Multimedia Systems
https://doi.org/10.1007/s00530-022-00980-0

REGULAR PAPER

A comprehensive survey on human pose estimation approaches

Shradha Dubey[1] · **Manish Dixit**[1]

Received: 23 November 2021 / Accepted: 2 July 2022

Abstract
The human pose estimation is a significant issue that has been taken into consideration in the computer vision network for recent decades. It is a vital advance toward understanding individuals in videos and still images. In simple terms, a human pose estimation model takes in an image or video and estimates the position of a person's skeletal joints in either 2D or 3D space. Several studies on human posture estimation can be found in the literature, however, they center around a specific class; for instance, model-based methodologies or human movement investigation, and so on. Later, various Deep Learning (DL) algorithms came into existence to overcome the difficulties which were there in the earlier approaches. In this study, an exhaustive review of human pose estimation (HPE), including milestone work and recent advancements is carried out. This survey discusses the different two-dimensional (2D) and three-dimensional human (3D) pose estimation techniques along with their classical and deep learning approaches which provide the solution to the various computer vision problems. Moreover, the paper also considers the different deep learning models used in pose estimation, and the analysis of 2D and 3D datasets is done. Some of the evaluation metrics used for estimating human poses are also discussed here. By knowing the direction of the individuals, HPE opens a road for a few real-life applications some of which are talked about in this study.

Keywords Human pose estimation · Deep learning · 2D/3D pose estimation · Activity recognition

1 Introduction

Human pose estimation (HPE) is described in the images or videos as the problem of locating human joints (keypoints such as elbows, wrists, etc.). It is one of the challenging tasks in the area of computer vision, its interpersonal occlusions, perfect estimation of the joints, and clothing make it difficult to estimate. In addition to this, it has several applications in real-life scenarios. Basically, it is a method of estimating the body configuration (pose) from a single, usually monocular image [1]. The problem of HPE may depend on many axes such as single person or multi- person human pose estimation, as the name suggests this estimation depends on the number of people present in a frame. The other factor on which HPE counts can be human body models which rely on shape, cone, and mesh-based models. The way in which the features are extracted also plays a vital role in estimating the human poses. The pose estimation for deep learning consists of views which are a holistic view (depends on joint position regression), local view (body parts detection), combining global and local information (body parts detection + joint position regression), and many others like motion features, pose recognition in videos, foreshortening [2]. In recent times, while digging deeper into this area, researchers are facing numerous obstacles like disorganization of background, complication in poses, foreground occlusion, and lack of robustness. Furthermore, high space and time complexity also restrict its application value. Therefore, various deep learning algorithms have been introduced to overcome these challenges. This paper extensively reviews how the HPE is enhanced from traditional ways to deep learning-based 2D/3D human posture estimation strategies from images or video recordings.

Communicated by R. Huang.

✉ Shradha Dubey
dubeyshradha29@gmail.com

Manish Dixit
dixitmits@mitsgwalior.in

[1] Department of Computer Science and Engineering, Madhav Institute of Technology & Science, Gwalior, MP, India

Published online: 16 August 2022

1.1 Problem definition

This is a difficult task due to strong articulations, small and hardly visible joints, occlusions, clothes, and illumination variations.

There are many more different problem axes on which human pose estimation models can be classified-

- Type of input modality (Red–Green–Blue (RGB) image, Depth (Time of Flight) image, Infra-red (IR) image)
- Number of cameras (Singleview or multiview)
- Human body models (Kinematic, planar, or volumetric)
- Type of images (Static or frames from video sequences)
- Number of people being tracked i.e. (single person or multiperson pose estimation)
- 2D/3D pose estimation

Each of these problem definitions is discussed in further sections of the paper.

1.2 Motivation

- A complete evaluation of modern deep learning and traditional 2D and 3D HPE approaches is offered by organizing them based on a 2D or 3D environment.
- A brief overview of major deep learning models such as openpose, deepcut, alphapose, mask RCNN, and Iterative Error Feedback are given
- An examination of a variety of applications of 2D/3D HPE, including gaming, training robots, medical, activity recognition, augmented/virtual reality, and animation.
- In consideration of key issues in HPE, an informative assessment of 2D and 3D HPE is offered, with future perspectives towards enhancing performance.
- In addition to this, different types of datasets that are used for the detection of 2D/3D pose estimation and performance metrics that help in evaluating a model are discussed in detail. Moreover, a separate description of 2D and 3D datasets are mostly missing in many of the surveys.
- The paper also reviews the different deep learning adversarial security attacks which may harm the well-implemented deep learning models.

The purpose of writing this survey is that it covers almost all the aspects of HPE along with the applications, challenges, research gaps, datasets, evaluation metrics, and security attacks which makes it different from the other studies in the literature. Thus, this paper is much more different and better than the other studies as nowhere in the published surveys are the discussion about the security attacks and challenges. The main pains of this area i.e., the effect of occlusion and crowd on time complexity, accuracy, and what are the measures to prevent this in an image are separately taken into consideration. Thus, covering all the areas in a single survey is not yet seen in any of the surveys in the literature hence, it is important. Last but not least most of the open areas for future researchers are well defined so that one could easily find a way for future findings in this particular field.

PE has comprehensive and multiple uses, response to circumstances of body reality. Some of the applications of HPE are as follows.

1.2.1 Activity recognition

Activity Recognition is an estimation of a person's identity, personality, knowledge in a range of application areas ranging from user interfaces to health monitoring, security, and protection. They have shown great advances, particularly for the task of pose estimation, as they can draw an appropriate feature when discriminating together [3]. One of the key topics for studying the science of computer vision and machine learning is the human capacity to perceive the actions of another.

Chengjun Chen et al. [4], The current study looks at how object recognition and pose prediction technology can be used in computer vision to identify a worker's assembly action. Furthermore, an efficient deep learning algorithm is used to estimate the running times of repeatable and tool-dependent assembly behavior. Instead of using the traditional behavior recognition algorithm, the YOLOv3 object detection algorithm is used in this analysis. It was discovered that the YOLOv3 scheme has a major impact on assembly action identification, with an accuracy rate of 92.8%. In addition, the CPM pose estimation algorithm is used to obtain the human body's joint knowledge. The judgment of running times for repeated assembly actions is 82.1% accurate.

Murilo Varges et al. [5], suggested a new way of depicting 2D poses in our paper. The 2D position is first translated to parameter space, where every other segment is allocated to a point, rather than utilizing straight line segments simply. Then, using a Bag-of-Poses technique, spatiotemporal characteristics from the parameter space are retrieved and encoded, and utilized to recognize human action in the movie. Experimental studies on two well-known public datasets, Weizmann and KTH, revealed that using 2D poses encoded in the spatial domain, the proposed approach would increase recognition rates while maintaining competitive accuracy rates as compared to other methods.

Francisco Javier Ordonez and Daniel Roggen [6], This paper are centered on convolutional and LSTM recurrent

units, and a generalized deep architecture for behavior recognition is suggested. This network is appropriate for multimodal wearable sensors; automatically performs sensor fusion; does not necessitate professional expertise in feature design, and specifically models the temporal dynamics of feature activations. The system is tested on two databases, one of which was used in a public activity identification competition. The framework works better than other deep learning models defined for activity recognition. Using only accelerometers, the model achieves a 0.69 F1 score. As accelerometers and gyroscopes are combined, accuracy increases by 15% on average, and by 20% when accelerometers, gyroscopes, and magnetic sensors are combined. This provides a trade-off for applications that include the automated processing of sensor data from various sources.

1.2.2 Medical

The pose measure has been used to classify postural problems, such as scoliosis, through the examination of irregularities in the posture of the patient and physical therapy [7].

Kenny Chen et al. [8], suggested a semi-automated method for optimizing upper-body pose prediction in noisy clinical areas, in which the expansion and development of a joint monitoring system (Caffe Heatmap) are done to enhance its reliability to uncertainties. The developed system employs subject-specific CNN models trained on a segment of a sufferer's RGB video recording selected to optimize each joint's function variance. Besides that, the expanded system yields more robust and precise posture annotations by remunerating cene lighting variations and optimizing the expected joint trajectories through a Kalman filter with fitted noise parameters. The system achieves an average accuracy of 96.5%, 96.8%, and 82.5% in all three subjects.

1.2.3 Training robots

Care robots are increasingly anticipated in the light of rising birth rates, an aging population, and a shortage of care staff. For instance, robots are supposed to monitor the condition of residents when patrolling the facility while taking care of nursing homes and other such establishments. Although the initial estimate of a position (standing, sitting, falling, etc.) is useful for assessing the condition of an individual, most methods to date use photos.

Md Jahidul Islam et al. [9], a framework presented for determining the 3D relative pose of pairs of talking robots using human pose-based key points as correspondences. To begin, an OpenPose is employed to extract pose-based 2D key points for the humans in the image. Following that, an iterative optimization algorithm is used to optimize the key-point correspondences based on their local structural similarity in the image space. Finally, the follower robots solve the perspective-n-point (PnP) problem to align the corresponding 2D predictions on their respective measured cameras and locate their relative poses. The model achieves an accuracy of 75.67, 57.82, and MAP 72.26, 54.91 on the Market-1501 dataset CUHK-03 dataset respectively.

Zimmermann et al. [10], developed a CNN-based method for predicting 3D human pose in real-world units by combining color and depth detail. The outperforms monocular 3D pose estimation approaches based on color and pose estimation based solely on depth. The method works on two datasets namely, MKV and captures dataset. The device is used in conjunction with a demonstration-based learning platform to teach a service robot without the use of markers. Studies in real-world environments show that the proposed method allows a PR2 robot to mimic the manipulation behavior of a human instructor.

Vasileiadis et al. [11], This paper explores a human pose prediction and monitoring system capable of precise and stable real-time output in real-life applications, with a focus on robotic-assisted living application fields. The method extends the effectiveness of the articulated-SDF monitoring approach by adding a set of complementary frameworks to address issues that occur in real-world situations, such as tracker activation and malfunction, pose monitoring from partial-views, broad object management, and body component convergence. Three publicly accessible human pose tracking datasets, SMMC-10, EVAL, and PDT, are used to test the overall evolved pose tracking process. On a practical human motion dataset, the experimental system achieves positive results: ADE 0.075, mAP 0.825, and outperforms the Kinect built-in pose estimator.

1.2.4 Animation

Character animation was historically a manual process. Nevertheless, postures may be specifically matched with an individual actor by advanced posture evaluation frameworks. More established frameworks depended on markers or special suits. Late advances in pose estimation and motion selection have made unmarkable, often real-time applications [12].

Kumarapu and Mukherjee et al. [13], In this work, AnimePose is introduced, a supervised multi-person 3D poses estimation and animation system for a given RGB video set. The work is divided into various modules like first estimating 2D poses then lifting 2D poses 3D poses. The work is done on a publicly accessible MuPoTS-3D dataset, the proposed solution achieves similar outcomes to previous state-of-the-art 3D multi-person pose prediction approaches, and it also outperforms previous competitive human pose monitoring methods by a substantial margin of 11.7 percent accuracy benefit on MOTA score on the Posetrack 2018 dataset.

Mohit Tiwari et al. [14], a method is introduced for estimating a human's 2D pose from a video that is both fast and effective. The method employs an affinity vector field (AVF) that discovers the relationships between body parts. The architecture uses global representation codes, allowing for a bottom-up approach with high accuracy in real-time. Moreover, the architecture is designed to understand and associate part locations together. The MPII dataset is used in this study.

For the application of Human Pose, Evaluation Amazon Go presents an important area. The camera track recognizes people and their actions, which is an important component of Pose Estimation. The services that monitor and measure human activities are largely based on an estimate of human employment. Other applications of HPE are gaming, motion tracking Casado García et al. [15], motion capture, augmented reality, and many more.

Albert Cleetus [16], In this paper, deep CNN is used for developing motion capture for animation. The images are captured live which will be helpful in real-time pose estimation for gaming applications. The captured video is then passed through various steps before mapping to 3D character animation. These steps are object detection, cropping region of interest, 3D pose reconstruction, and data normalization. The developed model can be used with ease and is affordable. The methods were put to the test on 10,000 images, and the results of the 3D pose prediction were compared to the GT data.

1.3 Challenges

The main challenges for human pose estimation are a variation of body poses, complicated background, and depth ambiguities. To solve these problems, considerable research efforts have been devoted to the related fields.

1.3.1 Lack of accuracy

The ultimate goal of every type of model is to achieve a better accuracy i.e., must perform its task efficiently and precisely. Therefore, the implemented algorithm must be good at its accuracy. Most of the time, it has been observed that accuracy has been neglected as one is focusing on other parameters while implementing an algorithm one may neglect accuracy.

1.3.2 Occlusion

Occlusion can occur in both single and multiple human pose estimation. It mainly creates problems in detecting accurate human poses in an image or video because the part of the human body gets occluded thus becoming an obstacle while detection. Thus, different occlusion handling techniques are proposed to overcome this problem.

1.3.3 Crowd

Alike occlusion, crowd also becomes an obstacle in detecting poses but it can occur only in the multiperson human pose estimation approach.

1.3.4 Time complexity

The algorithm should be designed in such a way that it should be time-efficient. It has been observed from the literature that most of the traditional approaches require much time to achieve better results and some of the deep learning approaches also do not have good results in case of time complexity if the selected data is not proper. Thus, time complexity constraints must always be in mind while designing an algorithm.

1.3.5 Preprocessing

Preprocessing is also one of the most challenging parts of HPE. The localization of body parts, background subtraction, data calibration, and image conversion thus plays a major role while detecting poses.

1.3.6 Data security

One of the most common concerns in ML/DL is security and privacy. Once a corporation has uncovered the data, security is a critical concern that must be addressed. To execute this accurately and efficiently, distinguishing between sensitive and insensitive data is critical.

These are some of the challenges which should not be ignored while implementing a human pose estimation algorithm.

1.4 Paper organization

The paper is divided into many sections. Section 2 gives a brief about the different human body models as they play a fundamental role in detecting different human poses. Next, Sect. 3 discusses the classical and deep learning approaches of 2D pose estimation and its classification into a single person and multiperson pose estimation. The single-person pose estimation is further classified into direct and heatmap regression. In the summary of 2D pose estimation, the comparison of all these approaches is done along with their advantages and disadvantages. Similarly, in Sect. 4, classical ad deep learning approaches to 3D human pose estimation are discussed and these approaches are also divided into monocular, a single person, and multiperson

3D pose estimation, and at last, the comparison of all these approaches is done.

In Sect. 5, some of the major deep learning models for HPE are discussed. Further in Sect. 6, the separate description of each of the 2D and 3D datasets are given, and also the evaluation metrics are discussed, at last, the paper binds up with the conclusion and future work for further research on this particular topic.

2 Human body models

The position of human body parts is required to develop a human body representation from visual input data in human pose estimation [17]. As a result, human body models are a critical component of human pose estimation. It addresses highlights and key points extracted from input images. To define and predict human body postures and generate 2D or 3D poses, a model-based technique is typically applied. There are three types of pose estimation based methods (Fig. 1).

- Kinematic modeling: This is also known as a skeleton-based model, and it is used to estimate 2D and 3D poses. To reflect human body structure, this adaptable and comprehensible human body model comprises a variety of joint configurations and limb orientations. As a result, this model is utilized to represent the relationships between various body parts [18]. The kinematic model, on the other hand, has limitations when it comes to capturing texture or shape information.
- Planar model: It's also referred to as a contour-based model, and it's primarily used to represent 2D body contours. This model is used to illustrate the human body's appearance and shape. Usually, body parts are represented by multiple rectangles approximating the human body contours [17].
- Volumetric model: Typically, body parts are depicted by a series of rectangles that resemble the contours of the human body.

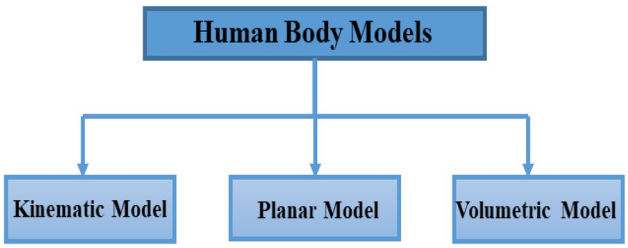

Fig. 1 Classification of Human body models

Figure 2 shows all three types of human body models. a shows how the various joints are determined with the help of key points represented by solid circles thus, helping in estimating poses. Similarly, b shows the rectangles.

3 2D pose estimation

The placement of key points in 2D space relative to an image or video frame is readily estimated with 2D pose estimation. It works by detecting and analyzing the X, and Y coordinates of human body joints in a picture [3]. 2D Pose Estimation is the process of identifying the location of body joints in an image (in terms of pixel values).

3.1 Classical approaches

There are different ways to deal with 2D human pose estimation and some general use classical approaches, for example, HOG, Edgelet, pictorial structure model (PSM) [19–21]. Some of the traditional works in the literature for analyzing the human stances are evaluated by making use of figure drawing for example by drawing cylinders for each body part and assessing the posture by joining these cylinders.

The pictorial structure parts are defined by pixel area and direction, which brings about the prevailing methodology for human pose estimation [19]. Pictorial structures have been proposed by M.A. Fischler [20]. It is a rearranged approach to portraying an item. The model has two components that

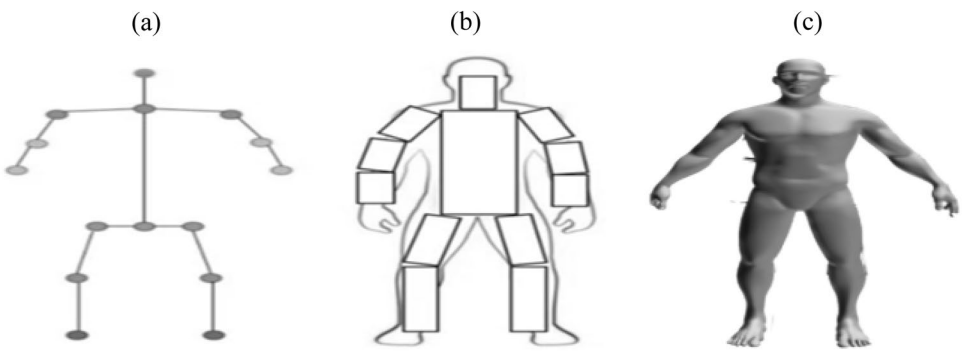

Fig. 2 Human body models from the left skeleton-based model; contour-based model; volume-based model [1]

comprise 2D image sections and structure which is an assortment of parts. An effective answer for pictorial structures was proposed by [2] as shown in Fig. 3. They showed how dynamic programming can be efficiently processed with pictorial structures if the representation has no cycles. It demonstrates 2D face recognition and human pose estimation applications. Thus, pictorial structures are appropriate for 2D human posture estimation. The above method, on the other hand, has the drawback of requiring a pose model that is independent of image data. Consequently, research has concentrated on improving the models' representational capability.

Interestingly, conventional models for object recognition use parts defined exclusively by areas, which disentangles both deduction and learning. Such models have been demonstrated to be exceptionally fruitful for object location. Along these lines, Yang [21] proposes a blended model of parts that presents complex connections, having a unary layout for and use in the assignment of recognizing explained individuals, and assessing their postures. The Part-based model excels at simulating articulations. However, this comes at the expense of restricted expressiveness and does not account for the global context (Fig. 4).

There are many shortcomings in traditional strategies, such as the tree model does not reflect all the limitations, the extraction execution of handcraft highlights is low and the complexity of reasoning is high. Due to certain parameters, the accuracy of the model gets disturbed and thus the model is not able to perform according to the expectations. Because of the limitations of conventional methodologies and to improve the presentation of assessing the postures, deep learning models have emerged, different models are dependent on the single person or multiperson. Some of the work based on deep learning approaches are reviewed below (Fig. 5).

3.2 Deep learning-based approaches

The traditional pipeline contains flaws, and pose estimation has been drastically altered by deep learning-based methodologies. These techniques are categorized as single person and multi-person pose estimation methods.

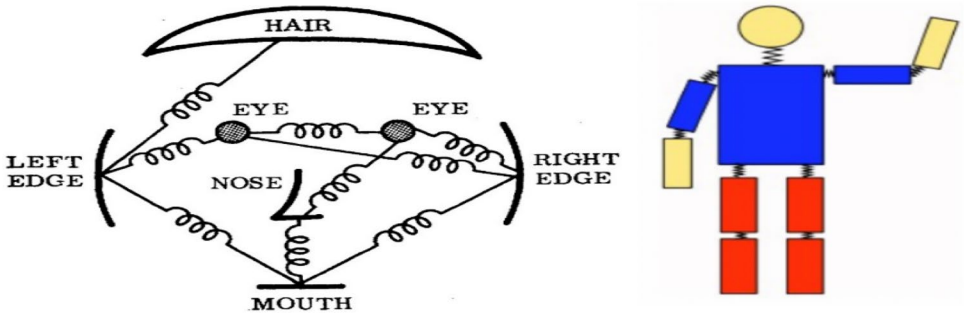

Fig. 3 Pictorial structural model [20]

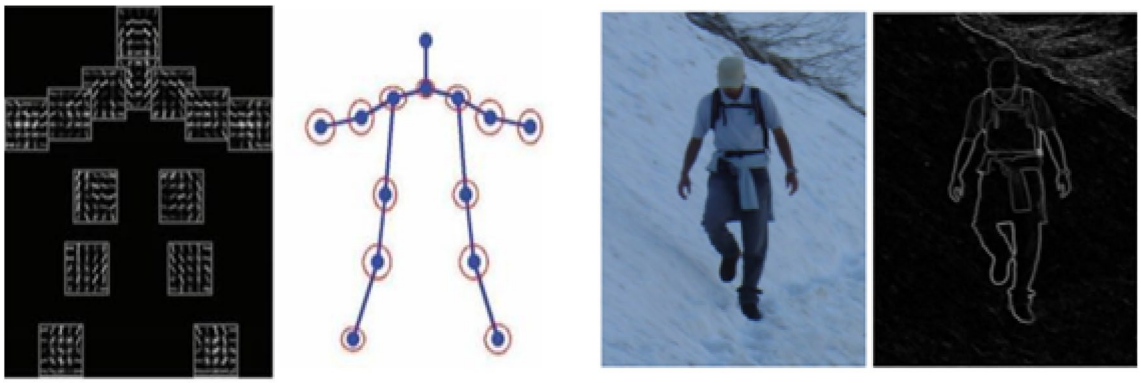

Fig. 4 Examples of **a** HOG features for keypoint detection [21] and contour Features from [22] **b** An original image; **c** Extracted contours

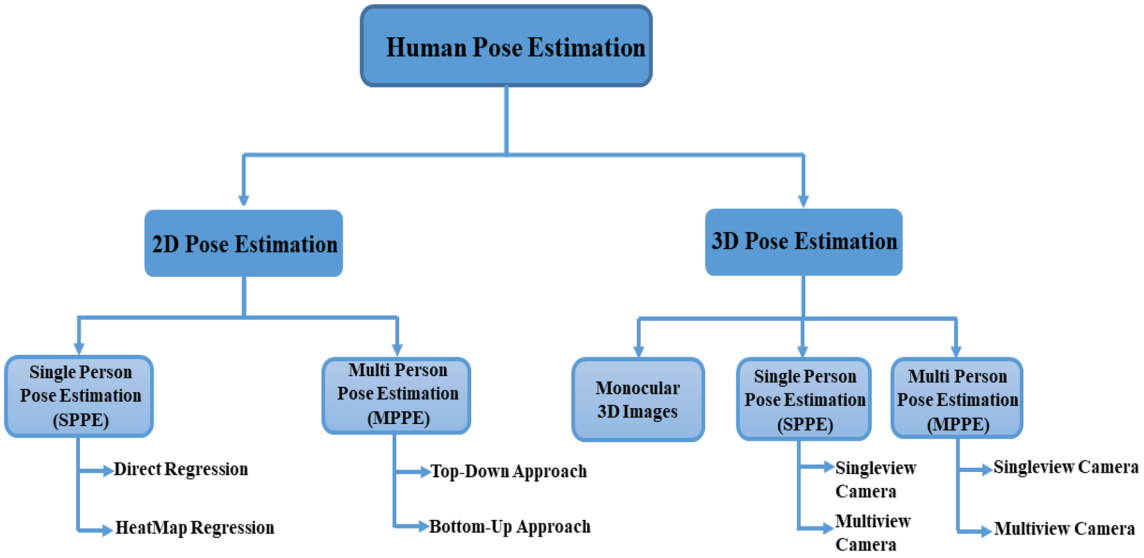

Fig. 5 The taxonomy used in this section

3.2.1 Single-person pose estimation (SPPE) methods

Single-person techniques detect a specific person's pose in an image. When an image contains multiple people, then the image is cropped in such a way that there is only one person left in the image. An upper-body detector [23] or a full-body detector [24] can perform this task automatically. The goal of single-person techniques is to locate the keypoint location in that area based on the given position information. Depending on how they predict key points, an SPPE pipeline is categorized into two categories.: KeyPoint Regression-based approaches and Heatmap-based approaches.

3.2.1.1 KeyPoint regression The model regresses the key body points directly from the feature maps in this manner, which is referred to as Direct Regression in some literature. The model's output will be a 17 by 2 vector comprising the X and Y coordinates of each anticipated key point if you want to estimate 17 key points for an individual using this method as shown in Fig. 6.

Several various models are presented to boost the effectiveness of the keypoint regression strategy in getting the correct points, such as Jaio Carreira et al. [26] suggested the Iterative Feedback method, which uses feedback to develop predictors that can efficiently manage complicated, structured output spaces. The primary goal is to learn several levels of feature extractors throughout their joint space to derive the 2D regions of a variety of key points from an RGB image, such as the ankle joint, shoulder, and so forth. Rogez et al. [27] developed the LCR network, in which each individual is first located, then categorized using a set of anchor postures, and finally regressed. A downside of this method is the large number of posture anchors required to achieve reliable results. Toshev and Szegedy [28] proposed and introduced a cascaded DNN regressor for directly predicting human keypoints. However, understanding mapping

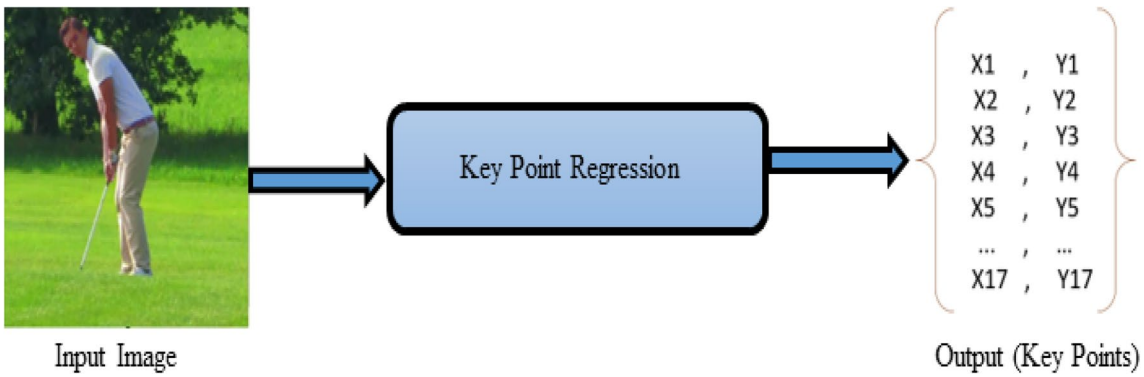

Fig. 6 KeyPoint regression [25]

straight from feature maps without the use of other processes is difficult.

3.2.1.2 Challenge Imagine a situation where the model chooses a specific key-point location with a deviation of one or two pixels from the underlying data. This little discrepancy in the model's estimation generates an error that disrupts the training process and prohibits the model from settling into an optimal solution; nonetheless, in many applications, even a small divergence in estimation is tolerable. As a result, training a model to explicitly identify the exact location increases the problem's complexity and sensitivity, as well as the model's training instability.

3.2.2 Heat map regression

Heatmap-based frameworks are now widely used in 2D HPE tasks. Heatmap-based framework regresses heatmaps first. In this framework, the probability of the existence of a key point in each pixel of the image is estimated. A more probable keypoint zones using a heat map is demonstrated. Many researchers have used heatmap regression models like Shih-En Wei et al. [29] propose Convolutional Pose Machines (CPMs) for the task of articulated pose estimation that consistently produces 2D conviction maps. At each phase in a CPM, features of the image and the conviction maps created by the past stage are given as input. The proposed multi-stage design is completely differentiable and accordingly can be prepared by utilizing back-propagation. The method is evaluated on the FLIC Dataset [30] and achieves PCK@0.2 with 97.59% on elbows and 95.03% on wrists. Newell et al. [31] reviewed that due to a stacked block architecture with numerous intermediary supervisions, a stacked hourglass network with convolutions in multi-level features was developed thus, allowing re-evaluation of prior estimations. Adrian Bulat et al. [32] suggested that even when there is considerable partial occlusion, the human pose can be detected by following their approach of pose estimation detection. On the heatmaps, this method employs detection and regression. Recently, Zhengxiong Luo et al. [33] proposes the scale-adaptive heatmap regression (SAHR) method, which can modify the standard deviation for each keypoint flexibility. SAHR is followed by weight-adaptive heatmap regression (WAHR), which helps in balancing the foreground and background samples. Extensive testing has shown that combining SAHR and WAHR increases the accuracy of bottom-up human posture assessment significantly. Chen and Yuille [34] used DCNN to learn pairwise relations using the graphical model (parts type and pairwise parts relationships) (Fig. 7).

3.2.2.1 Challenges Regression of heat maps faces two obstacles. First, there's the keypoint extraction issue utilizing heat maps (decoding problem), which can be solved by selecting the maximum or average of each heat map as a key-point position. The other issue is generating a ground truth [36]; because the model outcome is in the form of a heatmap, therefore, it is needed to convert the ground truth (which is made up of keypoint coordinates) to the same format (encoding problem).

3.2.3 Comparison of direct regression vs heat map regression

Direct regression of joint locations is very nonlinear, making learning mapping challenging [27]. It also can't be used in a multi-person situation. The heatmap-based system, on the other hand, regresses heatmaps first. Heatmaps can be used to aid human comprehension and model more complex scenarios. Direct regression is more trustworthy and has certain virtues because it is fast and simple [26–28]. When direct regression is used, the ultimate result can be determined from stem to stern without the need for heatmaps. It can also be used in 3D scenarios with minimal adjustments. Furthermore, integrating heatmaps with huge convolutional kernels

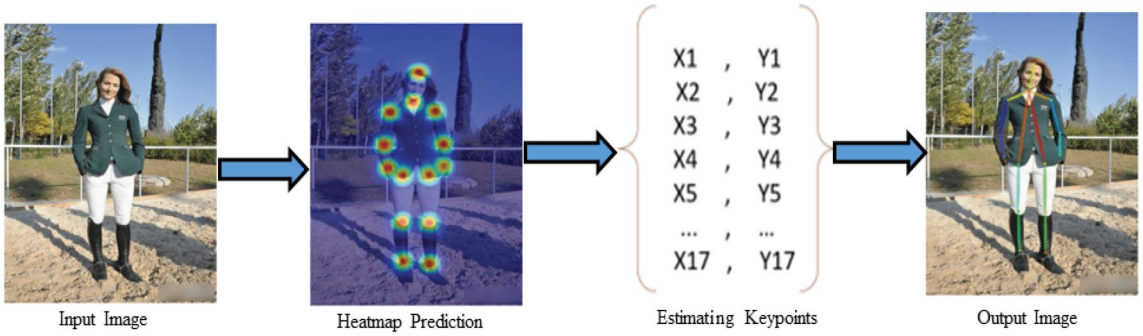

Fig. 7 Pose estimation using heat map regression [35]

and deeper models increases performance by expanding the relevant receptive field and hence the context acquired. [29, 31]. As the training proceeds, the erroneous response map is gradually suppressed, and the correct response map has gradually become strong. As a result, there seems to be no absolute answer to this question, and each paradigm has its own set of benefits and drawbacks.

3.2.4 Multi-person pose estimation

Because the position and volume of people in an image are uncertain, multi-person pose estimation is more complex than single-person posture estimation. In most cases, the problem can be addressed in one of two ways:

- The simplest method is to start with human detection, then estimate the parts, and finally calculate the pose for every individual [37, 38]. This approach is called the top-down approach.
- Another method is to identify all portions in the image (i.e., parts of each individual), then group parts that belong to different people. This is referred to as the bottom-up approach [39].

3.2.4.1 Top-down approach The top-down technique of HPE works in stages: initially, it detects the person from an observable frame, then it obtains that discovered object, and finally, it attempts to estimate 18 critical points for every identified person from in that frame before attempting to form the skeleton. He et al. merged segmentation and keypoint prediction in the Mask-RCNN model. To obtain a one-hot mask for each keypoint, the authors attach keypoint heads on top of RoI-aligned feature maps. For multi-scale inference, Chen et al. built globalnet on top of Feature Pyramid Networks and refined the predictions with hyper-features. [38]. Papandreou et al. proposed a model which works in two stages. Firstly, a faster RCNN detector is used to detect the location and scale of boxes and then the key points of the person are estimated [37]. In this approach, a novel aggregation is used which is used to aggregate these results and produce highly localized keypoint predictions by combining these outputs. Nelson Rodrigues et al., make use of time of flight (ToF) images and take out the ROI. Detecting joints on padded ROIs did significantly change the results and enabled the system to be more effective for joints near the ROI boundary [40].

3.2.4.2 Bottom-up approach On the other hand, the bottom-up approach firstly tries to estimate the 18 key points from the image for each person [39]. After that, it will try to estimate a person from those key points i.e. joining the key points and making the skeleton. Muhammed Kocabas et al. [41] use a pose residual network (PRN) that takes keypoint and the human detection process and assigns keypoints to human instances to generate valid poses. Iqbal et al. offer a new method for modeling multi-person pose computation and tracking in one formulation. The study resolves the densely connected graphical model locally, resulting in a significant increase in time efficiency. Insafutdinov et al. proposed DeeperCut [42] which improves DeepCut using deeper ResNet and employs image-conditioned pairwise terms to get better performance.

3.2.4.3 Comparison of top-down and bottom-up approaches In general situation, the top-down approach consumes time much more than the bottom-up, because the top-down approach needs N-times to pose estimation by person detector results. However, the bottom-up approach misses the opportunity to zoom into the details of each person's instance [38, 39]. As a result, there is a discrepancy in accuracy between these two approaches. Deep neural networks have been used to investigate both bottom-up and top-down techniques in the latest days. However, as the present scenarios are taken into account, it is impossible to say which strategy is better than the other. For multi-person pose estimate assessment, accuracy and speed are two main parameters.

Accuracy: In terms of the bottom-up pipeline, since the network cannot acquire reliable features from the frames, the scale variability of people might create issues with HPE. During the training phase, the average resolution of a single person in a bottom-up pipeline is relatively low than in a top-down pipeline using the same network and GPU storage [43]. As a result, what truly limits the accuracy of the bottom-up pipeline could be an equipment limitation.

Speed: Every human pose in the top-down pipeline is evaluated individually, which takes linear time as the number of people increases. Every person's pose in the top-down pipeline is evaluated individually, which takes linear time as the majority of individuals increases. As a result, in the bottom-up pipeline, increased speeds could be possible [43].

Therefore, deep learning-based approaches have achieved a breakthrough in HPE by improving performance significantly. In the following, section the comparative analysis of different deep learning-based 2D HPE methods concerning a single person and multi-person scenarios is done (Table 1).

4 3D pose estimation

Determining the articulated 3D joint (keypoints) regions of a human body from an image or video is known as three-dimensional (3D) HPE [52]. It calculates the 3D posture (x, y, z) of an RGB image. The purpose of 3D HPE is to use a

Table 1 Comparative analysis of 2D human pose estimation techniques

Technique	Single person	Multiperson	Advantage	Disadvantage	Performance
Real-time open pose [44]	–	Bottom-up approach	The first real-time system for detecting key points. Furthermore, it merges the body and the foot into a single model, increasing the accuracy and reducing the inference time	Greedy multi-person parsing fails in highly crowded images	Generates strong body pose parses and maintains efficiency regardless of the number of persons
Deep cut subset partitioning and labeling problem (SPLP)	–	Bottom-up approach	Empirical results on the four datasets show improvement in both single person and multi-person dataset	The presence of several occlusions in the background may degrade the result	The proposed model infers the number of persons, poses, spatial closeness, and part-level occlusions all at once
RMPE alpha pose [45]	–	Top-down approach	It improves pose estimation in the context of imprecise human boundary boxes and repetitive detections	it would be interesting if the framework is trained together with the human detector in an end-to-end manner	Works better in terms of accuracy and efficiency
Object localization using CNN [46]	Heatmap	–	Estimate the joint offset location within a small region of the image	The model is sensitive to poor lightning conditions, cluttered background	Quicker and more computationally efficient
Convolutional pose machines [29]	–	Multiperson	Without the use of explicit graphical model-style inference, the model delivers increasingly advanced estimates for part positions	Gets confused when multiple people are nearby	Achieve state-of-the-art results on standard benchmarks including the MPII, LSP, and FLIC datasets
Stacked hourglass network [31]	–	Repeatedly use Top Down and Bottom-up	Continuous performance across a range of conditions, including obstruction and multiple individuals in proximity	Sensitive to camera characteristics	Low computational cost
Deep high-resolution network HRNet [47]	Heatmap	–	Reliable, high resolution	The multiresolution representations are complex	Less complex, cost-effective
Cascaded pyramid network (GlobalNet, Refinenet) [16]	–	Top Down	A separate network is used to simply handle hard keypoints	Only capable of dealing with head pose changes, occlusion, and illustration	Low Computational cost
Real-time lightweight openpose[48]	–	Bottom-up	Suitable for real-time performance on edge devices	High Complexity	Heavily optimized network design and post-processing code
Deeply learned compositional model (DLCM)[49]	–	Top Down and Bottom-up	Use compositional patterns for pose estimation	Multiple humans cannot be identified in the same frame	Lower complexities, compactly encode orientations, scales, and shapes of parts
Transformer (TF) pose [50]	Improved regression	–	Removes different problems such as quantization error and non-differentiable post-processing	Does not able to produce better results for occluded images	Simple and direct framework
MoDeep CNN [51]	–	FLIC-motion	Incorporates both color and motion features to detect human poses	Moving objects in the backdrop can have a significant impact on the outcome	Even extremely modest temporal signals can boost performance with only a slight rise in complexity

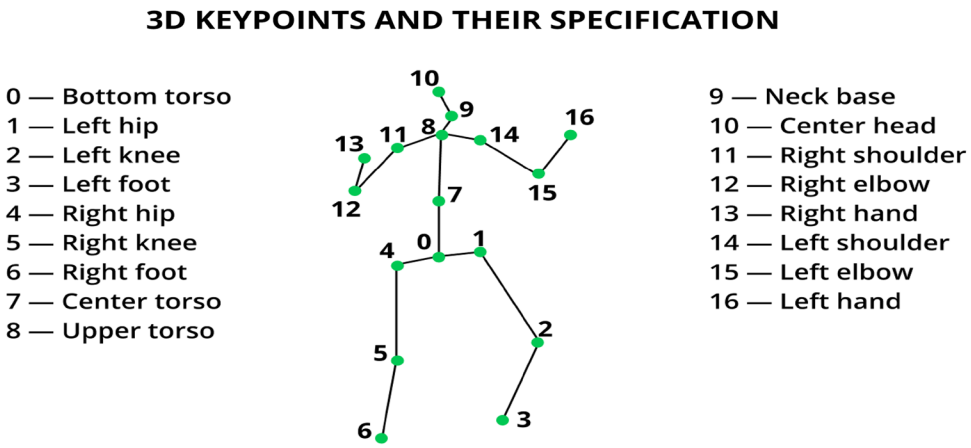

Fig. 8 Keypoints in three dimensions and their definition [53]

picture of a person to determine the XYZ coordinates of a particular set of keypoints on the human body. 3D keypoints are visually observed as follows (Fig. 8).

Following the extraction of joint positions, the activity assessment model examines a person's posture and examines the person's real motion in a series of frames from a video stream. Generally, it is harder to recuperate 3D poses from 2D RGB images because of more ambiguities in its estimation and it requires huge space as compared to 2D images [47].

The selection of the dataset images in a 3D pose is also a challenging task in its estimation. Moreover, the algorithm has to be invariant to many other factors such as textures, the skin color of the selected image, imperfections in an image background, human occlusions, and many more. The traditional and recent deep learning approaches for three-dimensional pose estimation are reviewed below.

4.1 Classical approaches

Pictorial structure models (PSM) are the accepted norm for 2D human posture estimation. It proposes a multi-view pictorial structures model that expands an ongoing advance in 2D estimation. The 3D PSM is non-exclusive and relevant to both single and multiple human posture estimations. It is essentially a generative model for pose estimation. PSM does not give many accurate results as they have the conditions between the output factors, subsequently, to consider these conditions SVM Hanguen Kim et al. [54], Ke Chen et al. [55] techniques are acquainted for the structured SVM utilization to learn the mapping from segmentation features to joint locations.

3D human stance estimation utilizing HOG highlights portrays the state of the object and is utilized to examine 3D human posture steadily. As the HOG highlights are registered over the whole picture, the HOG features measurement is high along these lines as shown in Fig. 3 and hence principal component analysis (PCA) is used on each HOG block [30]. Therefore, the 3D human pose can be assessed by the linear regression of HOG features.

This approach took first place in the inaugural COCO 2016 keypoints challenge, outperforming the previous state-of-the-art outcome on the MPII MultiPerson benchmark by a wide margin, both in performance and efficiency. MPII dataset result on the testing subset achieves a Map of 75.6% and results on the COCO 2016 keypoint challenge is 61.8% [52].

4.2 Monocular 3D human pose estimation

Among the most essential and difficult topics in computer vision is vision-based monocular HPE, which wants to acquire human postures from input images or frames. Due to depth uncertainties and occlusions, 3D HPE from monocular photos is inadequate.

In both 2D and 3D contexts, the monocular camera is by far the most extensively utilized detector for HPE. Researchers have been able to expand their findings to 3D HPE due to recent advances in deep learning-based 2D HPE from monocular photos and videos. Specifically, deep learning-based 3D HPE methods are further separated into two main groups: single-view 3D HPE and multi-view 3D HPE. Hallquist and Zakhor [56] use single-view pose estimation using mobile devices with a distance of 10 m over the query images. Fei et al. [57], calculate physical distances among individuals from just a single RGB image or video acquired by a camera observing a 3-D scene from a stable viewpoint.

3D keypoints are generally deduced using single-view photos, irrespective of methodology (image → 2D → 3D or image → 3D). Conversely, multi-view imaging can be employed, in which each frame is taken from multiple cameras focusing on the target object from various perspectives. The multi-view approach improves the perception of depth and is useful when certain sections of the body are obliterated in the image [58]. As a result, predicted values become even more precise.

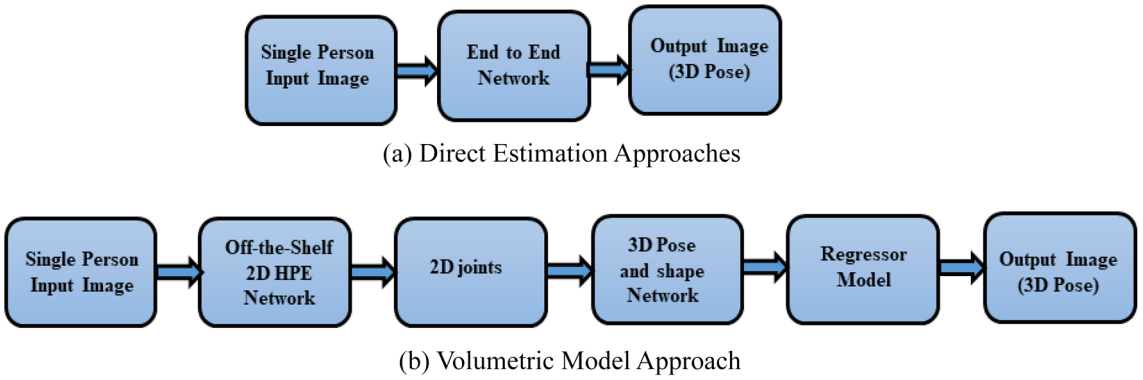

Fig. 9 Block diagram of **a** direct estimation approaches and **b** volumetric model approach for SPPE [61]

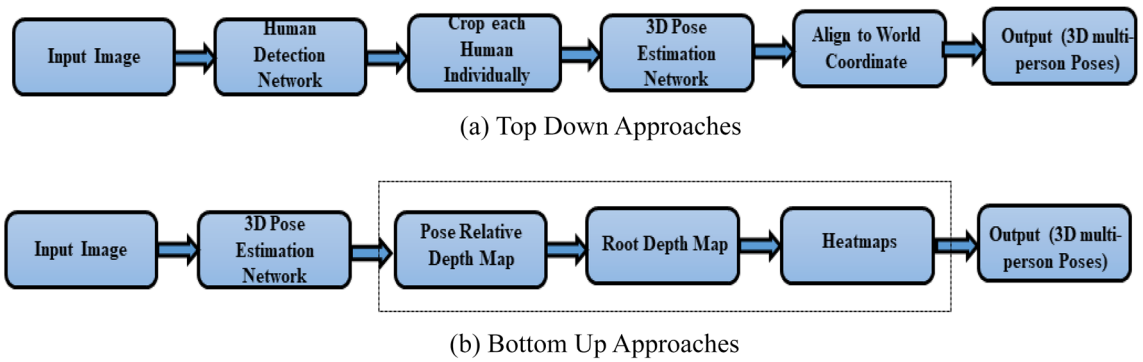

Fig. 10 Block diagram of **a** Top-down approach and **b** Bottom-up approach for 3D human pose estimation

Usually, this technique necessitates camera synchronization. Nevertheless, other researchers show that even asynchronous and uncalibrated video streams from several cameras may be used to predict 3D keypoints. Furthermore, there are other methods for estimating 3D human pose:

- To train a model that can infer 3D joints directly from the images given. For Eg, A multi-view system named EpipolarPose is trained to predict the locations of 2D and 3D keypoints simultaneously. The fascinating part is that it only requires 2D keypoints for training instead of ground truth 3D data [58]. Rather, it creates the 3D ground truth using epipolar geometry for 2D projections in a self-supervised manner. It's useful since a scarcity of elevated 3D pose annotations is a major issue when developing 3DHPE algorithms.
- Identifying 2D keypoints and then translating them into 3D is the most popular strategy because 2D keypoint projection is well-studied, and using a pre-trained framework for 2D predictions improves the system's accuracy rate [59]. Furthermore, several existing models offer reasonable accuracy and speed of inference in real-time (for example, PostNet, HRNet, Mask R-CNN, Cascaded Pyramid Network [16, 41, 47, 60]).

4.3 Single-person 3D pose estimation

Most works for estimating human pose for a single person use a single image/video. Despite the ambiguity in the depth dimension, models trained on 3D ground truth (GT) show pretty good performance for the case of a single person without occlusions. The figure below shows the direct estimation, 2D to 3D lifting approaches, and volumetric model approach (Fig. 9).

4.4 Multiperson 3D pose estimation

The main challenge in multi-person 3D pose estimation is occlusions. Due to a limitation of suitable datasets, progress on multi-person 3D pose estimation is inherently limited [62]. In addition, unfortunately, there are almost no annotated multi-person 3D pose datasets like the Human3.6 datasets. Most multi-person datasets either do not have good

Table 2 Comparative analysis of 3D human pose estimation techniques

Technique	Single person multiperson	Advantage	Disadvantage	Performance
Occlusion-robust pose-maps (ORPM) [64]	Monocular RGB	The proposed method works well even under strong inter-person occlusions and human interactions better than previous approaches	Inaccurate predictions when joints of the same type are nearby	Overall pose quality is improved
Deep convolutional neural network [65]	Monocular images	The network used has disentangled the dependencies among different body parts which makes it simpler to detect human poses	Not feasible for the multi-view pose estimation	The network achieves significant improvement over baseline methods
Single-view-multi-angle consistency (SVMAC) [66]	single view	It is desirable if the image contains only one human object	Not feasible When there is an occlusion in the images	Faster computation; Easy to implement;
Location-maps-based model [67]	Equirectangular Images	Implements the enhanced location maps-based model which takes both distorted and disconnected images into consideration	Not much robust to all the values of location map variances	the model indicates better performance with respect to accuracy and computation complexity
3D pictorial structures (3DPS) model [68]	Single person/Multiperson	Self and natural occlusions can be easily handled	Doest not work with real-time images	The 3D PCP scores for single human and multiple humans is 76 and 75.6 respectively
Efficient Pose [69]	Single-person	Practically used in-the-wild images even when end-to-end training is not feasible	Can take only a single image at a time	The 3D PCP scores for single human and multiple humans is 76 and 75.6 respectively
Deep Depth Pose (DDP) model [70]	Single view/Multiview	Deep Depth Model may generate a 3D posture by linearly combining prototype poses	The model is much more complex	Provides better accuracy on the dataset used
Fully Convolutional Model [71]	Video Based	The model is practical in scenarios where motion capture is challenging	Confused with occlusions and lightning conditions	It improves performance when labeled data is scarce
fully connected neural network with the SMPL [72]	Multiview	Improves image generalizations	Cannot restore posture distortion	the joints' average error of the proposed method is the smallest

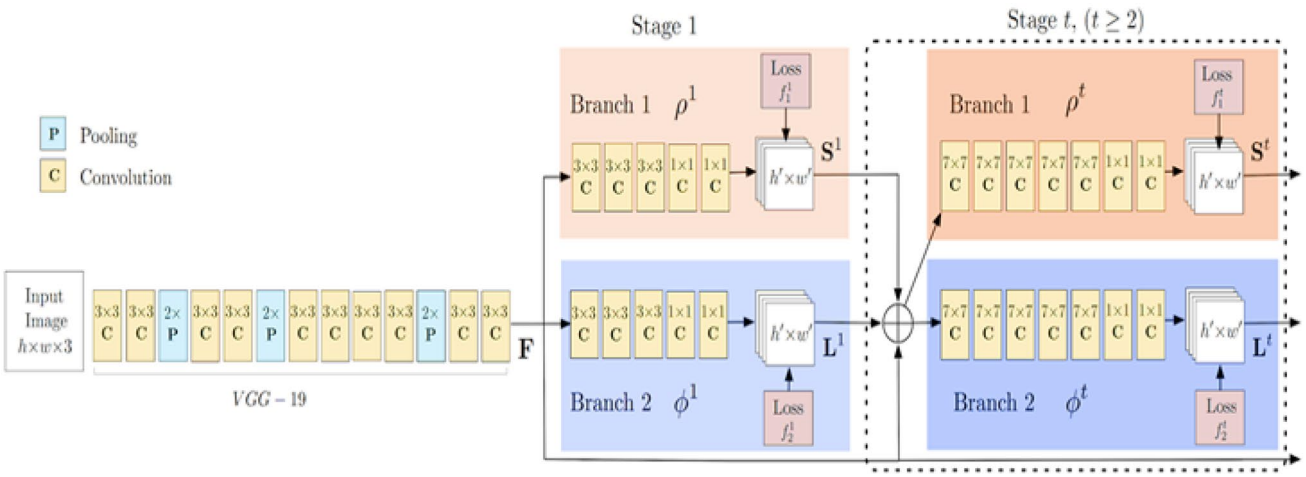

Fig. 11 OpenPose architecture [44]

ground truth or are not realistic. Abdallah Benzine uses PandaNet (Pose Estimation and Detection Anchor-based Network) for estimating multiperson 3D poses [63] (Fig. 10) (Table 2).

5 Major deep learning models in human pose estimation

5.1 OpenPose

OpenPose is the first real-time bottom-up multi-person system to detect human body key points (several 135 key points) on single images. Researchers at Carnegie Mellon University proposed it [44] (Fig. 11).

- Initially, an image is processed through a CNN network to retrieve the input's feature maps. The first ten layers of the VGG-19 network are used in the model.
- The part confidence maps (CM) and part affinity fields (PAF) are generated from the feature map using a multi-stage CNN pipeline.
- Each branch's forecasts are refined over successive stages. Bipartite graphs are created between pairs of parts using part confidence maps (as shown in the above image) [74].
- In multi-stage CNN, The PAF refine L^t from the extracted features of the base network F in the first set of phases.

$$L^t = \varphi^t\left(F, L^{t-1}\right), \forall 2 \le t \le T_{P,}$$

- The next step uses the prior layers' output PAF to refine the confidence map recognition.

$$S^{T_P} = \rho^t\left(F, L^{T_P}\right), \forall t = T_P,$$

$$S^t = \rho^t\left(F, L^{T_P}, S^{t-1}\right), \forall T_P < t \le T_P + T_c,$$

The greedy approach is then used to analyze the final S (CM) and L (PAF).

5.1.1 Advantages

- OpenPose is significantly more precise since it's designed to run on GPUs. OpenPose is free for non-profit usage and can be transferred below these terms.
- High accuracy without impacting implementation quality

5.1.2 Disadvantages

- The fundamental disadvantage of OpenPose is that its low-resolution findings restrict the level of information in keypoint predictions.
- It does not provide any statistics about the depth and is built on DNN, which demands a powerful machine.
- Speed and precision are slightly hampered.

5.2 DeepCut

DeepCut is a bottom-up approach for estimating multi-person human poses. The authors tackled the job at hand by identifying the significant information:

- Make a list of D body part candidates. This integration gives all probable body part locations for each person in

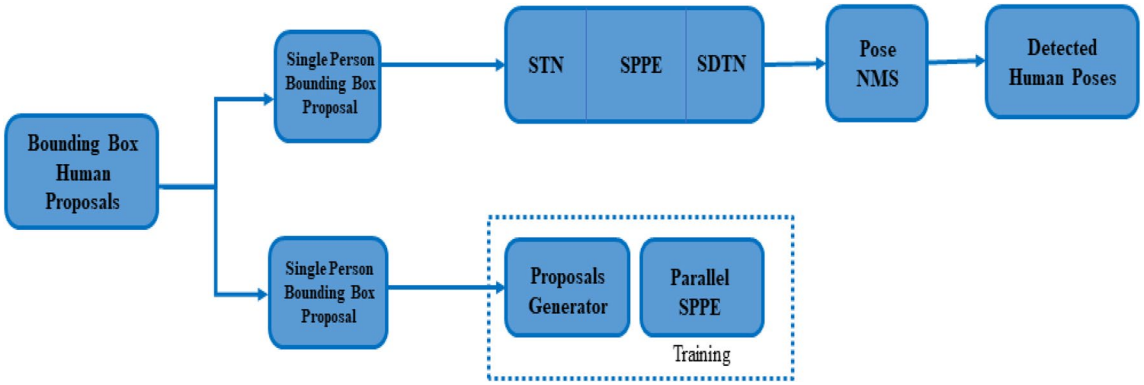

Fig. 12 RMPE Framework [73]

the image. Choose a subset of body parts from the list of candidates

- Each selected body part should be labeled with one of the C body part classes. These classes represent the various types of parts, such as "arm," "leg," "torso," and so on.
- Body parts that belong to the same person should be separated [75].

5.3 AlphaPose

AlphaPose is an accurate top-down multi-person pose estimator, which is the first open-source system. The model is used to solve the problem of multi-person pose estimation in the wild [73]. The AlphaPose framework consists of three components:

- Symmetric spatial transformer network (SSTN).
- Parametric pose non-maximum-suppression (NMS)
- Pose-guided proposals generator (PGPG).

- First, it has the bounding box proposals by the human detector. According to the paper, the authors used a VGG based SSD512 detector for detecting humans.
- These bounding box proposals are then fed into the Symmetric STN (Spatial Transformer Network) + SPPE module. This step generates the pose proposals.
- Now, there is also see a Parallel SPPE module in Fig. 12. This module is used during training to avoid the local minimums [76].
- The detected poses may also contain many redundant detections. To reduce these, the authors used a parametric Pose-NMS to eliminate the redundant poses.
- Also, to augment the training samples, the authors used a Pose Guided Proposals Generator (PGPG) during training.

5.3.1 Advantages

- Designed to eliminate flaws in the traditional system such as erroneous identification or localization.
- To boost performance, optimization of the network's hyperparameter was applied.
- When contrasted to commonly used one-stage process frameworks, the two-step framework results in better accuracy.

5.3.2 Disadvantages

- In instances, the two-step structure compromises speed or runtime efficiency.
- Does not work well as opposed to other traditional pose estimation approaches.

5.4 Mask R-CNN

MaskRCNN is a DNN designed to solve the problem of instance segmentation in ML or computer vision. Specifically, it can distinguish between various objects in an image or video. The fundamental design starts by applying a CNN to extract feature maps from an image [60]. A region proposal network (RPN) [24] uses the feature maps to find bounding box alternatives for the existence of objects. Because the bounding box candidates can be of different sizes, a layer named Region of Interest (RoI) Align would be used to minimize the dimension of the features extracted such that they were all of the same sizes. The obtained features are now forwarded to the parallel branches of CNNs for the last prediction of the bounding boxes and segmentation masks. The schematic diagram of Mask RCNN is shown in Fig. 13.

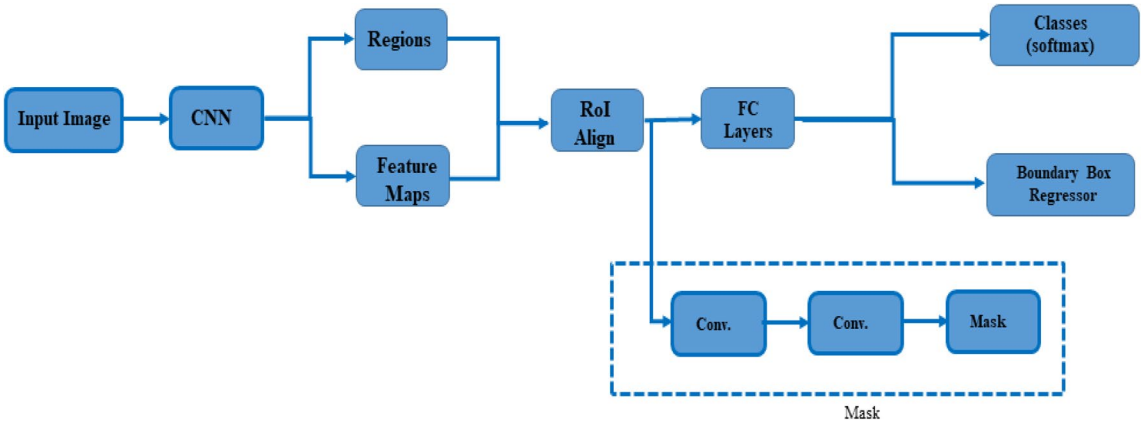

Fig. 13 Mask RCNN [60]

5.4.1 Advantages

- Experience and expertise are not required.
- For each case, creates a segmentation mask.
- Pipeline concurrently anticipates boxes and key points

5.4.2 Disadvantages

- Semantic segmentation can result in erroneous person bounding boxes, and thus erroneous joint postures.
- Quick, but not velocity optimized
- Leaves no room for a few failure circumstances and uncommon positions.

5.5 Iterative error feedback

IEF [77] is based on the idea of identifying what is problematic with the idea of prediction and resolving it iteratively, which is accomplished through a top-down feedback strategy. IEF used a system that integrates both input and output regions and expands the hierarchical feature extractor (ConvNet). Input consists of an image I and the originally predicted keypoints y_0 (representing the preceding output y_{t-1}) can be found on the left side of Fig. 13. Consider three key points: the head (red), right wrist (green), and left wrist (blue).

Next set input, $X_t = I \oplus g\ (y_{t-1})$,

where, I denote the image and y_{t-1} is the output from the previous stage.

The output known as correction ε_t is produced by the function $f\ (X_t)$, which is treated as a ConvNet, and this output is combined with the current output y_t to generate y_{t+1}, which signifies the correction is taken into account t [26]. Each keypoint is converted into one Gaussian heatmap channel using the function g (y_{t+1}), which may then be used as part of the picture input for the subsequent iteration. This technique is performed T times until a refined y_{t+1} is obtained that is extremely near to the underlying data. This method assessed their effectiveness on two datasets (LSP and MPII) using a PCKh@0.5 metric. IEF brought in new ideas and high-quality work. Both f and g are learnable functions, and f is also a ConvNet. This means that f can learn features across the joint input–output space.

By considering the system above it can be observed that deep learning-based approaches have achieved a breakthrough in HPE by improving the performance significantly.

5.5.1 Advantages

- The solution is constructed and improved by a series of steps iteratively. As a result, one can detect flaws in the preliminary phase.
- Iterative models allocate minimal time to documentation and more duration to develop.

5.5.2 Disadvantages

- There are no crossovers between the phases of an iteration.
- Since not all needs are obtained up front for the full lifecycle, there could be costly system architecture or design concerns.

These are some of the famous deep learning models which are used for the pose estimation. As we have seen from the above details each of the model has its own advantages and disadvantages thus, each of these models is better for the type of application they are applied on.

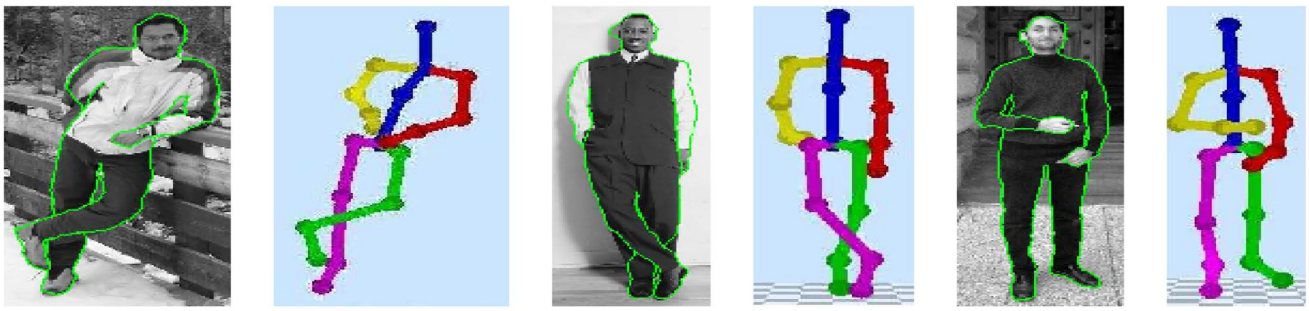

Fig. 14 Different ways of self-occlusion [82]

6 Occlusion and crowd detection human pose estimation

6.1 Occlusion detection and handling

Occlusion, in which one or even more joints are obscured from either the camera due to numerous variables such as self-occlusion, zooming, blurred objects in an image, or interference by random items, is a major difficulty for human posture assessment [78]. In a large percentage of real-world photographs, occlusion happens spontaneously. Occlusion has been a significant problem in HPE algorithms, particularly those that use deep learning [79]. This problem leads to inconsistencies in training and testing, which results in inaccuracy. The ability to handle occlusion, including self-occlusion [80] and occlusion between people, is a critical subject in this area. This all depends on the type of data, including the quality and quantity of the image.

In general, there are three types of occlusions are there in an image- self-occlusion, inter-object occlusion, and background occlusion.

6.1.1 Self-occlusion

In virtual settings, self-occlusion is a difficult challenge to solve. It arises when a section of an image overlaps the self, obscuring a part of the image that is ordinarily obscured from view by another part of the object [81, 82]. It is not that problematic in static models, but in malleable objects that are changed in real-time while processing.

6.1.2 Inter-object occlusion

When the relevant object is obscured by other objects, this is known as inter-object occlusion. Alike objects, in HPE multiple persons in crowded environments are difficult to detect since they are frequently partially or completely obscured by each other. Therefore, it is very important to handle this occlusion otherwise it will become one of the major drawbacks of this field [83].

6.1.3 Background occlusion

When a background object obscures the monitored objects, this is known as background occlusion (Fig. 14).

The review of each of these ways through which occlusion arises using different traditional and deep learning methods is discussed below and then the challenges and drawbacks of the suggested methods are observed.

Ghafoor et al. [78], 2022, Occlusion guidance, which offers additional details regarding the lack or existence of a joint, is used to address missing joints. Temporal data was also used to improve the estimation of missing joints. a huge series of researches are conducted to quantify the presented method's occlusion processing capability on three databases in multiple configurations, such as arbitrary missing joints and frames.

Qiankun Liu et al. [84], 2022, The suggested occlusion estimating module aims to forecast where occlusions occur, which are then utilized to approximate the locations of objects that are missing. This module tracks multi-object using unsupervised learning and it is observed from the results that it is better than the supervised learning method. The occlusion prediction module can help with the missing tracking problem.

Renshu Gu et al. [85], 2021, use a gated convolution temporal network to handle the occlusion in 3D pose estimation images. The authors also designed their dataset known as MMHuman to handle the occluded images in real-world scenarios.

Yan Di et al. [80], 2021, In this paper authors have overcome the challenge of directly regressing the 6 Degree of freedom object pose an RGB image. This can be done by developing a two-layer model for 3D objects that significantly improves end-to-end pose estimation accuracy. The developed SO-Pose, 6D pose regression framework

(a) Detection using 1 classifier (b) Detection using 17 classifiers

Fig.15 Inter occlusion detected using different classifiers [83]

outperforms other single-layer adversaries on a variety of tough datasets.

Zhou et al. [86], 2020, implemented a Siamese network for the understanding of occlusion among the different images. This method has the concept of erasing and reconstruction i.e., at first erasing step will erase the occluded part which may sometimes lead to the deletion of the important features in images therefore, the reconstruction technique is applied to recover those useful features. Thus, this method has considered all aspects and implemented a robust model. Moreover, it has better performance as it has been trained on a variety of occluded images from different datasets (MPII, LSP, COCO).

Rohit Kumar Jena [87], 2019 In this research, we present a new dataset to better assess algorithms and an innovative and effective way to tackle the issues of crowd pose estimation. They also defined a new dataset based on real-world applications. The sole major drawback of this work is that it still does not fully exploit the Part Affinity fields' potential.

Cheng et al. [88], 2019, to handle occlusion in 3D images a deep learning-based framework is designed in which partial 2D keypoints are there when occlusion occurs, which are fed into 2D and 3D CNN. The keypoints selected here are incomplete as occlusion keypoints errors are less in it. This network offers a "Cylinder Man Model" to estimate the utilization of bodily parts in 3D space since no such database exists.

Shihui Zhang [81] et al. 2016, This paper proposes a novel method for estimating movements in 3D space using depth information of a visual image. To monitor the efficiency of dynamic self-occlusion obfuscation, an assessment condition called "effective avoidance rate" is introduced. The experimental findings reveal that the suggested method meets the purpose of the camera intelligently avoiding self-occlusion when an object moves (Fig. 15).

The paper in the literature shows that despite wide research in occlusion handling there is still a need for future research in this field to enhance the reliability and robustness of the 3D pose estimation. There is still much work required in handling the 3D pose estimation dataset with occlusion. Therefore, future researchers should develop an efficient and simple DL algorithm that handles occlusion easily.

Handling Occlusion- Image segmentation and tracking of objects are the two approaches that can be used to deal with occlusions.

6.2 Crowd detection

The problem of crowded scenes in human pose estimation occurs only while detecting poses in multiperson pose estimation. It is a challenging task because overlaps and occlusions make it much harder to recognize human enclosing boxes and derive pose indications from particular key points, MPPE in overcrowded settings is problematic [89]. The study below shows how to handle the occlusions in crowded scenarios using various approaches. Not only in HPE, crowd monitoring is necessary but it also plays a major in many other applications such as population monitoring, public event management, suspicious activity detection, military management, disaster management, and many more[90].

This work presents a bounding box detection and keypoint grouping-free direct pose-level inference technique known as PINet [89]. It immediately infers a person's whole pose cues from his or her observable body components, rather than suggesting individual key points. This method is more efficient than other proposed methods but must be improved to achieve a higher level of accuracy.

Shuning Chang et al. [91], 2020, In this work, the focus is on enhancing HPE in cluttered scene films. The work is started by detecting people and doing SPPE using a top-down approach is applied for better results.

The results are then processed after applying an algorithm using optical flow which increases the robustness of the model.

Cheng Chi et al. [92], 2020, The suggested technique, PedHunter, adds significant occlusion managing capabilities

to current region-based detection networks without requiring additional inference procedures. We create a mask-guided component that uses head knowledge to optimize the network's function learning algorithm. The algorithm uses robustness and tests its results on heterogeneous datasets thus, achieves high accuracy.

Jiefeng Li et al. [93], 2019, In this research, we present a new dataset for better-evaluating algorithms as well as an innovative and fast way of tackling the problem of crowd pose estimation.

A.S. Elons et al.[94], 2017, The research uses a combination of LSTM and CNN for managing crowded occlusion thus, using a hybrid deep learning model. The model uses real-time video frames to analyze and implement a model with better accuracy.

Extensive studies in the field of crowd analysis have been conducted recently. There have been numerous datasets released. But most of these datasets are not able to resolve the localization and behavior analysis issues of crowd detection. As a result, there seems to be a lot of room for an in-crowd assessment of the available database.

7 Datasets and evaluation metrics

7.1 Dataset

In this section different datasets for 2D/3D pose estimation modeling are elaborated. The content in Tables 3 and 4 shows the full description of the 2D and 3D datasets respectively.

7.2 Performance metrics

Metrics act as an indicator for the pose estimation performance [68]. It is difficult to assess the performance of an estimation of human poses because many factors must be taken into account [36].

7.2.1 Percentage of correctly estimated body parts (PCP)

PCP evaluates the stick predictions [28]. For a specific part, PCP can be evaluated as follows,

$$\text{PCP} = \frac{\text{No. of correct parts for the entire dataset}}{\text{No. of total parts for the entire dataset}} \tag{1}$$

The higher the value of PCP, the better it works.

7.2.2 Percentage of correct key points (PCK)

PCK can be applied in the computation of both 2D and 3D (PCK3D), in PCK an appendage is viewed as identified (a right part) if the separation between the two anticipated joint areas and the actual appendage joint areas is not exactly 50% of the appendage length (ordinarily meant as PCP at 0.5). The variation of PCK is PCKh (PCK head) [112]. The limit of PCKh at 0.5 is half of the head bone band.

$$\text{PCK at } 0.2 = \text{ Distance } < 0.2 * \text{ Torso width among expected and true joint} \tag{2}$$

7.2.3 Percentage of detected joints (PDJ)

If the difference between the expected and the actual articulation is within a certain part of the torso diameter, the revealed joint is considered correct [28]. The cons of PDJ are that it alleviates the shorter limb as the shorter limb has smaller torsos. PDJ at 0.2 implies that the interval among expected and actual joints is less than 0.2 * torso diameter. It is typically used for 2D Pose Estimation. PDJ works better for the higher values.

7.2.4 Mean per joint position error (MPJPE)

Mean per joint position error is the average error for all N joints per joint position (typically, $N = 16$). The root joints (typically the pelvis) of the approximate and groundwater-3D poses are determined after alignment by the method of similarity transformation. For better understanding, we can write it as follows,

Per joint position error = calculated Euclidean distance of ground truth and expected value for a joint.

Average error per joint position = Average error for all N joints per joint position (typically N = 16).

In MPJPE, the joints 'J' is also normalized concerning the root joint $\hat{\mathrm{J}}$. It can be formalized as follows:

$$\text{MPJPE} = \frac{1}{M}\frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N}\left\|\left(\mathrm{J}_n^{(m)} - \mathrm{J}_{\text{root}}^{(m)}\right) - \left(\hat{\mathrm{J}}_n^{(m)} - \hat{\mathrm{J}}_{\text{root}}^{(m)}\right)\right\|_2 \tag{3}$$

where, M, N is no. of samples and no. of joints respectively.

MPJPE is mainly used for 3D pose estimation and it gives better results with lower values.

Table 3 Description of 2D Datasets used for HPE

Year	Dataset name	Dataset distribution	Type of DATASET	Type of physique	No. of joints	Description	Applications
2008	Buffy [95]	472 frames training 276 frames testing	Video Based	Upper body	6 body parts	Only one person is annotated in each image	Unconstrained movie and TV videos
2014	MPII Human Pose Dataset [96]	Training- 3844 Testing- 1758	Image Based	Full Body	16 Keypoints	Several interactive people in strongly articulated positions with a varying number of parts are grouped	Identify different activities such as "Home Activities", Garden" Washing windows, "Picking fruit" or "Rock climbing"
2010	Leeds Sports Pose [97]	Training- 1000 Testing- 1000	Image Based	Full body	14 Keypoints	Resize images. Only one person is annotated in each image	Used to identify the different types of sports such as "volleyball", "baseball", "gymnastics", "parkour" and many others
2013	Frames Labeled in Cinema(FLIC) [98]	Training- 3987 Testing- 1016	Image Based	Upper body	10 Keypoints	Hollywood movies are used to collect. The people who are obstructed or significantly non-frontal are removed	It is used for Human body joint detection
2014	Frames Labeled in Cinema (FLIC) Plus [99]	17 k images	Image Based	Upper body	10 Keypoints	FLIC-plus is a cleaned and simple version with no difficult poses	Enhanced version
2013	BBC Pose [100]	Training- 2000 Validation- 1000 Testing- 1000	Video Based	Upper body	4 body parts	BBC Pose is a collection of 20 videos obtained from the BBC with a sign language translator superimposed	Joint Prediction
2017	MSCOCO [101]	Training- 118,287 Validation- 5000 Testing-40,670	Image Based	Full body	17 Keypoints	The data was gathered from the internet. It has a wide range of actions	Object detection, segmentation, and captioning
2017	COCO 17 [102]	Training-64 k Validation-2.7 k Testing-40 k	Image Based	Full body	17 Keypoints	Various annotations from Google, Flickr are taken	Category detection, instance spotting and instance segmentation
2016	COCO 16	Training-45 k Validation-22 k Testing-80 k	Image Based	Full body	17 Keypoints		
2017	PoseTrack [35]	Training- 300 Validation- 50 Testing- 208	Video Based	Full body	15 Keypoints	This dataset focuses on 3 aspects: single-frame MPPE, MPPE in videos, and articulated tracking	Articulated multiperson pose tracking, annotated 15 body parts for each body pose
2014	Parse	Training-100 Testing- 205	Image Based	Full body	14 Keypoints	It's a small dataset with several annotations, like facial features, gender, gaze direction	Focuses on direction of gaze and facial features

Table 4 Description of 3D datasets used for HPE

Year	Name of dataset	Number of images	Type of dataset	No. of joints	Description	Applications
2018	3D Poses in the Wild (3DPW) [103]	60 video sequences ≈ 51 k frames	Video Based	18 Keypoints	3DPW is the one to integrate ten video clips from animations. For training and testing, the entire dataset can be exploited	Used to detect future human trajectories and skeleton poses
2014	Human 3.6 M [104]	Training≈1.5 M Validation≈0.6 M Testing ≈ 1.5 M	Video Based	17 Keypoints	3D human representations are created through visual effects and put into complicated real-world surroundings, which are observed with movable cameras and in an indoor space with occlusion	Human activities like talking on phone, capturing images, posing etc. It also includes some more synchronized images
2017	MPI-INF-3DHP [105]	≈1.3 M	Image Based	15 Keypoints	This dataset requires 25 GB and 7 GB spaces for training and testing respectively. Includes indoor and outdoor scenes	Deep kinematic pose. Good for image augmentation and segmentation
2018	JTA (Joint Track Auto) [106]	Training—256 Testing—256 ≈5 00 k frames	Video Based	–	Occlusion Annotation, approx. 10 M body poses	Pedestrian pose estimation and tracking in urban scenarios
2017	Total Capture [107]	1,892,176	Image Based (render from videos)	26 Keypoints	5 subjects, 5 actions, IMU and Vicon data, indoor environment	Yoga, walking, giving directions, bending over, freestyle and crawling
2010	HumanEva I [108]	Training ≈ 6.8 k Validation≈ 6.8 k Testing ≈ 24 k	Video Based	15 Keypoints	3 color + 4 grayscale video cameras, indoor environment, used for training, validation, and testing	Jogging, walking, standing, gesturing etc
2010	HumanEva II [109]	Testing-2.5 k	Video Based	15 Keypoints	3color video cameras, mainly used for testing purposes	
2016	TNT15 [110]	≈ 13 K frames	Video Based	15 Keypoints	4 subjects, 5 actions, IMU data, indoor environment, 3D body scans	
2015	Panoptic [111]	1.5 M	Video Based	15 Keypoints	Multi-annotation), internal environment	Activity recognition

Table 5 Evaluation of 2D HPE methods on LSP and MPII dataset

Author	Method	LSP	MPII
PCP			
Carreira et al. [77]	IEF	72.5%	–
Chen et al. [113]	Single CNN	75.0%	–
Chu et al. [114]	Single CNN	81.1%	–
Lifshitz et al. [115]	RNN	**84.2%**	–
PCK			
Bulat et al. [116]	Multi-stage CNN	82.7%	83.5%
DeepCut. [75]	Single CNN	82.4%	87.1%
Lifshitz et al. [115]	RNN	85.0%	85.0%
Belagiannis et al. [68]	RNN	88.1%	85.2%
Chu et al. [114]	Multi-stage CNN	91.5%	92.6%
Chen et al. [113]	GAN	**92.1%**	93.1%
Chou et al. [117]	GAN	91.8%	**94.0%**
AUC			
DeepCut. [75]	Single CNN	63.5	56.5
Tompson et al. [99]	CNN	47.3	51.8
Carreira et al. [77]	IEF		49.1
Wei et al. [29]	CNN	**65.4**	**61.4**

Higher PCP value on LSP dataset using RNN is 84.2%, Higher PCK value on LSP dataset using GAN method is 92.1% and on MPII dataset is 94.0%, and AUC value on LSP dataset using CNN is 65.4 and on MPII dataset is 61.4 (in bold)

7.2.5 Object keypoint similarity (OKS)

The difference between expected points and ground truth points is measured as averaged by an individual's scale. The OKS metric is more difficult to compute than the PDJ one. The OKS can be calculated as,

$$\mathrm{OKS} = \exp\left(\frac{-d_i^2}{2s^2k_i^2}\right) \tag{4}$$

where, di: toground truth; s: scale the area of the bounding box divided by the total image area; k: per-key point constant that controls fall off; OKS = 1 means that the prediction done is perfect or you can say that perfect prediction.

Table 7 Performance analysis on Human 3.0M dataset

References	Walking			
	S1	S2	S3	Avg
Yasin et al. [121]	35.8	32.4	41.6	36.6
Kostrikov and Gall [62]	44.0	30.9	41.7	38.9
Wang et al. [119]	71.9	75.7	85.3	77.6
Simo Serra et al. [122]	**99.6**	**108.3**	**127.4**	**111.8**
Bo and Sminchisescu [123]	38.2	32.8	40.2	37.1

Higher S1, S2, S3 and Average value on Human 3.0 M dataset for Simo Serra et al. [122] (in bold)

7.2.6 Average precision (AP)

Average Precision (AP) metrics are used for the evaluation of per-frame multi-person pose estimation. This measures the average recall value of 0 to 1. The other variation of AP is the Mean Average Precision (MAP) [112].

Here, the below section gives the performance analysis of different evaluation metrics on the various datasets. The Tables 5, 6, 7, 8, 9 shows the different results based on the method used.

8 Security attacks

DL and ML models are subject to adversarial instances, malevolent inputs that have been altered to produce incorrect predictive modeling while being unaltered to human vision [129].

There are various categories of security attacks. One of them is an attack based on specificity. There are two types of attacks based on this namely, targeted and untargeted attacks.

Table 6 Performance analysis on Human 3.6M and PoseTrack dataset

MPJPE (Average value of all the parameters- eat, greet, phon, pos, pur, sit, wait, walk etc.)			
		Human 3.6 M	PoseTrack
Chen et al. [113]	GAN	**41.6**	–
Cai et al. [118]	CNN	39.0	–
Wang et al. [119]		32.7	–
Average precision (AP)			
Guanghan Ning et al. [120]	Ground truth detections	–	**81.7%**
	Deform FPN (ResNet101)	–	74.6%
	Deform R-FCN (ResNet101)	–	73.7%

Higher MPJPE value (41.6) on Human 3.6 M dataset and Higher average precision of 81.7% on PoseTrack dataset (in bold)

Table 8 Evaluation of algorithms for identifying 3D human posture using various inputs on the MPI-INF-3DHP dataset

References	Method	PCK	MPJPE	AUC
Zhou et al. [124]	Supervised	69.2	–	32.5
Mehta et al. [105]	Improved CNN	76.5	–	40.8
Habibie et al. [125]	New DL based method	69.6	**127.0**	35.5
Xu et al. [126]	DenseRaC	**89.0**	83.5	49.1
Wandit and Rosenhahn et al. [127]	RepNet	82.5	97.8	**58.5**

Higher PCK, MPJPE and AUC value using DenseRac [126], New DL based method [125] and RepNet [127] method respectively (in bold)

Table 9 Outcomes from the dataset for the MSCOCO keypoint challenge (AP)

References	Method	AP
Cao et al. [128]	CMU-Pose	61.8
Papandreou et al. [37]	G-RMI	68.5
He et al. [60]	Mask R-CNN	63.1
Chen et al. [38]	Megvii	72.1
Fang et al. [45]	Faster RCNN with softnms, Pyranet	**72.3**

Higher AP value using Faster RCNN with softnms, Pyranet [45] (in bold)

In a targeted attack [130], there is one targeted class that has to be misclassified by the attackers on the other hand there is no specific targeted class to misclassify their main concern is to misclassify the model with the help of adversarial example [131].

On the basis of the study, it has been analyzed that targeted attacks have high time complexity but are more accurate as compared to untargeted attacks [130, 131].

The other category of adversarial security attack is a black box and white box attack. In a black Box attack, an attacker has no prior knowledge of the model in which it is attacking [132] but in its white box attacks, the attacker has prior knowledge of the model, and model attributes are easily accessible to them [133].

Apart from these are other common attacks are,

Integrity attack—A efficient resource assault that is identified as normal traffic due to false negatives [134].

Availability attack—A broad class of an attack that makes the system unusable with classification errors, denial of service, false negatives and positives, etc. [135].

Privacy violation attack—The discrepancy in confidential information if not handled properly results in a privacy violation attack [136].

Hou et al. [137], 2022, the authors have developed a deep learning-based anomaly detection model known as the integrity protection method (IPDLS). The main objective of the method is to determine the feature similarity between the skeptical sample and the original sample. The algorithm is evaluated on MNIST and CIFAR10 datasets and achieved better performance.

Jiawang Bai et al. [130], 2021, To achieve stealth capabilities, the goal is to misidentify a specific instance into a target class without changing it, but without lowering the predictive performance of other samples much. The problem is framed as binary integer programming since the parameters are stored as bits (i.e., 0 and 1). The defined technique is resistant to different parameters and is superior when it comes to attacking DNNs.

Xinghao Yang et al. [135], 2020, here the targeted attention attack is designed on one of the real-world computer vision applications (road sign recognition attack). The effective universal assault that optimizes a single perturbation based on a collection of training is created with the help of pre-trained images.

Guowen Xu et al. [138], 2019, the possible risks posed by deep learning are investigated in this study and the most up-to-date solutions are determined. Moreover, the authors have developed the SecureNet which can withstand a variety of security and privacy risks during a prognosis phase.

Arjun Nitin Bhagoji et al.[134], 2018, The author has designed the gradient-based black-box attacks without transferability. In addition to this, methods for decoupling the number of questions needed to produce each adversarial instance from the input's dimensionality are also introduced. The approach is tested on MNIST and CIFAR 10 datasets and achieves a better accuracy,

Yi Shi and Yalin E. Sagduyu [136], 2017, The method shows that the evasive and causal attacks are initiated after the exploratory attack, considerably raising the inaccuracy and also introducing defense techniques. A defense approach is described that involves changing a limited number of labels in the actual classifier to avoid the adversary from reliably inferring its identity and using it in evasion and causal assaults. The method is reliable and takes the varieties of a dataset into consideration (text and image), thus, able to find out the vulnerabilities in both types of datasets.

Moustapha Cisse et al. [129], 2017, a versatile method Houdini is introduced for creating adversarial instances that are specially customized for the task's final evaluation, whether combinatorial or non-decomposable. This approach achieves a significantly higher rate of success in many different areas such as pose estimation, semantic segmentation, and speech recognition. As a result, the usage of adversarial instances is extended beyond picture categorization.

9 Conclusion

In this survey, it is observed that as compared to other computer vision problems, the location of parts of the human body from images and their assembly based on a predefined human body configuration are involved in estimating the human pose. By reviewing so many articles it has been observed that humans and commodities are often considered isolated in classic human-object interaction evaluation techniques. The new HPE methodology relies on the interplay between people and things.

Depending on this, a comprehensive review of various classical and deep learning 2D/3D single pose and multi-pose techniques for HPE is studied and it is observed that each algorithm changes according to its environment and the availability of the dataset. These techniques and strategies have their benefits and shorts. Furthermore, the various human pose datasets for a single person and multi-person 2D/3D have been summed up along with some of the evaluation metrics. At last, the use of these pose estimation techniques, and responses to circumstances of body reality are discussed. Finally, it can be concluded that despite the extraordinary advancement of HPE with deep learning, there still stay some uncertain difficulties and holes among research and viable applications, for example, crowd and occlusion. The most essential concerns for deep learning-based methodologies are efficient organizations and sufficient training of data. The different security attacks in deep learning models are also discussed so that the researchers should be aware of these attacks and their abnormal behavior.

9.1 Current and future developments

From the study, it can be depicted that the human posse estimation trend is lying towards the deep learning approaches as it can be noticed that these deep learning techniques achieve better performance over the activities and the dataset as compared to other state-of-art approaches. The success of deep learning approaches to the HPE technique is the availability of a huge amount of the dataset which is one of the limitations of the DL application. Despite various databases having been created for unbiased HPE assessments, additional datasets with adequate examination methodologies are still desired. Additional body sensors can be used in the long term to record raw data from diverse postures. The domain can be separated into two categories: 2D and 3D pose estimate. While 2D PE has attained an adequate degree of precision, 3D PE takes a lot of effort unless more balancing models are developed, particularly for interpretation from a single image and without depth details.

Moreover, the future in HPE is massive and has a great application area that is important in our daily life. There is also scope for achieving good results on the higher dimensional datasets (higher than 2D/3D) such as on 6D pose estimation which estimates the position and direction of the 6D poses. These poses are useful in robotic applications. Although massive work endeavor efforts are devoted to identifying human poses from videos or photos, there is indeed a significant gap between theoretical study and real-world applications.

Funding Not applicable.

Declarations

Conflict of interest No conflict of interest, financial or otherwise.

Consent for publication Not applicable.

References

1. Chen, Y., Tian, Y., He, M.: Monocular human pose estimation: a survey of deep learning-based methods. Comput. Vis. Image Underst. (2020). https://doi.org/10.1016/j.cviu.2019.102897
2. Szczuko, P.: Deep neural networks for human pose estimation from a very low resolution depth image. Multimed. Tools Appl. **78**, 1–21 (2019). https://doi.org/10.1007/s11042-019-7433-7
3. Liu, Y., Xu, Y., Li, S.: 2-D Human Pose Estimation from Images Based on Deep Learning: A Review," 2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Xi'an, China. 462-465. (2018). https://doi.org/10.1109/IMCEC.2018.8469573
4. Chen, C., Wang, T., Li, D., Hong, J.: Repetitive assembly action recognition based on object detection and pose estimation. J. Manuf. Syst. **55**, 325–333 (2020). https://doi.org/10.1016/j.jmsy.2020.04.018
5. Silva, D., Varges, M., Marana, A.N.: "Human action recognition in videos based on spatiotemporal features and bag-of-poses. Appl. Soft Comput. **95**, 106513 (2020). https://doi.org/10.1016/j.asoc.2020.106513
6. Ordóñez, F., Roggen, D.: Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. Sensors **16**, 1–25 (2016). https://doi.org/10.3390/s16010115
7. Christian, S., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision." In Proceedings of the IEEE conference on computer vision and pattern recognition. 2818–2826, (2016). https://doi.org/10.1109/CVPR.2016.308
8. Chen, K., Paolo Gabriel, Alasfour, A., Gong, C., Doyle, W.K., Devinsky, O., Friedman, D., et al.: Patient-specific pose estimation in clinical environments. IEEE J. Transl. Eng. Health Med. **6**, 1–11 (2018). https://doi.org/10.1109/JTEHM.2018.2875464
9. Islam, M.J., Mo J., Sattar. J.: Robot-to-robot relative pose estimation using humans as markers. arXiv preprint arXiv:1903.00820 (2019).

10. Zimmermann, C., Tim, W., Christian, D., Wolfram, B., and Thomas, B.: 3d human pose estimation in Rgbd images for robotic task learning. In 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 1986–1992. IEEE (2018). https://doi.org/10.1109/ICRA.2018.8462833
11. Vasileiadis, M., Sotiris, M., Dimitrios, G., Christos-Savvas, B., Dimitrios, T.: "Robust human pose tracking for realistic service robot applications." In Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 1363–1372. (2017). https://doi.org/10.1109/ICCVW.2017.162
12. Preim, B., Meuschke, M.: A survey of medical animations. Comput. Graph. **90**, 145–168 (2020). https://doi.org/10.1016/j.cag.2020.06.003
13. Kumarapu, L., Mukherjee, P.: "AnimePose: Multi-person 3D pose estimation and animation", arXiv:2002.02792v1, pp 1–5, (2020). https://doi.org/10.1016/j.patrec.2021.03.028
14. Tiwari, M.M., Tiwari, M.T., Rajendran, G., Suson, R.: Deep learning approach for generating 2D pose estimation from video for motion capture animation. Int. J. Future Gener. Commun. Netw. **13**(2), 1556–1561 (2020)
15. Casado García, F., Luis, Y.,Pérez Losada, D., Santana Alonso, A.: "Pose estimation and object tracking using 2D images", In 2017-27th International Conference on Flexible Automation and Intelligent Manufacturing, Modena, Italy, (2017). https://doi.org/10.1016/j.promfg.2017.07.134
16. Cleetus, A.: Real-time multiple human pose estimation for animations in game engines. Int. Res. J. Eng. Technol. (IRJET) **7**(5), 7923–7928 (2020)
17. https://mobidev.biz/blog/human-pose-estimation-ai-personal-fitness-coach. Accessed 2 Sept 2021
18. https://viso.ai/deep-learning/pose-estimation-ultimate-overview/. Accessed 14 Aug 2021
19. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. Int. J. Comput. Vision **61**, 55–79 (2005). https://doi.org/10.1023/B:VISI.0000042934.15159.49
20. Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE Trans. Comput. **C22**, 67–92 (1973). https://doi.org/10.1109/T-C.1973.223602
21. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2878–2890 (2013). https://doi.org/10.1109/TPAMI.2012.261
22. Wu, J., Christopher, G., James M.R.: "Real-time human detection using contour cues." In 2011 IEEE international conference on robotics and automation, pp. 860–867. IEEE, (2011). https://doi.org/10.1109/ICRA.2011.5980437
23. Micilotta, AS., Eng-Jon, O., Richard, B.: "Real-time upper body detection and 3D pose estimation in monoscopic images." In European Conference on Computer Vision, pp. 139–150. Springer, Berlin, Heidelberg, (2006). https://doi.org/10.1007/11744078_11
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2016). https://doi.org/10.1109/TPAMI.2016.2577031
25. https://neuralet.com/article/human-pose-estimation-with-deep-learning-part-i/. Accessed 15 Sept 2021
26. Munea, T.L., Jembre, Y.Z., Weldegebriel, H.T., Chen, L., Huang, C., Yang, C.: The progress of human pose estimation: a survey and taxonomy of models applied in 2D human pose estimation. IEEE Access **8**, 133330–133348 (2020). https://doi.org/10.1109/ACCESS.2020.3010248
27. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net: localization-classification-regression for human pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3433–3441. (2017)
28. Toshev, A., Szegedy, DC.: Human pose estimation via deep neural networks, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, WI, USA, pp. 1653–1660. (2014).
29. Wei, S., Ramakrishna, V., Kanade, T., Sheikh, Y.: "Convolutional Pose Machines," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 4724–4732. (2016). https://doi.org/10.1109/CVPR.2016.511
30. Onishi, K., Takiguchi, T., Ariki, Y.: "3D human posture estimation using the HOG features from monocular image." In 2008 19th International Conference on Pattern Recognition, pp. 1–4. IEEE. (2008). DOI:https://doi.org/10.1109/ICPR.2008.4761608
31. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation, European Conference on Computer Vision (ECCV) (2016) 483–499 (2016).
32. Bulat, A., Georgios, T.: "Human pose estimation via convolutional part heatmap regression." In European Conference on Computer Vision, pp. 717–732. Springer, Cham, (2016)
33. Luo, Z., Zhicheng, W., Yan, H., Liang, W., Tieniu, T., Erjin, Z.. "Rethinking the Heatmap Regression for Bottom-up Human Pose Estimation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13264–13273. (2021).
34. Chen, X., Yuille, A. L.: Articulated pose estimation by a graphical model with image dependent pairwise relations, in Advances in Neural Information Processing Systems, pp. 1736–1744. (2014).
35. Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., Schiele, B.: "PoseTrack: A Benchmark for Human Pose Estimation and Tracking", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5167–5176. (2018). https://doi.org/10.1109/CVPR.2018.00542
36. Dang, Qi., Yin, J., Wang, B., Zheng, W.: Deep learning based 2D human pose estimation: a survey. Tsinghua Sci. Technol. **24**, 663–676 (2019). https://doi.org/10.26599/TST.2018.9010100
37. Papandreou, G., Tyler, Z., Nori, K., Alexander, T., Jonathan, T., Chris, B., Kevin M.: Towards accurate multi-person pose estimation in the wild." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4903–4911. (2017). https://doi.org/10.1109/CVPR.2017.395
38. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.:"Cascaded pyramid network for multi-person pose estimation." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7103–7112. (2018). https://doi.org/10.1109/CVPR.2018.00742
39. Gamra, M.B., Akhloufi, M.A.: A review of deep learning techniques for 2D and 3D human pose estimation. Image Vis. Comput. (2021). https://doi.org/10.1016/j.imavis.104282
40. Rodrigues, N., Torres, H.D.R., Oliveira, B., Borges, J., Queirós, S.F.M., Mendes, J.A., Fonseca, J.C., Coelho, V., Brito, J.H.: Top-down human pose estimation with depth images and domain adaptation. SCITEPRESS (2019)
41. Kocabas, M., Karagoz, S., Akbas, E.: "Multiposenet: Fast multi-person pose estimation using pose residual network." In Proceedings of the European conference on computer vision (ECCV), pp. 417–433. (2018). https://doi.org/10.1007/978-3-030-01252-6_26
42. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: "Deepercut: A deeper, stronger, and faster multi-person pose estimation model." In European Conference on Computer Vision, pp. 34–50. Springer, Cham, (2016). https://doi.org/10.1007/978-3-319-46466-4_3

43. Zheng, C., Wu, W., Yang, T., Zhu, S., Chen, C., Liu, R., Shen, J., Kehtarnavaz, N., Shah, M.: "Deep learning-based human pose estimation: A survey." arXiv preprint arXiv:2012.13392 (2020).
44. Cao, Z., Simon, T., Wei, S. E., Sheikh, Y.: "OpenPose: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. (2017). https://doi.org/10.1109/CVPR.2017.143
45. Fang, H. S., Xie, S., Tai, Y. W., Lu, C.: "Rmpe: Regional multi-person pose estimation." In Proceedings of the IEEE international conference on computer vision, pp. 2334–2343. (2017). https://doi.org/10.1109/ICCV.2017.256
46. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: "Efficient object localization using convolutional networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 648–656. (2015). https://doi.org/10.1109/CVPR.2015.7298664
47. Sun, K., Xiao, B., Liu, D., Wang, J.,: "Deep high-resolution representation learning for human pose estimation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5693–5703. (2019). https://doi.org/10.1109/CVPR.2019.00584
48. Osokin, D.: "Real-time 2d multi-person pose estimation on CPU: Lightweight OpenPose." arXiv preprint arXiv:1811.12004 (2018).
49. Tang, W., Yu, P., Wu, Y.: "Deeply learned compositional models for human pose estimation." In Proceedings of the European conference on computer vision (ECCV), pp. 190–206. (2018). https://doi.org/10.1007/978-3-030-01219-9_12
50. Mao, W., Ge, Y., Shen, C., Tian, Z., Wang, X.: "Tfpose: Direct human pose estimation with transformers." arXiv preprint arXiv:2103.15320 (2021).
51. Jain, A., Tompson, J., LeCun, Y., Bregler, C.: "Modeep: A deep learning framework using motion features for human pose estimation." In: Asian conference on computer vision, pp. 302–315. Springer, Cham. (2014). https://doi.org/10.1007/978-3-319-16808-1_21
52. Alzughaibi, A., Chaczko, Z.: "Human detection model using feature extraction method in video frames," 2016 International Conference on Image and Vision Computing New Zealand (IVCNZ), pp. 1-6. (2016) https://doi.org/10.1109/IVCNZ.2016.7804424
53. https://mobidev.biz/wp-content/uploads/2020/07/3d-keypoints-human-pose-estimation.png. Accessed 20 Aug 2021
54. Hanguen, K., Lee, S., Lee, D., Choi, S., Ju, J., Myung, H.: Real-time human pose estimation and gesture recognition from depth images using superpixels and SVM classifier. Sensors (Basel) (2015). https://doi.org/10.3390/s150612410
55. Chen, K., Gong, S., Xiang, T.: "Human pose estimation using structural support vector machines", 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, pp. 846–851. (2011). https://doi.org/10.1109/ICCVW.2011.6130340
56. Hallquist, A., Zakhor, A.: "Single view pose estimation of mobile devices in urban environments." In 2013 IEEE Workshop on Applications of Computer Vision (WACV), pp. 347–354. IEEE, (2013).
57. Fei, X., Wang, H., Cheong, L. L., Zeng, X., Wang, M., Tighe, J.: "Single View Physical Distance Estimation using Human Pose." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12406–12416. (2021)
58. Wang, J., Tan, S., Zhen, X., Xu, S., Zheng, F., He, Z., Shao, L.: Deep 3D human pose estimation: a review. Comput. Vis. Image Underst. (2021). https://doi.org/10.1016/j.cviu.2021.103225
59. https://www.kdnuggets.com/. Accessed 30 Aug 2021
60. He, K., Gkioxari, G., Dollár, P., Girshick, R.: "Mask r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 2961–2969. (2017). https://doi.org/10.1109/ICCV.2017.322
61. Su, J.-Y., Cheng, S.-C., Chang, C.-C., Chen, J.-M.: Model-based 3D pose estimation of a single rgb image using a deep viewpoint classification neural network. Appl. Sci. **9**(12), 2478 (2019). https://doi.org/10.3390/app9122478
62. Kostrikov, I., Gall, J.: Depth sweep regression forests for estimating 3D human pose from images. BMVC **1**(2), 5 (2014). https://doi.org/10.5244/C.28.80
63. Benzine, A., Chabot, F., Luvison, B., Pham, Q. C., Achard, C.: "Pandanet: Anchor-based single-shot multi-person 3d pose estimation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6856–6865. (2020).
64. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: "Single-shot multi-person 3d body pose estimation from monocular rgb input." arXiv preprint arXiv:1712.03453 (2017).
65. Li, S., Chan, A.B.: "3d human pose estimation from monocular images with deep convolutional neural network." In Asian Conference on Computer Vision, pp. 332–347. Springer, Cham. (2014). https://doi.org/10.1007/978-3-319-16808-1_23
66. Deng, Y., Sun, Y., Zhu, J.: "SVMA: A GAN-based model for Monocular 3D Human Pose Estimation." arXiv preprint arXiv:2106.05616 (2021).
67. Miura, T., Sako, S.: 3D human pose estimation model using location-maps for distorted and disconnected images by a wearable omnidirectional camera. IPSJ Trans. Comput. Vis. Appl. **12**(1), 1–17 (2020). https://doi.org/10.1186/s41074-020-00066-8
68. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S., "3D Pictorial Structures for Multiple Human Pose Estimation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH. pp. 1669–1676. (2014). https://doi.org/10.1109/CVPR.2014.216
69. Groos, D., Ramampiaro, H., Ihlen, E.A.F.: EfficientPose: scalable single-person pose estimation. Appl. Intell. **51**(4), 2518–2533 (2021). https://doi.org/10.1186/s41074-020-00066-8
70. Marin-Jimenez, M.J., Romero-Ramirez, F.J., Munoz-Salinas, R., Medina-Carnicer, R.: 3D human pose estimation from depth maps using a deep combination of poses. J. Vis. Commun. Image Represent. **55**, 627–639 (2018). https://doi.org/10.1016/j.jvcir.2018.07.010
71. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: "3d human pose estimation in video with temporal convolutions and semi-supervised training." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7753–7762. (2019). https://doi.org/10.1109/CVPR.2019.00794
72. Meng, Lu., Gao, H.: 3D human pose estimation based on a fully connected neural network with adversarial learning prior knowledge. Front. Phys. **9**, 3 (2021). https://doi.org/10.3389/fphy.2021.629288
73. https://inblog.in/Human-Pose-Estimation-Using-Alpha-Pose-XyPPEbNTAO. Accessed 10 Sept 2021
74. https://analyticsindiamag.com/guide-to-openpose-for-real-time-human-pose-estimation/. Accessed 10 Sept 2021
75. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P. V., Schiele, B.: "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. (2016). https://doi.org/10.1109/CVPR.2016.533
76. https://debuggercafe.com/real-time-pose-estimation-using-alpha pose-pytorch-and-deep-learning/. Accessed 5 Sept 2021
77. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: "Human Pose Estimation with Iterative Error Feedback," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, pp. 4733–4742. (2016). https://doi.org/10.1109/CVPR.2016.512

78. Ghafoor, M., Mahmood, A.: "Quantification of Occlusion Handling Capability of 3D Human Pose Estimation Framework." IEEE Transactions on Multimedia. (2022). DOI: https://doi.org/10.48550/arXiv.2203.04113
79. Wu, B., Ramakant N.: "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors." In Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, vol. 1, pp. 90–97. IEEE, (2005). https://doi.org/10.1109/ICCV.2005.74
80. Di, Y., Manhardt, F., Wang, G., Ji, X., Navab, N., Tombari, F.: "SO-Pose: Exploiting Self-Occlusion for Direct 6D Pose Estimation." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12396–12405. (2021). DOI: https://doi.org/10.1109/ICCV48922.2021.01217
81. Zhang, S., He, H., Zhang, Y., Li, X., Sang, Y.: Dynamic self-occlusion avoidance approach based on the depth image sequence of moving visual object. Math. Probl. Eng. (2016). https://doi.org/10.1155/2016/4783794
82. Jacques, J. C., Dihl, L. L., Jung, C. R., Musse, S. R.: "Self-occlusion and 3D pose estimation in still images." In 2013 IEEE International Conference on Image Processing, pp. 2539–2543. IEEE. (2013). DOI: https://doi.org/10.1109/ICIP.2013.6738523
83. Veld, R. M., Wijnhoven, R. G. J., Bondarev, Y.: "Detection and handling of occlusion in an object detection system." In Video Surveillance and Transportation Imaging Applications 2015, vol. 9407, pp. 184–195. SPIE. (2015). DOI: https://doi.org/10.1117/12.2077175
84. Liu, Q., Chen, D., Chu, Q., Yuan, L., Liu, B., Zhang, L., Yu, N.: Online multi-object tracking with unsupervised re-identification learning and occlusion estimation. Neurocomputing (2022). https://doi.org/10.1016/j.neucom.2022.01.008
85. Gu, R., Wang, G., Hwang, J. N.: "Exploring severe occlusion: multi-person 3d pose estimation with gated convolution." In 2020 25th International Conference on Pattern Recognition (ICPR), pp. 8243–8250. IEEE. (2021). DOI: https://doi.org/10.1109/ICPR48806.2021.9412107
86. Antol, S., Lawrence Zitnick, C., Parikh, D.: "Zero-shot learning via visual abstraction." In European conference on computer vision, pp. 401–416. Springer, Cham. 2014. https://doi.org/10.1007/978-3-319-10593-2_27
87. Jena, R.: "Out of the Box: A combined approach for handling occlusion in Human Pose Estimation." arXiv preprint arXiv:1904.11157 (2019).
88. Cheng, Y., Yang, B., Wang, B., Yan, W., Tan, R. T.: "Occlusion-aware networks for 3d human pose estimation in video." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 723–732. (2019). DOI: https://doi.org/10.1109/ICCV.2019.00081
89. Wang, D., Zhang, S., Hua, G.: "Robust Pose Estimation in Crowded Scenes with Direct Pose-Level Inference." Advances in Neural Information Processing Systems 34 (2021).
90. Khan, K., Albattah, W., Khan, R.U., Qamar, A.M., Nayab, D.: Advances and trends in real time visual crowd analysis. Sensors (2020). https://doi.org/10.3390/s20185073
91. Chang, S., Yuan, L., Nie, X., Huang, Z., Zhou, Y., Chen, Y., Yan, S.: "Towards accurate human pose estimation in videos of crowded scenes." In Proceedings of the 28th ACM International Conference on Multimedia, pp. 4630–4634. (2020). DOI: https://doi.org/10.1145/3394171.3416299
92. Chi, C., Zhang, S., Xing, J., Lei, Z., Li, S. Z., Zou, X.: "Pedhunter: Occlusion robust pedestrian detector in crowded scenes." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 10639-10646. (2020). DOI: https://doi.org/10.1609/AAAI.V34I07.6690
93. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H. S., Lu, C.: "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10863–10872. (2019). DOI: https://doi.org/10.1109/CVPR.2019.01112
94. Elons, A.S., Abol-Ela, M.: "Occlusion resolving inside public crowded scenes based on social deep learning model," 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS). pp. 218-223. (2017). DOI: https://doi.org/10.1109/INTELCIS.2017.8260050
95. Ferrari, V., Marin-Jimenez, M., Zisserman, A., "Progressive search space reduction for human pose estimation." In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE. (2008). https://doi.org/10.1109/CVPR.2008.4587468
96. Shafaei, A., James J.L.: "Real-Time Human Motion Capture with Multiple Depth Cameras", Proceedings of the 13th Conference on Computer and Robot Vision. (2016). https://doi.org/10.1109/CRV.2016.25
97. Johnson, S., Everingham, M.: "Learning Effective Human Pose Estimation from Inaccurate Annotation", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 1465–1472. (2011). https://doi.org/10.1109/CVPR.2011.5995318
98. Sapp, B., Taskar, B.: "MODEC: Multimodal Decomposable Models for Human Pose Estimation", In 2013 IEEE Conference on Computer Vision and Pattern Recognition, NW Washington DC, United States, pp. 3674–3681. (2013). https://doi.org/10.1109/CVPR.2013.471
99. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. Adv. Neural. Inf. Process. Syst. **27**, 1799–1807 (2014)
100. Charles, J., Pfister, T., Everingham, M., Zisserman, A.: Automatic and efficient human pose estimation for sign language videos. Int. J. Comput. Vision **110**(1), 70–90 (2014). https://doi.org/10.1007/s11263-013-0672-6
101. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, Microsoft coco: Common objects in context, in European Conference on Computer Vision, Zurich, Switzerland, 2014, pp. 740–755.
102. Bin, Y., Chen, Z. M., Wei, X. S., Chen, X., Gao, C., Sang, N.: "Structure-aware Human Pose Estimation with Graph Convolutional Networks", Vol. 106, pp.107410, Pattern Recognition. (2020). https://doi.org/10.1016/j.patcog.2020.107410
103. Von Marcard, T., Henschel, R., Black, M. J., Rosenhahn, B., Pons-Moll, G.: "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. (2018).
104. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, No. 7. (2014). https://doi.org/10.1109/TPAMI.2013.248
105. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: "Monocular 3d human pose estimation in the wild using improved cnn supervision." In 2017 international conference on 3D vision (3DV), pp. 506–516. IEEE. (2017). https://doi.org/10.1109/3DV.2017.00064
106. Fabbri, M., Lanzi, F., Calderara, S., Palazzi, A., Vezzani, R., Cucchiara, R.: "Learning to detect and track visible and occluded body joints in a virtual world." In Proceedings of the European conference on computer vision (ECCV), pp. 430–446. (2018). https://doi.org/10.1007/978-3-030-01225-0_27
107. Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J.P.: Total capture: 3D human pose estimation fusing video and inertial sensors. BMVC **2**(5), 1–13 (2017). https://doi.org/10.5244/C.31.14

108. Sigal, L., Balan, A., Black, M.J.: HumanEva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. Int. J. Comput. Vis. **87**, 1–2 (2010)
109. Sigal, L., Black, M. J.: HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion, Techniacl Report CS-06–08, Brown University. (2006).
110. Marcard, T.V., Pons-Moll, G., Rosenhahn, B.: "Multimodal motion capture dataset TNT15". Leibniz Univ. Hannover, Hanover, Germany, and Max Planck for Intelligent Systems, Tübingen, Germany. Tech. Rep. (2016). https://doi.org/10.13140/RG.2.1.4162.0248
111. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: "Panoptic studio: A massively multiview system for social motion capture." In Proceedings of the IEEE International Conference on Computer Vision, pp. 3334–3342. 2015. DOI: https://doi.org/10.1109/ICCV.2015.381
112. Haque, A., Peng, B., Luo, Z., Alahi, A., Yeung, S., Fei-Fei, L.: "Towards Viewpoint Invariant 3D Human Pose Estimation", arXiv:1603.07076v3 [cs.CV], pp.1–20. (2016). https://doi.org/10.1007/978-3-319-46448-0_10
113. Chen, Y., Shen, C., Wei, X. S., Liu, L., Yang, J.: "Adversarial posenet: A structure-aware convolutional network for human pose estimation." In Proceedings of the IEEE International Conference on Computer Vision, pp. 1212–1221. (2017). DOI: https://doi.org/10.48550/arXiv.1705.00389
114. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. CVPR (2017). https://doi.org/10.48550/arXiv.1702.07432
115. Lifshitz, I., Fetaya, E., Ullman, S.: Human pose estimation using deep consensus voting. ECCV (2016). https://doi.org/10.1007/978-3-319-46475-6_16
116. Bulat, A., Tzimiropoulos, G.: Human pose estimation via convolutional part heatmap regression. ECCV (2016). https://doi.org/10.1007/978-3-319-46478-7_44
117. Chou, C. J., Chien, J. T., Chen, H. T.: "Self adversarial training for human pose estimation." In 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 17–30. IEEE. (2018). DOI: https://doi.org/10.48550/arXiv.1707.02439
118. Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.-J., Yuan, J., et al.: Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. In: International conference on computer vision p. 227–2281. (2019). doi: https://doi.org/10.1109/ICCV.2019.00236 48.
119. Wang, J., Yan, S., Xiong, Y., Lin, D.: "Motion guided 3d pose estimation from videos." In European Conference on Computer Vision, pp. 764–780. Springer, Cham. (2020). DOI: https://doi.org/10.48550/arXiv.2004.13985
120. Ning, G., Liu, P., Fan, X., Zhang, C.: "A top-down approach to articulated human pose estimation and tracking." In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 0–0. (2018). DOI: https://doi.org/10.1007/978-3-030-11012-3_20
121. Yasin, H., Iqbal, U., Kruger, B., Weber, A., Gall, J.: "A dual-source approach for 3d pose estimation from a single image." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4948–4956. (2016). DOI: https://doi.org/10.1016/j.cviu.2018.03.007.
122. Simo-Serra, E., Quattoni, A., Torras, C., Moreno-Noguer, F.: "A joint model for 2d and 3d pose estimation from a single image." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3634–3641. (2013). DOI: https://doi.org/10.1109/CVPR.2013.466
123. Bo, L., Sminchisescu, C., Kanaujia, A., Metaxas, D.: "Fast algorithms for large scale conditional 3D prediction." In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE. (2008). DOI: https://doi.org/10.1109/CVPR.2008.4587578
124. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3d human pose estimation in the wild: a weakly-supervised approach. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 398–407. (2017). DOI: https://doi.org/10.48550/arXiv.1704.02447
125. Habibie, I., Xu, W., Mehta, D., Pons-Moll, G., Theobalt, C.: In the wild human pose estimation using explicit 2D features and intermediate 3D representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10905–10914. (2019). DOI: https://doi.org/10.48550/arXiv.1904.03289
126. Xu, Y., Zhu, S.-C., Tung, T.: DenseRaC: Joint 3D Pose and Shape Estimation by Dense Render-and-Compare. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7760–7770. (2019). DOI: https://doi.org/10.48550/arXiv.1910.00116
127. Wandt, B., Rosenhahn, B.: RepNet: Weakly supervised training of an adversarial reprojection network for 3D human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7782–7791. (2019). DOI: https://doi.org/10.48550/arXiv.1902.09868
128. Chen, X., Lin, K., Liu, W., Qian, C., Lin, L.: Weakly-supervised discovery of geometry-aware representation for 3D human pose estimation. In: Conference on computer vision and pattern recognition p. 10895–904. (2019).
129. Cisse, M. M., Adi, Y., Neverova, N., Keshet, J.: Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. In: Advances in neural information processing systems, vol. 30. (2017)
130. Bai, J., Wu, B., Zhang, Y., Li, Y., Li, Z., Xia, S. T.: "Targeted attack against deep neural networks via flipping limited weight bits." arXiv preprint arXiv:2102.10496. (2021).
131. Rathore, P., Basak, A., Nistala, S. H., Runkana, V.: "Untargeted, Targeted and Universal Adversarial Attacks and Defenses on Time Series." In 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE. (2020). DOI: https://doi.org/10.1109/IJCNN48605.2020.9207272
132. Guo, S., Zhao, J., Li, X., Duan, J., Mu, D., Jing, X.: A black-box attack method against machine-learning-based anomaly network flow detection models. Secur. Commun. Netw. (2021). https://doi.org/10.1155/2021/5578335
133. Wang, Y., Liu, J., Chang, X., Wang, J., Rodríguez, R. J.: "DI-AA: An Interpretable White-box Attack for Fooling Deep Neural Networks." arXiv preprint arXiv:2110.07305. (2021).
134. Bhagoji, A. N., He, W., Li, B., Song, D.: "Exploring the space of black-box attacks on deep neural networks." arXiv preprint arXiv:1712.09491. (2017).
135. Yang, X., Liu, W., Zhang, S., Liu, W., Tao, D.: Targeted attention attack on deep learning models in road sign recognition. IEEE Internet Things J. **8**(6), 4980–4990 (2021). https://doi.org/10.1109/JIOT.2020.3034899
136. Shi, Y., Sagduyu, Y. E.: "Evasion and causative attacks with adversarial deep learning," MILCOM 2017 - 2017 IEEE Military Communications Conference (MILCOM). pp. 243–248. (2017). doi: https://doi.org/10.1109/MILCOM.2017.8170807.

137. Hou, R., Ai, S., Chen, Q., Yan, H., Huang, T., Chen, K.: Similarity-based integrity protection for deep learning systems. Inf. Sci. (2022). https://doi.org/10.1016/j.ins.2022.04.003
138. Xu, G., Li, H., Ren, H., Yang, K., Deng, R.H.: Data security issues in deep learning: attacks, countermeasures, and opportunities. IEEE Commun. Mag. **57**(11), 116–122 (2019). https://doi.org/10.1109/MCOM.001.1900091