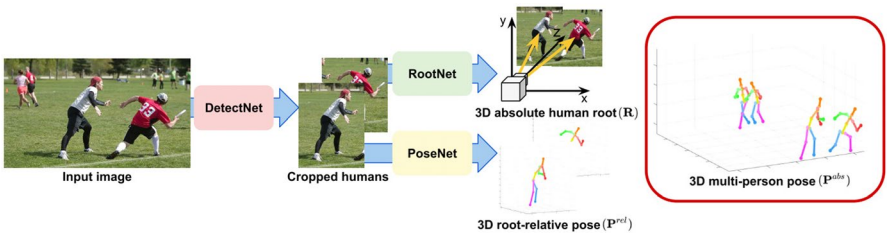

Fig. 9 Overall pipeline of the proposed framework for 3*D* multi-person pose estimation from a single RGB image [64]

The network architecture of the proposed work, which consists of three components, is visualized in Fig. 9:

- A human detection network (*DetectNet*) detects the bounding boxes of humans in an input image.
- The proposed 3*D* human root localization network (*RootNet*) estimates the detected human roots' camera-centered coordinates.
- A root-relative 3*D* single-person pose estimation network (*PostNet*) estimates the root-relative 3*D* pose for each detected human.

3.3 Volumetric representation using depth information

Wu et al. proposed a depth image's volumetric representation that surpassed the existing hand-crafted descriptor-based methods in 3*D* shape classification and retrieval problems. Each voxel was represented as a binary random variable and employed a convolutional deep belief network to learn the probability distribution for each voxel. Recent work [116] also represented 3*D* input data as a volumetric form for 3*D* object classification and detection. In [67], several types of volumetric representations were proposed to fully utilize the rich source of 3*D* information and efficiently deal with large amounts of point cloud data. Their presented *CNN* architecture and occupancy grids outperform state-of-the-arts in several available datasets.

4 3*D* human hand and mesh estimation methods

3*D* human pose and mesh estimation models were proposed to recover 3*D* human joint and mesh vertex locations simultaneously. This is a challenging task due to the depth and scale ambiguity, complexity of the human body and hand articulation. Recent deep learning-based methods have shown distinct performance improvement in solving this problem and outperformed all previous approaches. The deep learning-based methods rely on human mesh models and can be generally categorized into two approaches: model-based approach and model-free approach.

Model-based approach In the model-based approach, a network is trained to predict the model parameters and to generate a human mesh by decoding them [4, 6, 7, 38, 49, 77, 79].

Model-free approach In contrast, the model-free approach regresses a 3*D* human mesh coordinates directly [28, 50, 51]. Both given approaches compute the 3*D* human pose by multiplying the output mesh with a joint regression matrix defined in the human mesh models [16, 29, 32, 46, 86, 87, 92, 94].

Deep learning-based 3*D* human pose and mesh estimation models regress the pose and shape parameters of human mesh models. Even though they recently have shown significant improvement, they still have weaknesses: suffering from domain gap and unapproprite parameters.

Suffering from domain gap When tested on in-the-wild data, the models suffer from the gap that exists between the controlled and in-the-wild environment data. The data obtained from the controlled environments [7, 33, 60, 90, 113, 114] are considered as valuable train data in 3*D* human pose and estimation because it contains accurate 3*D* annotations. However, due to the significant difference in image appearance between the two domains, such as backgrounds and clothes, an image-based approach cannot fully benefit from the data.

Unapproprite parameters The pose parameters of the human mesh models might not be an appropriate regression target [50, 51]. The SMPL [6, 56] pose parameters represent 3*D* rotations in an axis-angle, resulting from the non-unique problem (i.e., periodicity). Although scientists [38, 77] tried to avoid the periodicity by utilizing a rotation matrix as the prediction target, it still has a non-minimal representation problem. Choi et al. [16] proposed *Pose2Mesh* as a solution to the above problems. *Pose2Mesh* is a graph convolutional network that recovers 3*D* human pose and mesh from the 2*D* human pose, in a model-free fashion. It has two advantages over existing methods: 2*D* poses from controlled and in-the-wild environments, and avoiding the representation issues. Tables 3, 4, 5 give a complete understanding of the discussed models and describe the taxonomy of the models, exploited networks, and experimental details, respectively.

2D poses from controlled and in-the-wild environments The proposed method benefits from a relatively homogeneous geometric property of the input 2*D* pose from controlled and in-the-wild environments. They alleviate the appearance of domain gap problem and provide essential geometric information on human articulation. 2*D* poses can be estimated accurately from in-the-wild images since many well-performing methods [15, 65, 99, 115] are trained on large-scale in-the-wild 2*D* human pose datasets [1].

Avoiding the representation issues The next advantage is that the proposed method avoids the pose parameters' representation issues while exploiting the human mesh topology. *Pose2Mesh* directly regresses the 3*D* coordinates of mesh vertices using a graph convolutional neural network (*Graph CNN*) with graphs constructed from the mesh topology.

Pose2Mesh. *Pose2Mesh* is designed in a cascaded architecture, which consists of *PoseNet* and *MeshNet*. The *PoseNet* lifts the 2*D* human pose to the 3*D* human pose. Moreover, the *MeshNet* takes both 2*D* and 3*D* human poses to estimate the 3*D* human mesh in a coarse-to-fine manner. The mesh features are initially processed in a coarse resolution and gradually upsampled to a fine resolution during the forward propagation. The overall pipeline of the proposed *Pos2Mesh* system is represented in Fig. 10.

4.1 Depth-based 3*D* hand pose estimation

Depth-based 3*D* hand pose estimation methods can be divided into: generative methods; discriminative methods; hybrid methods.

Generative methods appropriate a pre-defined hand shape and fit it to the input depth image by minimizing hand-crafted cost functions. Particle swam optimization

Table 3 Taxonomy of pose estimation models

Models	Key points	Advantages	Drawbacks
3D Human Pose Estimation			
Bayesian Capsule Network [88]	Novel Bayesian Capsule networks estimates the 3*D* human pose from a single 2*D* image	Minimize the homoscedastic uncertainty; self-balancing for a contribution of multiple losses to the total loss	For comparison with state-of-the-arts used a straightforward and much simpler approach over the *Human*3.6*M* data set
*V*2*V* − *Posenet* [67]	3*D CNN* provides accurate hand and pose estimation in real-time; first place in the *HANDS* 2017 frame-based 3*D* hand pose estimation challenge	Using 3*D* voxelized grid estimates the per-voxel likelihood for each key point; maps a single depth into a voxel-to-voxel prediction	Converting voxel-to-voxel to pixel-to-voxel changes the model from the 3*D CNN* to the 2*D CNN*, which may cause the performance to degrade
PoseFix [65]	*PoseFix* is trained independently of the pose estimation model	Can take the pose estimation result of any pose detection method as the input; does not require any code or knowledge about other methods	Takes an input pose in a coarse form and estimates the refined pose in a finer form; could not directly estimate from the image
DetectNet, *RootNet*, *PoseNet* [64]	Human detection, 3*D* human root localization, and root-relative 3*D* single-person pose estimation models	Fully learning-based camera distance-aware top-down approach; compatible with most of the previous human detection and 3*D HPE* models	Could not directly estimate from the image
DeepPose [105]	Implemented Deep Learning (*CNN*) to *HPE* that pretty much kicked off research in this direction	Based on a convolutional Deep Neural Network (*DNN*) which consists of several layers—each being a linear transformation followed by a non-linear one	Regressing to *XY* locations is difficult; adding learning complexity weakens generalization and performs poorly
3*D* mesh, pose estimation			
PoseNet, *MeshNet*, *Pose2Mesh* [16]	3*D* single-person pose estimation models, directly regresses 3*D* coordinates of a human mesh using *GraphCNN*	Can recover various body shapes from the 2*D* pose	Not image-based, uses the 2*D* human pose as an input
PoseNet, *MeshNet* [66]	3*D* single-person pose estimation models, directly regresses 3*D* coordinates of a human mesh using *GraphCNN*	Can recover various body shapes from the 2*D* pose	Not image-based, uses the 2*D* human pose as an input

Table 4 Taxonomy of networks used in pose estimation models

Network	Input	Output	Loss functions
3D Human Pose Estimation			
Bayesian Capsule Network [88]	2*D* image	3*D* coordinates of human body joints	Use multiple losses for different complementarytasks
V2V − Posenet [67]	2*D* depth map (image)	Regressed 3*D* coordinates of key points, hand or human body joints	Mean square error loss between ground-truth and estimated pose one $L = \sum_{n=1}^{N} \sum_{i,j,k} \left\|H_n^*(i,j,k) - H_n(i,j,k)\right\|^2$
PoseFix [65]	Pose estimation results of any other method with an input image, 2*D* coordinates of the human body key points	Refined pose, refined 2*D* coordinates of the human body key points	Cross entropy-based integral loss $L = L_H + L_C$, $L_H = \frac{1}{N} \sum_{n=1}^{N} \sum_{i,j} H_n^*(i,j) log H_n(i,j)$, $L_C = \frac{1}{N} \sum_{n=1}^{N} \left\|C_n^* - C_n\right\|_1$, where L_C sum of L_1 losses, L_H cross entropy loss
DetectNet [64]	Single *RGB* image	Cropped human image	Fast—$R - CNN$ with mean binary cross entropy $L = L_{cls} + L_{box} + L_{mask}$, where $L = L_{cls}$ classification loss, L_{box} bounding box loss, L_{mask} segmentation mask loss
RootNet [64]	Cropped human image	Localized root of the human $R = (x_R, y_R, Z_R)$, x_R, y_R are pixel coordinates, Z_R is an absolute depth value	L_1 distance between the estimated and ground-truth coordinates $L_{root} = \|R - R^*\|_1$, where R estimated pose coordinate and R^* ground-truth one
PoseNet [64]	Localized root of the human $R = (x_R, y_R, Z_R)$, x_R, y_R are pixel coordinates, Z_R is an absolute depth value	Absolute 3*D* pose $P_j^{abs} = (x_j, y_j, Z_j^{abs})$ from a cropped human image	L_1 distance between the estimated and ground-truth coordinates $L_{pose} = \frac{1}{J} \sum_{j=1}^{J} \left\|P_j^{rel} - P_j^{rel_*}\right\|_1$, where P_j^{rel} estimated pose, $P_j^{rel_*}$ ground-truth one
PoseNet [16]	2*D* pose outputs from Sun et al. [99, 115]	Root joint-relative 3*D* pose	L_1 distance between the predicted 3D pose P^{3D} and ground-truth P^{3D^*} $L_{pose} = \left\|P^{3D} - P^{3D^*}\right\|_1$
DeepPose [105]	Image of predefined size and has a size equal to the number of pixels times three color channels	Target values of the regression, 2*k* joint coordinates	Used L_2 loss for regression $arg\min_\theta \sum_{(x,y) \in D_N} \sum_{i=1}^{k} \left\|y_i - \psi_i(x;\theta)\right\|_2^2$, where x input image, θ model parameter, y_i ground-truth and ψ_i estimated pose coordinates
PoseNet [66]	2*D* pose outputs from Sun et al. [99, 115]	Root joint-relative 3*D* pose	L_1 distance between the predicted pose and ground-truth $L_{pose}^{PoseNet} = \left\|P^{3D} - P^{3D_*}\right\|_1$

Table 4 (continued)

Network	Input	Output	Loss functions
3*D* human mesh and poseestimation			
MeshNet [16]	3*D* single-person pose estimation models	3*D* human mesh and pose estimation from a 2*D* human pose	Vertex coordinate loss $L_{vertex} = \left\|M^C - M^{C*}\right\|_1$, Joint coordinate loss $L_{joint} = \left\|JM^C - P^{C*}\right\|_1$, Surface normal loss $L_{normal} = \sum_f \sum_{\{i,j\}\subset f} \left\|\left\langle \frac{m_i - m_j}{\|m_i - m_j\|_2}, n_f^* \right\rangle\right\|$, where M^C estimated and M^{C*} coordinates, $*$ indicates ground-truth
Pose2Mesh [16]	2*D* pose outputs from Sun et al. [99, 115], 3*D* single-person pose estimation models	3*D* human mesh and pose estimation from a 2*D* human pose	Surface edge loss $L_{edge} = \sum_f \sum_{\{i,j\}\subset f} \left\|\left\|m_i - m_j\right\|_2 - \left\|m_i^* - m_j^*\right\|_2\right\|$, $L_{mesh} = \lambda_v L_{vertex} + \lambda_j L_{joint} + \lambda_n L_{vnormal} + \lambda_e L_{edge}$, where L_{vertex} vertex loss, L_{joint} loss, and $\lambda_v = 1$, $\lambda_j = 1$, $\lambda_n = 0.1$, $\lambda_2 = 20$
MeshNet [66]	3*D* single-person pose estimation models	3*D* human mesh and pose estimation from a 2*D* human pose	$L = L_{pose}^{PoseNet} + L_{pose}^{MeshNet} + L_{vertex} + \lambda_v L_{normal} + L_{edge}$

Table 5 Experiment data of Networks used in pose estimation models

Network	Datasets	Evaluation metrics	Implementation details
3D Human Pose Estimation			
Bayesian Capsule Network [88]	*HUMAN*3.6*M* dataset	averaged errors of the Euclidean Distances of the 17 joints	AdamOptimizer $lr = 10^{-5}$ and $b - size = 1, 11, 20$, trained on Intel i7-7700, 3.6 GHz CPU with 16 GB of memory and NVIDIA GTX1080TI
$V2V - Posenet$ [67]	*ICVL* [103], *NYU*, *MSRA* [101], *HANDS* 2017 *Frame-based* [121] 3*D* hand, *ITOP* [30] 3*D HPE* datasets	3*D* distance error and percentage of success frame metrics, mean average precision (mAP) that is defined as the detected ratio of all human body joints based on 10 *cm* rule	Weights are initialized from the zero-mean *Gaussian* distribution with $\sigma = 0.001$, updated by the *RMSProp*, mini-batch size 8, 10 epochs, learning rate $2.5 * 10^{-4}$
PoseFix [65]	*ImageNet*, *COCO*	*OKS*-based *AP* metric is used to evaluate the accuracy of the key point localization	Weights are initialized from the zero-mean *Gaussian* distribution with $\sigma = 0.01$ and updated by *Adam optimizer* mini-batch size of 128, learning rate is set to $5 * 10^{-4}$ and reduced by a factor of 10 at 90 and 120th epoch
DetectNet [64]	Object detection datasets, *COCO* dataset	Mean per joint position error (*MPJPE*), *PA-MPJPE* calculates *MPJPE* after further alignment	Minibatch size is 16, 160 k iterations, with a learning rate of 0.02 which is decreased by 10 at the 120 kiteration, weight decay of 0.0001 and momentum of 0.9.
RootNet [64]	*ImageNet*, *Human*3.6.*M*	Mean per joint position error (*MPJPE*), *PA-MPJPE* calculates *MPJPE* after further alignment	Weights are initialized by *Gaussian* distribution with $\sigma = 0.001$ and updated by *Adam optimizer* with mini-batch size of 128, learning rate $1 * 10^{-3}$ after 17 epochs $1 * 10^{-4}$
PoseNet [64]	*MPI-INF*-3*DHP* [60], *MuPoTS*-3*D* [61]	Mean per joint position error (*MPJPE*), *PA-MPJPE* calculates *MPJPE* after further alignment	Used L_1 loss function, which is not recommendable

Table 5 (continued)

Network	Datasets	Evaluation metrics	Implementation details
PoseNet [16]	*MuCo*-3*DHP* [61, 62], *FreiHAND* [127]	Mean per joint position error (*MPJPE*), *PA-MPJPE* calculates *MPJPE* after further alignment	*Rmsprop* optimization updated weights by with mini-batch size of 64, pre-rained 60 epochs with learning rate $1 * 10^{-3}$ after 30 epochs $1 * 10^{-4}$
DeepPose [105]	Leeds sports dataset(*LSP*), Frames Labeled In Cinema (*FLIC*)	Percentage of Correct Parts (*PCP*), Percent of Detected Joints (*PDJ*)	θ is optimized for using *Backpropagation*, mini-batch size 128, learning rate is 0.0005, left/right flips as well as *DropOut* regularization for the *F* layers set to 0.6
PoseNet [66]	*MuCo*-3*DHP* [61, 62], *FreiHAND* [127], *ImageNet*	Mean per joint position error (*MPJPE*), *PA-MPJPE* calculates *MPJPE* after further alignment	*Rmsprop* optimization updated weights by with mini-batch size of 64, pre-rained 60 epochs with learning rate $1 * 10^{-3}$ after 30 epochs $1 * 10^{-4}$
3*D* human mesh and pose estimation			
MeshNet [16]	*Human*3.6.*M*, 3*DPW* [57], *COCO*	Mean per joint position error (*MPJPE*), *PA-MPJPE* calculates *MPJPE* after further alignment	*Rmsprop* optimization updated weights by with mini-batch size of 64
*Pose*2*Mesh* [16]	*Human*3.6.*M*, 3*DPW* [57], *COCO*	Mean per joint position error (*MPJPE*), *PA-MPJPE* calculates *MPJPE* after further alignment	*Rmsprop* optimization updated weights by with mini-batch size of 64, trained 15 epochs with learning rate $1 * 10^{-3}$ after 12 epochs $1 * 10^{-4}$
MeshNet [66]	*Human*3.6.*M*, 3*DPW* [57], *COCO*	mean per joint position error (*MPJPE*), *PA-MPJPE* calculates *MPJPE* after further alignment	*Rmsprop* optimization updated weights by with mini-batch size of 64

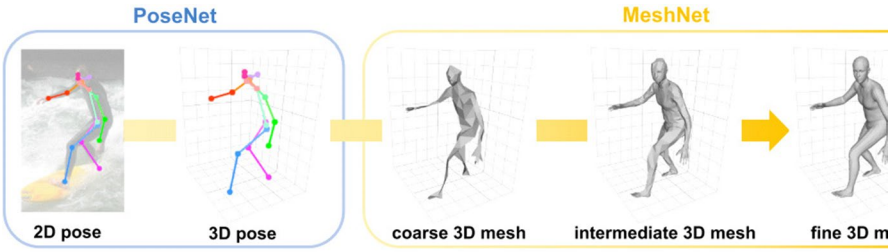

Fig. 10 Overall architecture of the *Pose2Mesh* network [16]

(*PSO*), iterative closest point (*ICP*), and their combination are the familiar algorithms used to obtain optimal hand pose [47] results.

Discriminative methods directly localize hand joints from an input depth map. Random forest-based methods [45] provide quick and precise representation. The *CNN*-based approaches outperform the existing methods and can learn useful features by themselves. *CNN* primarily was utilized to localize hand key points by estimating 2*D* heatmaps for each hand joint and then was extended by exploiting multi-view *CNN* to estimate 2*D* heatmaps. In [26, 27], the 2*D* input depth map was transformed to the 3*D* form and the 3*D* coordinates were calculated directly via 3*D* *CNN*.

Hybrid methods are a combination of the generative and discriminative approach. Oberweger et al. [73] suggested training the discriminative and generative *CNNs* by a feedback loop. Zhou et al. [126] proposed defining a hand model and estimating the model's parameter then regress to 3*D* coordinates. Furthermore, in [119], the spatial attention mechanism and hierarchical *PSO* were utilized. Wan et al. [110] used two deep generative models with a shared latent space and training discriminator to estimate the posterior of the latent pose.

4.2 3D human hand and mesh estimation

A model-based approach trains a neural network to estimate the human mesh model parameters [56, 92]. The neural network has been widely used for the 3*D* human mesh estimation since it does not necessarily require 3*D* annotation for mesh supervision. Kanazawa et al. [38] used the adversarial loss to regress plausible *SMPL* parameters. Baek et al. [4] trained *CNN* to estimate the *MANO* model parameters using a neural renderer [40]. Omran et al. [77] introduced training a network with 2*D* joint coordinates, which takes human part segmentation as input. The advancement of fitting frameworks [6, 83] has motivated a model-free approach that estimates human mesh coordinates directly. Researchers could obtain 3*D* mesh annotation, which is essential for the model-free methods, from in-the-wild data. Ge et al. [26] utilized a *GraphCNN* to estimate vertices of hand mesh. Kolotouros et al. [50, 51] proposed a *GraphCNN*, which learns the template body mesh's deformation to the target body mesh. Moon et al. [66] introduced a new heatmap representation, called lixel, to recover 3*D* human meshes. Choi et al. [16] presented a novel method

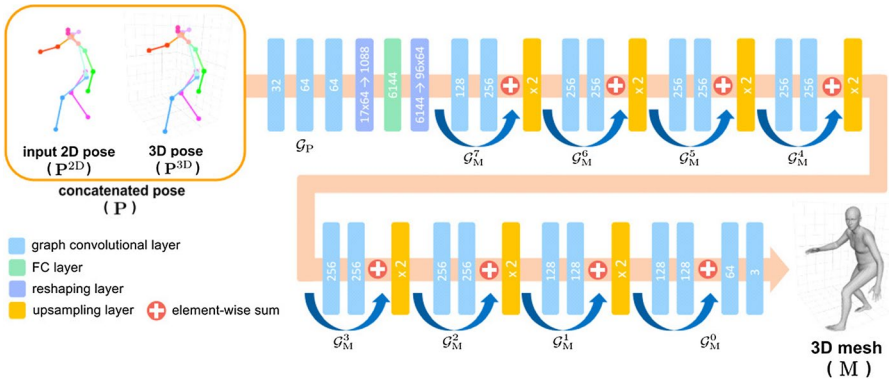

Fig. 11 Network architecture of *MeshNet* [16]

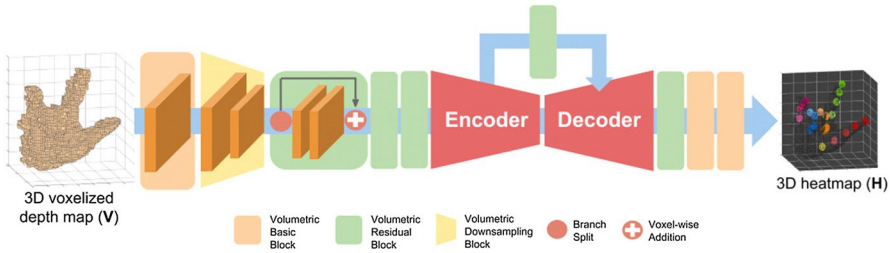

Fig. 12 Overall architecture of the V2V-PoseNet [67]

Pose2Mesh that differs from the previous models, which are image based, in that it uses the 2*D* human pose as an input. The proposed *Pose2Mesh* system can benefit from the data with 3*D* annotations captured from controlled environments. As described in Fig. 10, it consists of two networks:

- *PoseNet*
- *Meshnet*

The network architecture of the *MeshNet* network model is described in Fig. 11.

4.3 Input and output representation in 3*D* hand pose estimation

The most massive existing 3*D* hand pose estimation methods from a single depth map are based on taking a 2*D* depth image and directly regressing 3*D* coordinates. A 2*D* depth image was recently converted to a truncated signed distance function-based 3*D* volumetric form and directly regressed 3*D* coordinates [21]. In 3*D HPE* from an *RGB* image, the per-voxel likelihood for each body key point via 2*D CNN* was estimated by Pavlakos et al. [83]. The discretized depth value was treated as a channel of the feature map, which resulted in different kernels for each depth value. Moon et al. [67]

Table 6 Average 3*D* distance error (mm) [67]

InputOutput	3*D* Coordinates	Per-voxel likelihood
2*D* depth map	18.85 (21.1M)	13.01 (4.6 M)
3*D* voxelized grid	16.78 (457.5M)	**10.37 (3.4M)**

Table 7 Effect of localization refinement and epoch ensemble [67]

Methods	Average 3*D* distance error
Baseline	11.14 mm
+Localization refinement	9.22 mm
+Epoch ensemble	8.42 mm

proposed to estimate each key point's per-voxel likelihood from the voxelized input. They exploited the 3*D* fully convolutional network, and it was declared as the first model to generate voxelized output from voxelized input using 3*D CNN* for 3*D* hand pose estimation. To estimate the 3*D* coordinates of all key points, 2*D* depth images were converted to 3*D* volumetric forms. Then, $V2V-PoseNet$ takes the 3*D* voxelized data and estimates each key point's per-voxel likelihood (Fig. 12).

4.4 Refining target object localization

To localize key points, such as hand or human body joints, a cubic box that contains the hand or human body in 3*D* space is essential. This cubic box is usually placed around the reference point obtained using ground-truth joint position or the center of mass after simple depth thresholding around the hand region. However, in the real-world applications, using the ground-truth joint part is impractical. Moreover, the usage of center of mass calculated by simple depth thresholding does not guarantee that the object is correctly contained in the acquired cubic box due to the error in the center-of-mass calculations in cluttered scenes. When different items are near the target one, the simple depth thresholding method cannot filter the other objects correctly. So, the computed center-of-mass cubic box becomes incorrect. These weaknesses were overcome by training a simple 2*D CNN*; then, [72, 73] obtained a valid reference point. The network takes a depth image, whose reference point is calculated by the simple depth thresholding, and outputs 3*D* offset. The refined reference point can be obtained by adding the network's output offset value to the calculated one. Table 6 shows converting the input image type from the 2*D* depth map to 3*D* voxelized form (from 2*D CNN* to 3*D CNN*) substantially improves performance.

The power of the localization-refining procedure and the epoch ensemble are described in Table 7.

4.5 GraphCNN for mesh processing

In current methods, a mesh is considered as a graph structure. It is processed using the *GraphCNN* as it can fully exploit mesh topology. *GraphCNN* is utilized to learn a deformation from an initial ellipsoid mesh to the target object mesh in a coarse-to-fine manner [111]. Verma et al. [109] introduced a graph convolution operator and evaluated it on the correspondence problem. *GraphCNN-based VAE* was also proposed, which learns a latent space of the human face meshes hierarchically.

5 Datasets and evaluation metrics used in HPE

Datasets and evaluation protocols play a significant role in *DL*-based *HPE*. They are essential for fair comparison of different algorithms and bring more challenges and complexity through their expansion and improvement. With the maturity of commercial motion capture systems and crowdsourcing services, recent datasets are no longer limited by the data quantity or lab environments. This section will discuss some publicly available datasets for 2*D* and 3*D* HPE, with the most used evaluation metrics.

5.1 Datasets for 2D human pose estimation

Early created datasets for 2*D HPE* contain images with relatively simple backgrounds. However, *DL*-based models are unsuitable for these datasets because the number of images is too small for training. The common datasets used in *DL*-based approaches include *MSCOCO*, *MPII*, *LSP*, *FLIC*, *Pose Track*, and *AI Challenger*, which contain more images in more complicated scenes. The *HPE* datasets, such as *FLIC* and *LSP*, are relatively small and only contain specific activity categories. The images in the *FLIC* dataset are collected from Hollywood movies. The *LSP* dataset images are from sports scenes. Other datasets, such as *AIChallenger* and *MSCOCO*, are bigger in both size and number of image categories.

5.2 Datasets for 3D human pose estimation

In contrast with 2*D HPE* datasets, acquiring accurate 3*D* annotation for 3*D HPE* datasets is challenging. It requires motion capture systems such as *MoCap* and wearable *IMU*s. Therefore, many 3*D HPE* datasets are created in constrained conditions.

HumanEva dataset contains seven calibrated video sequences with ground-truth 3*D* annotation captured by a commercial *MoCap* system. It consists of four subjects performing six everyday actions: walking, jogging, gesturing, throwing and catching a ball, boxing, and combo. *Human*3.6*M* is the mainly used dataset for 3*D HPE* from monocular images and videos. It consists of 11 professional actors performing 17 activities such as smoking, taking photos, talking on the phone and etc. The dataset contains 3.6 million 3*D* human poses with 3*D* ground-truth annotation captured by an accurate marker-based *MoCap* system. *TNT*15 dataset consists of synchronized

data streams from eight *RGB* cameras and ten *IMU*s. It has been recorded in an office environment. It records four actors performing five activities: walking, running on the spot, rotating arms, jumping and skiing, and dynamic punching. The dataset contains about 13k frames, including binary segmented images obtained by background subtraction, 3*D* laser scans and registered meshes of each actor. *MPI-INF-3DHP* dataset was collected with a marker- less multi-camera *MoCap* system and includes both indoor and outdoor scenes. It contains over 1.3 M frames from 14 different views. Eight subjects are recorded performing eight activities such walking/standing, exercise, sitting, crouch/reach, on the floor, sports, miscellaneous. *TotalCapture* dataset was captured indoors with eight calibrated *HD* video cameras. There are four male and one female subjects performing four diverse performances, repeated three times. The variation and body motions within the acting and freestyle sequences are very challenging with actions such as yoga, giving directions, bending over and crawling performed in both the train and test data.

Furthermore, the *MARCOnI* dataset contains sequences in a variety of uncontrolled indoor and outdoor scenarios. They vary according to different data modalities captured, in the numbers and identities of actors to track, the complexity of the motions, the number of cameras used, the existence and number of moving objects in the background and the lighting conditions. Cameras differ in the types, hence the frame resolutions, and the frame rates. *Panoptic* dataset was captured with a markerless motion capturing using multiple view systems. It contains 65 sequences of social interaction with 1.5 million 3*D* skeletons. The provided annotations include 3*D* key points, cloud points, optical flow, etc. 3*DPW* dataset was captured using a single hand-held camera in natural environments. 3*D* annotations are estimated from *IMU*s. All subjects in the dataset are provided with 3*D* scans. It consists of 60 video sequences with periodic actions, including walking in the city, going upstairs, having coffee, taking the bus, etc. The datasets were collected with *MoCap* systems. Table 8 shows popular 3*D HPE* datasets which we have described above.

5.3 Evaluation metrics used in 2D HPE

Different datasets have different features and task requirements (single/multi-pose). Therefore, several metrics are used evaluate the performance in 2*D HPE*, which is tricky due to many factors that need to be considered. We will describe some of the commonly used metrics in the following.

Percentage of Correctly estimated body Parts (*PCP*) metric evaluates stick predictions, and it was used in early research studies. *PCP* reports the localization accuracy for human limbs. A human limb is correctly localized if its two endpoints are within a threshold from the corresponding ground truth endpoints. Besides, a mean *PCP*, some limbs *PCP*, such as the torso, upper legs, lower legs, upper arms, forearms, head, are also reported. Moreover, percentage curves for each limb can be obtained with the threshold variation in the metric. The similar metrics *PCPm* use 50% of the mean ground-truth segment length over the entire test set as a matching threshold.

Table 8 Popular datasets for 3*D HPE* [14]

Dataset name	Evaluation metric	Highlights
Single person approach		
HumanEva − I	*MPJPE*	4/2 subjects, 6/1 actions, *Vicon* data, indoor environment
HumanEva − II	*MPJPE*	4/2 subjects, 6/1 actions, *Vicon* data, indoor environment
*Human*3.6*M*	*MPJPE*	11 subjects, 17 actions, *Vicon* data, indoor environment, multi-annotation
*TNT*15	*MPJPE*	4 subjects, 5 actions, *IMU* data, 3*D* body scans, indoor environment
MPI − INF − 3DHP	3*DPCK*	8 subjects, 8 actions, commercial markerless system, indoor and outdoor scenes
TotalCapture	*MPJPE*	5 subjects, 5 actions, *Vicon* and *IMU* data, indoor environment, multi-annotation
Multi-person approach		
Panoptic	3*DPCK*	Up to 8 subjects in each video, social interactions, markerless studio, multi-annotation, indoor environment
3*DPW*	*MPJPE*, *MPJAE*	7 subjects, daily actions, estimated 3*D* poses, 3*D* scans, in the wild

Percentage of Correct Key points (*PCK*) measures the accuracy of the localization of the human body joints. A human body joint is considered correct if it falls within the threshold pixels of the ground-truth joint. Moreover, with the variation in a threshold, Area Under the Curve (*AUC*) can be generated for further analysis.

The Object Key point Similarity (*OKS*) and Average Precision (*AP*) of *OKS* consider scale and introduce the per-point constant to control falloff.

AP, Average Recall (*AR*) and their variants are also metrics used in evaluating multi-person pose estimation results. *AP*, *AR* and their variants are reported based on an analogous similarity measure: object key point similarity (*OKS*), which plays the same role as the Intersection over Union (*IoU*). In addition, *AP*/*AR* with different human body scales are also reported in the *COCO* dataset.

5.4 Evaluation metrics used in 3D HPE

There are several evaluation metrics for 3*D HPE* with different limitation factors. In this subsection, we will give a list of widely used evaluation metrics.

Mean Per Joint Position Error (*MPJPE*) is one of the most popular metrics to evaluate the performance of 3*D HPE*. It is based on *Euclidean* distance and calculates the distance from the estimated 3*D* joints to the ground truth, averaged over all joints in one image. In the set of frames cases, the mean error is averaged over all frames. Different datasets and protocols have different data post-processing of estimated joints before computing the *MPJPE*.

PMPJPE measure called a Reconstruction Error is the *MPJPE* after rigid alignment by post-processing between the estimated pose and the ground-truth one.

NMPJPE is defined as the *MPJPE* after normalizing the predicted positions in scale to the reference.

Mean Per Vertex Error (*MPVE*) measures the Euclidean distances between the ground truth vertices and the predicted vertices.

3*DPCK* is a 3*D* extended version of the *PCK* metric used in 2*D HPE* evaluation. An estimated joint is considered correct if the distance between the estimation and the ground truth is within a certain threshold, and mainly the threshold is set to 150 mm.

6 Open issues and challenges

HPE is still a hot topic in computer vision which recently has evolved along with DL approaches. Despite the significant development of 2*D* and 3*D* human hand, pose and mesh estimation with *DL*, some unresolved open issues and challenges still exist between academia and industry, for example issue of the influence of human body part occlusion and crowded people. Effective practical models and sufficient training data are essential for *DL*-based methods. The massive interest in *HPE* and its importance can be seen from the workshops and challenges on *HPE*, increasing. They gather researchers from academia and industry on *HPE* and discuss the current state-of-the-art and future research directions. Here we give some of them

as we decide as recent and important ones: *ICCV* 2017—PoseTrack Challenge: Human Pose Estimation and Tracking in the Wild, *CVPR* 2018—3D humans 2018: 1st International workshop on Human pose, motion, activities and shape, ECCV 2018-PoseTrack Challenge: Articulated People Tracking in the Wild, CVPR 2019—Workshop On Augmented Human: Human-centric Understanding, CVPR 2019-3D humans 2019: 2nd International workshop on Human pose, motion, activities and shape, ACM Multimedia-2020 Large-scale Human-centric Video Analysis in Complex Events, CVPR 2020-Towards Human-Centric Image/Video Synthesis, ECCV 2020-3D poses in the wild challenge.

7 Conclusion

This paper reviewed and discussed recent published *DL*-based papers on the human pose, hand and mesh estimation approaches in great detail. We have comprehensively investigated the related theoretical and practical issues compared to existing methods in this *HPE* research field. Moreover, the pose estimation concepts and their applications are clearly explained in detail to provide readers with a deeper understanding of these topics. We also provided a clear taxonomy of the presented survey-based 2*D* and 3*D* pose, hand and mesh estimation, including single-person or multi-person, single-stage or double-stage categories. In addition, datasets and metrics used in the *HPE* research approaches are provided for both 2*D* and 3*D* HPE approaches. The taxonomy of the presented paper is based on the methodology, which includes single-person or multi-person, single-stage or double-stage pipelines. The comparisons are made among different frameworks and different pipelines of the *HPE* approaches. Moreover, we also summarized the datasets and evaluation metrics for *DL*-based 2*D* and 3*D HPE* approaches. We hope that the presented review work can motivate new research efforts to improve the *HPE* approaches with large-scale applications such as non-verbal and remote communication, including hand and body motion, *VR*, *AR*, human action recognition and computer games.

8 Future research directions

Despite the remarkable success in the *HPE* field, there are still various promising future directions to promote advances in *HPE* research. Further, we point out some of them:

3*D HPE* is usually used in visual tracking and analysis fields. Existing 3*D* human hand, pose and mesh estimation from the given videos is not smooth and continuous. It is because the evaluation metrics cannot evaluate the smoothness. Suitable frame-level evaluation metrics focusing on temporary consistency and action smoothness should be generated.

The slight noise can significantly affect the performance of the *HPE* network. *DL*-based networks in computer vision tasks are weak to adversarial attacks. The

researches against adversarial attacks can improve the robustness of models and promote real-world *HPE* applications.

Acknowledgements This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea *NRF*-2019*S*1*A*5*C*2*A*03081234 and Inha University.

References

1. Andriluka M, Iqbal U, Insafutdinov E, Pishchulin L, Milan A, Gall J, Schiele B (2018) Posetrack: a benchmark for human pose estimation and tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5167–5176
2. Averbukh V, Averbukh N, Vasev P, Gvozdarev I, Levchuk G, Melkozerov L, Mikhaylov I (2019) Metaphors for software visualization systems based on virtual reality. In: International Conference on Augmented Reality, Virtual Reality and Computer Graphics, pp 60–70. Springer
3. Babu SC (2018) A 2019 guide to Human Pose Estimation with Deep Learning. Accessed 3 Feb 2018. https://nanonets.com/blog/human-pose-estimation-2d-guide/
4. Baek S, Kim KI, Kim TK (2019) Pushing the envelope for RGB-based dense 3d hand pose estimation via neural rendering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1067–1076
5. Barreda M, Dolz MF, Castaño MA, Alonso-Jordá P, Quintana-Orti ES (2020) Performance modeling of the sparse matrix-vector product via convolutional neural networks. J Supercomput 76(11):8883–8900
6. Bogo F, Kanazawa A, Lassner C, Gehler P, Romero J, Black MJ (2016) Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. In: European Conference on Computer Vision, pp 561–578. Springer
7. Boukhayma A, Bem R, Torr PH (2019) 3d hand shape and pose from images in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 10843–10852
8. Bounouni M, Bouallouche-Medjkoune L (2018) Acknowledgment-based punishment and stimulation scheme for mobile ad hoc network. J Supercomput 74(10):5373–5398
9. Bulat A, Tzimiropoulos G (2016) Human pose estimation via convolutional part heatmap regression. In: 14th European Conference on Computer Vision (EECV 2016), 8–16 October 2016. Amsterdam, Netherlands
10. Bulat A, Tzimiropoulos G (2016) Human pose estimation via convolutional part heatmap regression. In: European Conference on Computer Vision. Springer, pp 717–732
11. Bulat A, Tzimiropoulos G (2016) Human pose estimation via convolutional part heatmap regression. Lecture Notes in Computer Science pp 717–732. https://doi.org/10.1007/978-3-319-46478-7_44
12. Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7291–7299
13. Carreira J, Agrawal P, Fragkiadaki K, Malik J (2016) Human pose estimation with iterative error feedback. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4733–4742

14. Chen Y, Tian Y, He M (2020) Monocular human pose estimation: a survey of deep learning-based methods. Comput Vis Image Underst 192:102897. https://doi.org/10.1016/j.cviu.2019.102897
15. Chen Y, Wang Z, Peng Y, Zhang Z, Yu G, Sun J (2018) Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7103–7112
16. Choi H, Moon G, Lee KM (2020) Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. arXiv preprint arXiv:2008.09047
17. Choi S, Kim C, Kang YS, Youm S (2021) Human behavioral pattern analysis-based anomaly detection system in residential space. J Supercomput 77:1–18
18. Chu X, Yang W, Ouyang W, Ma C, Yuille AL, Wang X (2017) Multi-context attention for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1831–1840
19. Clark A, Absher J (2018) Cyber-surveillance analysis for supercomputing environments. In: Surveillance in action. Springer, pp. 395–412
20. Dang Q, Yin J, Wang B, Zheng W (2019) Deep learning based 2d human pose estimation: a survey. Tsinghua Sci Technol 24(6):663–676. https://doi.org/10.26599/TST.2018.9010100
21. Deng X, Yang S, Zhang Y, Tan P, Chang L, Wang H (2017) Hand3d: hand pose estimation using 3d neural network. arXiv preprint arXiv:1704.02224
22. Ding P, Zhang J, Zhou H, Zou X, Wang M (2020) Pyramid context learning for object detection. J Supercomput 76:1–14
23. Du X, Kuang D, Ye Y, Li X, Chen M, Du Y, Wu W (2018) Comparative study of distributed deep learning tools on supercomputers. In: International Conference on Algorithms and Architectures for Parallel Processing. Springer, pp 122–137
24. Fang B, Fang L (2020) Concise feature pyramid region proposal network for multi-scale object detection. J Supercomput 76(5):3327–3337
25. Fieraru M, Khoreva A, Pishchulin L, Schiele B (2018) Learning to refine human pose estimation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). https://doi.org/10.1109/cvprw.2018.00058
26. Ge L, Liang H, Yuan J, Thalmann D (2017) 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1991–2000
27. Ge L, Liang H, Yuan J, Thalmann D (2018) Robust 3d hand pose estimation from single depth images using multi-view CNNs. IEEE Trans Image Process 27(9):4422–4436
28. Ge L, Ren Z, Li Y, Xue Z, Wang Y, Cai J, Yuan J (2019) 3d hand shape and pose estimation from a single RGB image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10833–10842
29. Habermann M, Xu W, Zollhofer M, Pons-Moll G, Theobalt C (2020) Deepcap: monocular human performance capture using weak supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5052–5063
30. Haque A, Peng B, Luo Z, Alahi A, Yeung S, Fei-Fei L (2016) Towards viewpoint invariant 3d human pose estimation. In: European Conference on Computer Vision. Springer, pp. 160–177
31. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. CoRR arXiv:abs/1512.03385
32. Hesse N, Pujades S, Black M, Arens M, Hofmann U, Schroeder S (2019) Learning and tracking the 3d body shape of freely moving infants from RGB-D sequences. IEEE Trans Pattern Analysis Mach Intell 42:2540
33. Hidalgo G, Raaj Y, Idrees H, Xiang D, Joo H, Simon T, Sheikh Y (2019) Single-network whole-body pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 6982–6991
34. Huang S, Gong M, Tao D (2017) A coarse-fine network for keypoint localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3028–3037
35. Ichimura T, Fujita K, Yamaguchi T, Hori M, Wijerathne L, Ueda N (2020) Fast multi-step optimization with deep learning for data-centric supercomputing. In: Proceedings of the 2020 4th International Conference on High Performance Compilation, Computing and Communications, pp 7–13
36. Insafutdinov E, Pishchulin L, Andres B, Andriluka M, Schiele B (2016) Deepercut: a deeper, stronger, and faster multi-person pose estimation model. In: European Conference on Computer Vision. Springer, pp 34–50

37. Jifara W, Jiang F, Rho S, Cheng M, Liu S (2019) Medical image denoising using convolutional neural network: a residual learning approach. J Supercomput 75(2):704–718
38. Kanazawa A, Black MJ, Jacobs DW, Malik J (2018) End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7122–7131
39. Kaster J (2020) Training convolutional neural network classifiers using simultaneous scaled supercomputing. Tech. rep., University of Dayton Dayton United States
40. Kato H, Ushiku Y, Harada T (2018) Neural 3d mesh renderer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3907–3916
41. Kaur M, Kaur G, Sharma PK, Jolfaei A, D Singh (2020) Binary cuckoo search metaheuristic-based supercomputing framework for human behavior analysis in smart home. J Supercomput 76(4):2479–2502
42. Kawana Y, Ukita N, Huang JB, Yang MH (2018) Ensemble convolutional neural networks for pose estimation. Comput Vis Image Underst 169:62–74. https://doi.org/10.1016/j.cviu.2017.12.005
43. Ke L, Chang MC, Qi H, Lyu S (2018) Multi-scale structure-aware network for human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 713–728
44. Ke S, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. pp 5686–5696. https://doi.org/10.1109/CVPR.2019.00584
45. Khamparia A, Gupta D, de Albuquerque VHC, Sangaiah AK, Jhaveri RH (2020) Internet of health things-driven deep learning system for detection and classification of cervical cells using transfer learning. J Supercomput 76(11):8590–8608. https://doi.org/10.1007/s11227-020-03159-4
46. Kim M, Pons-Moll G, Pujades S, Bang S, Kim J, Black MJ, Lee SH (2017) Data-driven physics for human soft tissue animation. ACM Trans Gr 36(4):1–12
47. Kim S, Jang SW, ho Park J, Kim G (2019) Robust hand pose estimation using visual sensor in IoT environment. J Supercomput 76:5382–5401
48. Kocabas M, Karagoz S, Akbas E (2018) Multiposenet: fast multi-person pose estimation using pose residual network. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 417–433
49. Kolotouros N, Pavlakos G, Black MJ, Daniilidis K (2019) Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2252–2261
50. Kolotouros N, Pavlakos G, Daniilidis K (2019) Convolutional mesh regression for single-image human shape reconstruction** supplementary material
51. Kolotouros N, Pavlakos G, Daniilidis K (2019) Convolutional mesh regression for single-image human shape reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4501–4510
52. Kondratyuk N, Smirnov G, Agarkov A, Osokin A, Nikolskiy V, Semenov A, Stegailov V (2019) Performance and scalability of materials science and machine learning codes on the state-of-art hybrid supercomputer architecture. In: Russian Supercomputing Days. Springer, pp 597–609
53. Lee D, Kang S, Choi K (2018) Compend: computation pruning through early negative detection for ReLU in a deep neural network accelerator. In: Proceedings of the 2018 International Conference on Supercomputing, pp. 139–148
54. Li J, Liu M, Ma D, Huang J, Ke M, Zhang T (2020) Learning shared subspace regularization with linear discriminant analysis for multi-label action recognition. J Supercomput 76(3):2139–2157
55. Loghman M, Kim J, Choi K (2018) Fast depth estimation using semi-global matching and adaptive stripe-based optimization. J Supercomput 74(8):3666–3684
56. Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ (2015) SMPL: a skinned multi-person linear model. ACM Trans Gr 34(6):1–16
57. von Marcard T, Henschel R, Black MJ, Rosenhahn B, Pons-Moll G (2018) Recovering accurate 3d human pose in the wild using imus and a moving camera. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 601–617
58. Martinez J, Hossain R, Romero J, Little JJ (2017) A simple yet effective baseline for 3d human pose estimation. In: 2017 IEEE International Conference on Computer Vision (ICCV). https://doi.org/10.1109/iccv.2017.288
59. McDanel B, Zhang SQ, Kung H, Dong X (2019) Full-stack optimization for accelerating CNNs using powers-of-two weights with FPGA validation. In: Proceedings of the ACM International Conference on Supercomputing, pp 449–460

60. Mehta D, Rhodin H, Casas D, Fua P, Sotnychenko O, Xu W, Theobalt C (2017) Monocular 3d human pose estimation in the wild using improved CNN supervision. In: 2017 International Conference on 3D Vision (3DV), IEEE, pp 506–516
61. Mehta D, Sotnychenko O, Mueller F, Xu W, Sridhar S, Pons-Moll G, Theobalt C (2017) Single-shot multi-person 3d pose estimation from monocular RGB. arXiv preprint arXiv:1712.03453
62. Mehta D, Sotnychenko O, Mueller F, Xu W, Sridhar S, Pons-Moll G, Theobalt C (2018) Single-shot multi-person 3d pose estimation from monocular RGB. In: 2018 2017 International Conference on 3D Vision (3DV), IEEE, pp 120–130
63. Millar K, Cheng A, Chew HG, Lim CC (2019) Using convolutional neural networks for classifying malicious network traffic. Springer, Cham, pp 103–126. https://doi.org/10.1007/978-3-030-13057-2_5
64. Moon G, Chang JY, Lee KM (2019) Camera distance-aware top-down approach for 3d multi-person pose estimation from a single RGB image. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). https://doi.org/10.1109/iccv.2019.01023
65. Moon G, Chang JY, Lee KM (2019) Posefix: model-agnostic general human pose refinement network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7773–7781
66. Moon G, Lee KM (2020) 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
67. Moon G, Yong Chang J, Mu Lee K (2018) V2v-posenet: voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5079–5088
68. Mukhiddin T, Lee W, Lee S, Rashid T (2020) Research issues on generative adversarial networks and applications. In: 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), IEEE, pp 487–488
69. Newell A, Huang Z, Deng J (2017) Associative embedding: end-to-end learning for joint detection and grouping. In: Advances in Neural Information Processing Systems, pp 2277–2287
70. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. Springer, pp 483–499
71. Nie X, Feng J, Zhang J, Yan S (2019) Single-stage multi-person pose machines. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 6951–6960
72. Oberweger M, Lepetit V (2017) Deepprior++: improving fast and accurate 3d hand pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp 585–594
73. Oberweger M, Wohlhart P, Lepetit V (2015) Training a feedback loop for hand pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3316–3324
74. Oh SH, Han SW, Choi BS, Kim GW, Lim KS (2018) Deep feature learning for person re-identification in a large-scale crowdsourced environment. J Supercomput 74(12):6753–6765
75. de Oliveira CHR, Costa APF, Thomaz VF, Silva IA (2019) Low-cost deployment proposal to urban mobility in smart cities. J Supercomput 75(11):7265–7289
76. Oliveira D, Blanchard S, DeBardeleben N, Fernandes dos Santos F, Piscoya Dávila G, Navaux P, Favalli A, Schappert O, Wender S, Cazzaniga C et al (2020) Thermal neutrons: a possible threat for supercomputer reliability. J Supercomput 77:1–23
77. Omran M, Lassner C, Pons-Moll G, Gehler P, Schiele B (2018) Neural body fitting: unifying deep learning and model based human pose and shape estimation. In: 2018 International Conference on 3D Vision (3DV), IEEE, pp. 484–494
78. Pan H, Li Y, Zhao D (2021) Recognizing human behaviors from surveillance videos using the SSD algorithm. J Supercomput. https://doi.org/10.1007/s11227-020-03578-3
79. Panteleris P, Oikonomidis I, Argyros A (2018) Using a single RGB frame for real time 3d hand pose estimation in the wild. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, pp 436–445
80. Papandreou G, Zhu T, Kanazawa N, Toshev A, Tompson J, Bregler C, Murphy K (2017) Towards accurate multi-person pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4903–4911
81. Papandreou G, Zhu T, Kanazawa N, Toshev A, Tompson J, Bregler C, Murphy K (2017) Towards accurate multi-person pose estimation in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2017.395

82. Park S, Hwang J, Kwak N (2016) 3d human pose estimation using convolutional neural networks with 2d pose information. In: Computer Vision—ECCV 2016 Workshops, pp. 156–169. https://doi.org/10.1007/978-3-319-49409-8_15
83. Pavlakos G, Zhou X, Derpanis KG, Daniilidis K (2017) Coarse-to-fine volumetric prediction for single-image 3d human pose. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2017.139
84. Pishchulin L, Insafutdinov E, Tang S, Andres B, Andriluka M, Gehler P, Schiele B (2016) Deepcut: joint subset partition and labeling for multi person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4929–4937
85. Proietti R, Shang Y, Xiao X, Chen X, Zhang Y, Yoo SB (2019) Self-driving reconfigurable silicon photonic interconnects (flex-lions) with deep reinforcement learning. Supercomput Poster, 118
86. Prokudin S, Black MJ, Romero J (2020) Smplpix: neural avatars from 3d human models. arXiv preprint arXiv:2008.06872
87. Pujades S, Mohler B, Thaler A, Tesch J, Mahmood N, Hesse N, Bülthoff HH, Black MJ (2019) The virtual caliper: rapid creation of metrically accurate avatars from 3d measurements. IEEE Trans Vis Comput Gr 25(5):1887–1897
88. Ramírez I, Cuesta-Infante A, Schiavi E, Pantrigo JJ (2020) Bayesian capsule networks for 3d human pose estimation from single 2d images. Neurocomputing 379:64–73. https://doi.org/10.1016/j.neucom.2019.09.101
89. Rane C, Mehrotra R, Bhattacharyya S, Sharma M, Bhattacharya M (2020) A novel attention fusion network-based framework to ensemble the predictions of CNNs for lymph node metastasis detection. J Supercomput 77:1–20
90. Rhodin H, Salzmann M, Fua P (2018) Unsupervised geometry-aware representation for 3d human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 750–767
91. Rogez G, Weinzaepfel P, Schmid C (2017) LCR-Net: localization-classification-regression for human pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3433–3441
92. Romero J, Tzionas D, Black MJ (2017) Embodied hands: modeling and capturing hands and bodies together. ACM Trans Gr 36(6):245
93. Sabour S, Frosst N, Hinton G (2017) Dynamic routing between capsules
94. Sattar H, Krombholz K, Pons-Moll G, Fritz M (2019) Shape evasion: preventing body shape inference of multi-stage approaches. arXiv preprint arXiv:1905.11503
95. Sharma S, Varigonda PT, Bindal P, Sharma A, Jain A (2019) Monocular 3d human pose estimation by generation and ordinal ranking. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2325–2334
96. Sharma V, Srinivasan K, Kumar R, Chao HC, Hua KL (2017) Efficient cooperative relaying in flying ad hoc networks using fuzzy-bee colony optimization. J Supercomput 73(7):3229–3259
97. Shi D, Wei X, Yu X, Tan W, Ren Y, Pu S (2021) Inspose: instance-aware networks for single-stage multi-person pose estimation. arXiv preprint arXiv:2107.08982
98. Singha T, Pham DS, Krishna A, Dunstan J (2020) Efficient segmentation pyramid network. In: International Conference on Neural Information Processing. Springer, pp 386–393
99. Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5693–5703
100. Sun X, Shang J, Liang S, Wei Y (2017) Compositional human pose regression. In: 2017 IEEE International Conference on Computer Vision (ICCV). https://doi.org/10.1109/iccv.2017.284
101. Sun X, Wei Y, Liang S, Tang X, Sun J (2015) Cascaded hand pose regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 824–832
102. Sun X, Xiao B, Wei F, Liang S, Wei Y (2018) Integral human pose regression. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 529–545
103. Tang D, Chang HJ, Tejani A, Kim TK (2016) Latent regression forest: structured estimation of 3d hand poses. IEEE Trans Pattern Anal Mach Intell 39(7):1374–1387
104. Tekin B, Katircioglu I, Salzmann M, Lepetit V, Fua P (2016) Structured prediction of 3d human pose with deep neural networks. In: Proceedings of the British Machine Vision Conference 2016. https://doi.org/10.5244/c.30.130
105. Toshev A, Szegedy C (2014) Deeppose: human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1653–1660

106. Toshpulatov M, Lee W, Lee S (2021) Generative adversarial networks and their application to 3d face generation: a survey. Image Vis Comput 108:104119. https://doi.org/10.1016/j.imavis.2021.104119
107. Toutouh J, Esteban M, Nesmachnow S (2020) Parallel/distributed generative adversarial neural networks for data augmentation of covid-19 training images. In: Latin American High Performance Computing Conference. Springer, pp 162–177
108. Tseng KK, Zhang R, Chen CM, Hassan MM (2020) Dnetunet: a semi-supervised CNN of medical image segmentation for super-computing AI service. J Supercomput 77:1–22
109. Verma N, Boyer E, Verbeek J (2018) Feastnet: feature-steered graph convolutions for 3d shape analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2598–2606
110. Wan C, Probst T, Van Gool L, Yao A (2017) Crossing nets: Combining GANs and VAEs with a shared latent space for hand pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 680–689
111. Wang N, Zhang Y, Li Z, Fu Y, Liu W, Jiang YG (2018) Pixel2mesh: generating 3d mesh models from single RGB images. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 52–67
112. Wang Z, Peng Y, Yu ZZG, Sun J et al. (2018) Cascaded pyramid network for multi-person pose estimation
113. Wu E, Koike H (2019) Futurepose-mixed reality martial arts training using real-time 3d human pose forecasting with a RGB camera. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, pp 1384–1392
114. Xiang D, Joo H, Sheikh Y (2019) Monocular total capture: posing face, body, and hands in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 10965–10974
115. Xiao B, Wu H, Wei Y (2018) Simple baselines for human pose estimation and tracking. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 466–481
116. Xiao SSJ (2016) Deep sliding shapes for amodal 3d object detection in RGB-D images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 808–816
117. Xiao Y, Yu D, Wang X, Lv T, Fan Y, Wu L (2020) Spcnet: spatial preserve and content-aware network for human pose estimation. In: ECAI
118. Yang W, Ouyang W, Wang X, Ren J, Li H, Wang X (2018) 3d human pose estimation in the wild by adversarial learning. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2018.00551
119. Ye Q, Yuan S, Kim TK (2016) Spatial attention deep net with partial PSO for hierarchical hybrid hand pose estimation. In: European Conference on Computer Vision. Springer, pp 346–361
120. Yong B, Zhang G, Chen H, Zhou Q (2017) Intelligent monitor system based on cloud and convolutional neural networks. J Supercomput 73(7):3260–3276
121. Yuan S, Ye Q, Garcia-Hernando G, Kim TK (2017) The 2017 hands in the million challenge on 3d hand pose estimation. arXiv preprint arXiv:1707.02237
122. Yun K, Park J, Cho J (2020) Robust human pose estimation for rotation via self-supervised learning. IEEE Access 8:32502–32517. https://doi.org/10.1109/ACCESS.2020.2973390
123. Zhao L, Peng X, Tian Y, Kapadia M, Metaxas DN (2019) Semantic graph convolutional networks for 3d human pose regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3425–3435
124. Zheng C, Wu W, Yang T, Zhu S, Chen C, Liu R, Shen J, Kehtarnavaz N, Shah M (2020) Deep learning-based human pose estimation: a survey. ArXiv arXiv:abs/2012.13392
125. Zhou X, Huang Q, Sun X, Xue X, Wei Y (2017) Towards 3d human pose estimation in the wild: a weakly-supervised approach. In: 2017 IEEE International Conference on Computer Vision (ICCV). https://doi.org/10.1109/iccv.2017.51
126. Zhou X, Wan Q, Zhang W, Xue X, Wei Y (2016) Model-based deep hand pose estimation. arXiv:1606.06854
127. Zimmermann C, Ceylan D, Yang J, Russell B, Argus M, Brox T (2019) 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, pp 813–822

Authors and Affiliations

Mukhiddin Toshpulatov[1] · Wookey Lee[1] · Suan Lee[2] · Arousha Haghighian Roudsari[3]

Mukhiddin Toshpulatov
muhiddin1979@inha.edu

Suan Lee
suanlee@semyung.ac.kr

Arousha Haghighian Roudsari
arousha.haghighian@inha.edu

[1] Biomedical Science and Engineering, Inha University, 100 Inha-ro, Michuhol-gu, Incheon 22212, South Korea

[2] School of Computer Science, Semyung University, Jecheon 27136, South Korea

[3] Industrial Engineering, Inha University, 100 Inha-ro, Michuhol-gu, Incheon 22212, South Korea