Multimedia Systems
https://doi.org/10.1007/s00530-022-01019-0

SPECIAL ISSUE PAPER

2D Human pose estimation: a survey

Haoming Chen[1] · Runyang Feng[1] · Sifan Wu[1] · Hao Xu[2] · Fengcheng Zhou[1] · Zhenguang Liu[1]

Received: 6 September 2021 / Accepted: 3 March 2022

Abstract
Human pose estimation aims at localizing human anatomical keypoints or body parts in the input data (*e.g.*, images, videos, or signals). It forms a crucial component in enabling machines to have an insightful understanding of the behaviors of humans, and has become a salient problem in computer vision and related fields. Deep learning techniques allow learning feature representations directly from the data, significantly pushing the performance boundary of human pose estimation. In this paper, we reap the recent achievements of 2D human pose estimation methods and present a comprehensive survey. Briefly, existing approaches put their efforts in three directions, namely *network architecture design*, *network training refinement*, and *post processing*. Network architecture design looks at the architecture of human pose estimation models, extracting more robust features for keypoint recognition and localization. Network training refinement tap into the training of neural networks and aims to improve the representational ability of models. Post processing further incorporates model-agnostic polishing strategies to improve the performance of keypoint detection. More than 200 research contributions are involved in this survey, covering methodological frameworks, common benchmark datasets, evaluation metrics, and performance comparisons. We seek to provide researchers with a more comprehensive and systematic review on human pose estimation, allowing them to acquire a grand panorama and better identify future directions.

Keywords Human pose estimation · Pose estimation · Survey · Deep learning · Convolutional neural network

1 Introduction

As a compelling and fundamental problem in computer vision, human pose estimation (HPE) has attracted intense attention in recent years. As shown in Fig. 1, the goal of 2D HPE is to: 1) recognize different person instances within the multimedia data (RGB images, videos, RF signals, or radar) recorded by sensors, and 2) to localize a set of predefined human anatomical keypoints for each person. As the cornerstone of human-centric visual understanding, 2D HPE provides the groundwork for tackling multitudinous higher-order computer vision tasks such as 3D human pose estimation [16, 37–39, 111, 112, 176, 190, 206], human action recognition [4, 66, 167], human parsing [43, 44, 140], pose tracking [41, 172, 181], motion prediction [98, 100, 102], human motion retargeting [13, 74, 120], and vision-and-language conversion [22, 49–51, 117]. HPE supports a wide spectrum of applications including human behaviors understanding, motion capture, violence detection, crowd riot scene identification, human-computer interaction, and autonomous driving.

Earlier methods [143, 164, 174, 196] adopt the probabilistic graphical model to represent relations between joints. Unfortunately, these methods rely heavily on hand-crafted features which limit their generalization and performance. More recently, the deep learning techniques [81, 90, 103, 146, 147] enable learning feature representations automatically from data, which has significantly contributed to the advancement of human pose estimation. These deep learning-based approaches [9, 86, 91, 99, 101, 153, 161, 181], commonly building upon the success of convolutional neural networks, have achieved outstanding performance on this task.

Given the rapid development, this paper seeks to track recent progress and summarize their accomplishments to deliver a clearer panorama for 2D human pose estimation.

Haoming Chen and Runyang Feng have equal contribution.

✉ Fengcheng Zhou
zfc@zjsu.edu.cn

[1] Zhejiang Gongshang University, Hangzhou, China

[2] Zhejiang Lab, Hangzhou, China

Published online: 11 November 2022

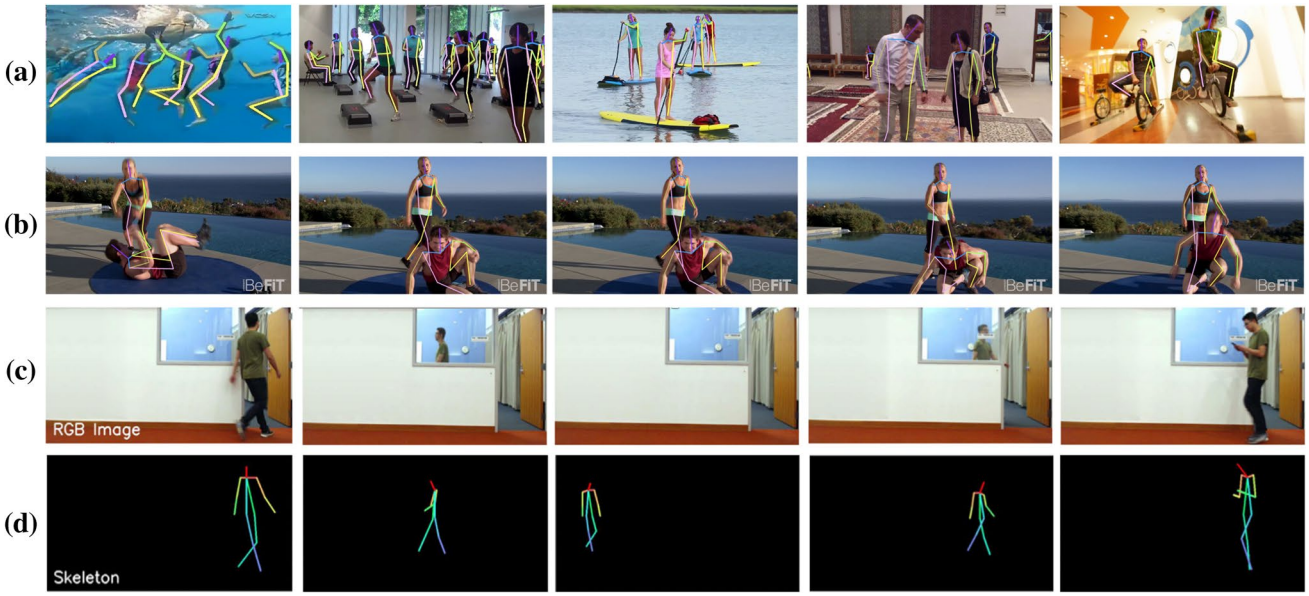

Fig. 1 An illustration of 2D human pose estimation on multimedia data, including images **a**, videos **b**, and RF signals **c**, **d**. Note that RGB images **c** are presented for visual reference of RF signals-based HPE, and **d** shows the skeleton extracted from the RF signals *alone*. The pictures in **c** and **d** are cited from [199]

Several excellent surveys related to human pose estimation have been published, as presented in the Table 1, involving studies in areas of human motion capture and analysis [67, 114, 115, 134], activity recognition and 2D/3D HPE [18, 97, 200], *etc*. However, few surveys are dedicated to 2D human pose estimation. On the other hand, most of existing surveys cast existing approaches into *single-person* and *multi-person* pose estimation methods. The *single-person pose estimators* typically focus on the model architectures for keypoint detection, and can perform well in the multi-person pose estimation scenarios by predicting pose for each individual person within his/her bounding box. Therefore, a pose estimation model can accommodate both *single-person* and *multi-person* scenes, and the division as above might be unnecessary. Moreover, while human pose estimation on images or videos has been widely concerned, to the best of our knowledge there is still no work that summarizes signal-based human pose estimation, *e.g.*, RF signals and radar signals.

In this paper, we roughly cast human pose estimation methods into three categories, each containing several sub-categories on a finer level. (1) *Network architecture design* approaches attempt to devise vigorous models that capture robust representations across different scenes to effectively detect keypoints. Methods in this category concentrate on extracting and processing human body features within a person bounding box [34, 181] or over the entire image [9, 19]. (2) *Network training refinement* approaches aim at optimizing neural network training, trying to improve the model ability without changing the network structure. Towards this aim, they engage in data augmentation techniques [6, 171], model training strategies [124, 180], loss function constraints [17, 203], and domain adaption methods [55, 183]. (3) *Post processing* methods focus on pose polishment upon the coarse pose estimates to improve the performance. The methods within this category usually behave as a model-agnostic plugin. Representative techniques for pose polishement include quantization error minimization [58, 193] and pose resampling [99, 172]. *Furthermore*, we also discuss the rarely involved topic of reconstructing 2D human poses from signals such as RF signals [165, 199] and radar signals [83], hoping to fill the knowledge gap.

1.1 Scope

Our scope is limited to 2D human pose estimation with deep learning, we do not consider the conventional non-deep-lear-ning methods. Topics such as the applications of 2D HPE [200] and the representations of human body models [18] that have been adequately covered by other reviews will not be detailed here either. Nevertheless, there are still a breathtaking number of papers on 2D HPE, hence it is necessary to establish a selection criterion, in such a way that we restrict our attention to the top journal and conference papers since 2014. In light of these constraints, we sincerely apologize to those authors whose works are not incorporated into this paper.

Table 1 Summary of previous surveys and reviews related to the human pose estimation

Survey title	Year	Venue	Content	Single-person	Multi-person
A survey of computer vision-based human motion capture [114]	2001	CVIU	A survey of different functionalities in motion capture system, including initialization, tracking, pose estimation, and recognition	✓	✓
A survey of advances in vision-based human motion capture and analysis [115]	2006	CVIU	A survey of advances in human motion capture and analysis from 2000 to 2006	✓	✓
Vision-based human motion analysis: an overview [134]	2007	CVIU	An overview of markerless vision-based human motion analysis	✓	✓
Advances in view-invariant human motion analysis: a review [67]	2010	TSMCS	A review of major issues in human motion analysis system, including human detection, view-invariant pose representation and estimation, and human behavior understanding	✓	✓
Visual analysis of humans [116]	2011	Book	A comprehensive overview of human analysis such as pose estimation and applications	✓	✓
Human pose estimation and activity recognition from multi-view videos: comparative explorations of recent developments [57]	2012	JSTSP	A review of multi-view based 3D human pose estimation and activity recognition	✓	
A survey of human pose estimation: the body parts parsing based methods [97]	2015	JVCIR	A survey of human parsing based 2D/3D human pose estimation	✓	✓
Human pose estimation from monocular images: a comprehensive survey [45]	2016	Sensors	A survey of conventional and deep learning methods for human pose estimation		✓
3d human pose estimation: a review of the literature and analysis of covariates [145]	2016	CVIU	A review of the advances in 3D human pose estimation from RGB images or image sequences	✓	
Monocular human pose estimation: a survey of deep learning-based methods [18]	2020	CVIU	A survey of monocular based 2D/3D human pose estimation employing deep learning methods	✓	✓
The progress of human pose estimation: a survey and taxonomy of models applied in 2D human pose estimation [119]	2020	IEEE access	A survey of researches on 2D human pose estimation	✓	✓
Deep learning-based human pose estimation: a survey [200]	2020	arXiv	A survey of deep learning-based 2D/3D human pose estimation	✓	✓

1.2 Outline

The rest of this paper is organized as follows. In Sect. 2 we provide problem formulations for 2D human pose estimation, and briefly discuss the technological challenges of 2D HPE. Then, we present works on network architecture design in Sect. 3, introduce network training refinement methods in Sect. 4, and review post processing approaches in Sect. 5. Subsequently, we summarize the common benchmark datasets, evaluation metrics, and performance comparisons in Sect. 6. We further provide discussions in Sect. 7, including open questions, signal-based 2D HPE, and future research directions. Finally, we conclude the paper in Sect. 8.

2 Problem statement

In this section, we first define the problem of 2D HPE on the image and video data, followed by the discussion of technological challenges in this task.

2.1 The problem

Formally, the human pose estimation problem can be formualted as follows. Given an image or a video as input, the goal is to detect the **poses** of all persons in the input data. Technically, presented with an observed image I, we aim to

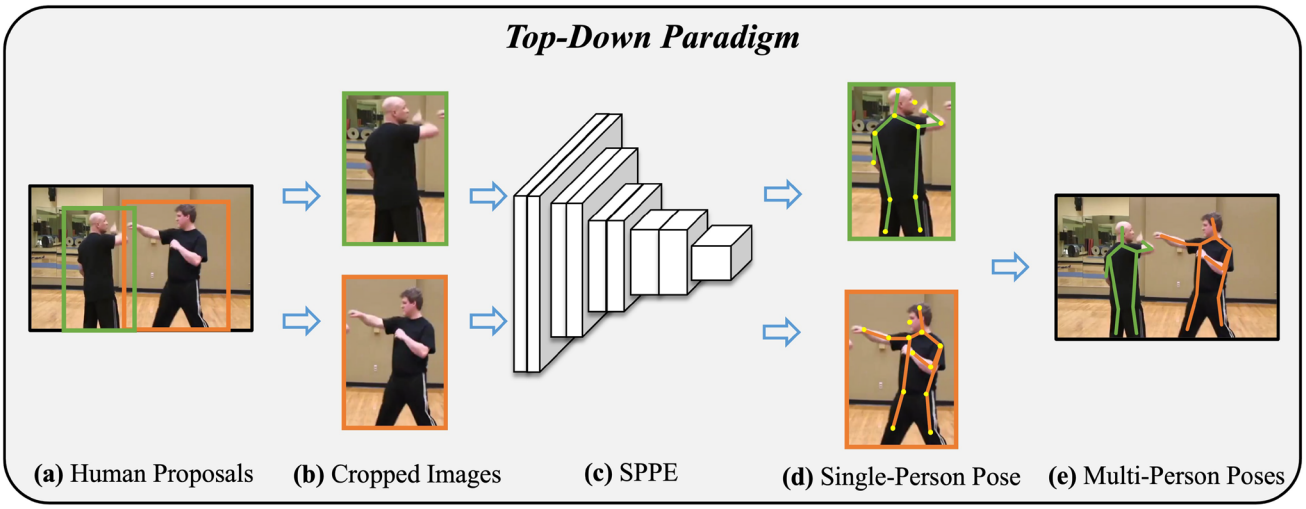

Fig. 2 A classical pipeline of top-down framework for human pose estimation. **a** Original image in the dataset. The goal is to detect the poses of all persons in the input image. An off-the-shelf object detector is employed to perform person detection and gives the human proposals. **b** The regions of human proposals are cropped from the original image to form the single person images. **c** Each cropped image is subjected to single person pose estimation (SPPE) to obtain the estimated pose, which is illustrated in **d**. **e** All estimated poses are projected to the original image and yield the final results

detect the pose of each person i in the image $\mathbf{P} = \{\mathbf{P}_i\}_{i=1}^{n}$, where n denotes the number of persons in I.

To describe human poses, skeleton-based model [35], contour-based model [73], and volume-based model [148] have been proposed in previous works. In particular, the contour-based representation contains rough body contour and limb width information while the volume-based representation describes 3D human shapes. The skeleton-based model, which characterizes the human body as a set of pre-defined joints, has been widely employed in 2D HPE.

2.2 Technical challenges

Ideally, an algorithm that is both highly accurate and efficient is desired to solve the problem of 2D HPE. High accuracy detection ensures a precise human body information to facilitate downstream tasks such as 3D HPE and action recognition, while high efficiency allows real-time computing in different devices such as desktops and mobile phones.

Challenges in *accurate* pose detection come from several aspects. (1) Nuisance phenomena such as under/over-exposure and human-objects entanglement frequently occur in real-world scenes, which may easily lead to detection failure. (2) Due to the highly flexible human kinematic chains, pose occlusions even self-occlusions in many scenarios are inevitable, which will further confuse keypoint detectors using visual features. (3) Motion blur and video defocus do frequently happen in videos, which deteriorates the accuracy of pose detection.

When the pose estimation algorithms are applied to practical applications, besides accurate estimation, the running speed (*efficiency*) is also important. However, high accuracy and high efficiency are often in conflict to each other since the high accuracy models tend to be deeper, requiring increased resources for computation and storage. For example, HRNet-W48 [153] has achieved state-of-the-art results on multiple benchmarks, which however has difficulties in achieving real-time pose estimation even with the help of powerful NVIDIA GTX-1080TI GPUs. Consequently, lightweight models with comparable precision are much coveted for mobile or wearable devices.

3 Network architecture design methods

A key advantage of modern deep learning methods is the ability to learn feature representations automatically from data. However, feature quality is closely related to the network architecture, therefore the topic of network design deserves to be investigated deeply. Correspondingly, *network architecture design* methods aim at extracting powerful features by investigating various network designs to address human pose estimation. In this section, we set out to introduce these approaches in detail with a focus on their network architectures.

On a high level, these approaches typically fall into two general frameworks, namely **top-down** framework [5, 34, 99, 122, 153, 178] and **bottom-up** framework [9, 40, 70, 78, 106, 177]. The top-down paradigm employs a two-step procedure that first detects human bounding boxes and then performs single person pose estimation for each bounding box, which is exemplified in Fig. 2. The bottom-up paradigm

adopts the part-based procedure that first locates identity-free keypoints and then groups them into different person instances. We may further divide different methods in these two paradigms into fine-grained sub-categories, where the top-down approaches are categorized into *regression-based* [12, 161], *heatmap-based* [153, 181], *video-based* [99, 104], and *model compressing-based* [187, 194] methods, and the bottom-up approaches are classified into *one stage* [40, 127] and *two-stage* methods [9, 78]. In what follows, we introduce these categories in detail.

3.1 Top-down framework

3.1.1 Regression-based methods

Earlier works [12, 33, 36, 89, 135, 154, 155, 161, 170, 192, 198] attempt to learn a mapping from input image to the pre-defined kinematic joints via an end-to-end network, and directly regress the keypoint coordinates, which we refer to as the *regression-based* approaches.

For instance, DeepPose [161] sets the precedent of human pose estimation with deep learning technique. It [161] first employs an iterative architecture to extract image features with the cascaded convolutional neural networks (AlexNet [79]), and subsequently regresses the joint coordinates with fully connected layers. Inspired by the remarkable performance of deep learning works such as DeepPose, researchers gradually turned from conventional methods to the deep learning ones. Building upon the GoogleNet [12, 156] proposes a self-correcting model, which progressively changes the initial joint coordinates estimations instead of directly predicting joint positions. [154] presents a structure-aware regression approach that utilizes a novel re-parameterized pose representation of bones. This method is constructed on the ResNet50 [52], and is able to capture more structural human body information such as joint connections, which enriches the pure joint-based pose descriptions.

Graph convolutional network (GCN) [76] has recently been widely explored, which employs nodes and edges to represent entities and their correlations. Upon convolutions on the graph, the feature of a node is enhanced by incorporating features from the neighboring nodes. Compared to traditional methods, GCN provides another competitive and novel model to characterize the human body. Qiu et al. [135] casts the human body as a graph structure where the nodes represent joints and the edges represent bones, and proposes to estimate invisible joints using an Image-Guided Progressive GCN module.

Attention mechanism has greatly advanced the representation learning, and the Transformer [11, 64, 163, 205] built upon self-attention has established new state-of-the-arts on multiple visual understanding tasks such as object detection, image classification, and semantic segmentation. Li et al. [87] presents a cascaded Transformers performing end-to-end regression of human and keypoint detection, which first detects the bounding boxes for all persons and then separately regresses all joint coordinates for each person.

The regression-based methods are highly efficient and show promising potential in real-time applications. Unfortunately, such approaches directly output a single 2D coordinates for each joint, failing to consider the area of the body part. To tackle this issue, heatmap-based approaches are introduced, which localize the keypoints by probabilistic heatmaps instead of determined coordinates.

3.1.2 Heatmap-based methods

In order to overcome the shortcomings of direct coordinate regression, heatmap-based joint representations have been widely adopted [131], which leads to an easier optimization and a more robust generalization. Specifically, the heatmap H_i is generated via a 2D Gaussian centered at each joint location (x_i, y_i), encoding the probability of the location being the i^{th} joint. During training, the goal is to predict N heatmaps $\{H_1, H_2, .., H_N\}$ for a total of N joints. Representative *heatmap-based* approaches include:

Iterative architecture Conventionally, the iterative architecture [12, 104, 137, 161, 178] is designed to produce and refine the keypoint heatmaps. Ramakrishna et al. [137] presents an inference machine model which gradually infers the locations of joints in multiple stages. Wei et al. [178] further extends the architecture of [137] and builds a sequential prediction framework, which employs sequential convolutions to implicitly model long-range spatial dependencies between human body parts. This approach harvests increasingly refined estimates for joint locations by operating on the results of previous stage, as shown in Fig. 3. Wei et al. [178] additionally proposes intermediate supervision to alleviate the inherent problem of *vanishing gradients* in the iterative architectures.

Although the intermediate supervision strategy relieves the *vanishing gradients* of multi-stage models, each stage still fails to build a deep sub-network to extract effective semantic features, which greatly limits their fitting capabilities. This issue has been tackled with the emergence of residual network (ResNet) [52], which introduces a shortcut and allows the errors at deeper layers to be back-propagated. Benefiting from such a way, numerous large models [8, 17, 20, 68, 75, 96, 122, 152, 153, 158, 181, 184] have been devised, which greatly boost the process of 2D HPE.

Symmetric architecture The deep models generally employ a *high-to-low* (downsampling) and *low-to-high* (upsampling) framework, where *high* and *low* denote the resolution of feature representations. Newell et al. [122] proposes a novel stacked hourglass architecture based on the successive steps of pooling and upsampling, which

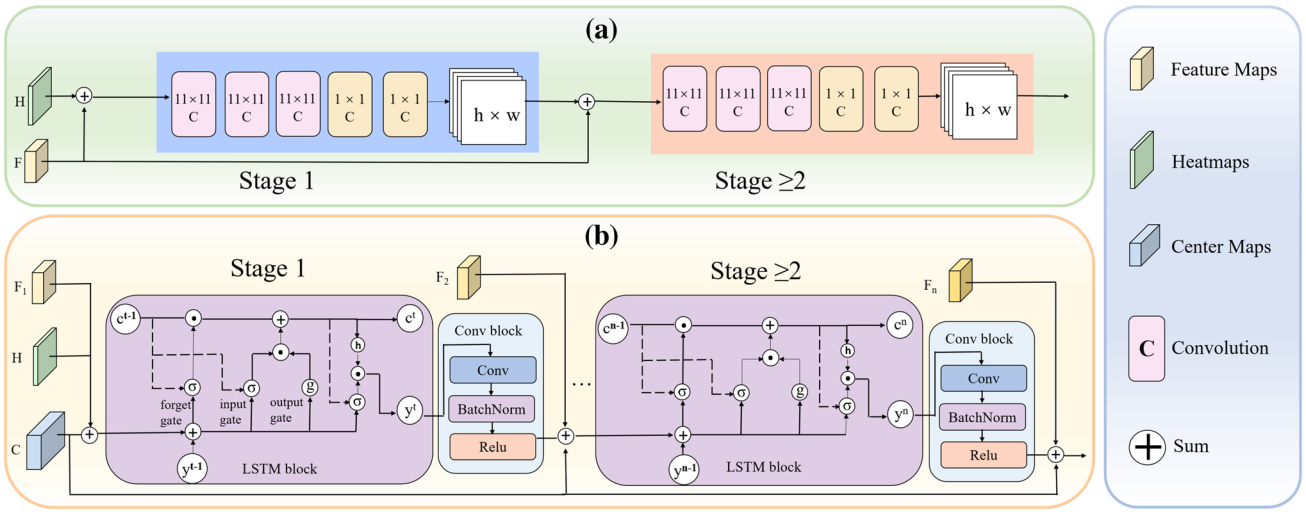

Fig. 3 Illustration of the networks based on iterative architecture. The top portion **a** of the figure depicts the structure of Convolutional Pose Machine [178] while the bottom part **b** shows the network of LSTM Pose Machines [104]. In [178], the prediction of each stage and image features are concatenated for the subsequent stage. [104] extends [178] with LSTM. The heatmaps predicted at the previous stage, frame features, and a center map are concatenated to fed into the subsequent stage. Note that different stages in [178] aim at optimizing pose estimation of the same image, while different stages in [104] process various video frames

incorporates features across all scales to capture the various spatial relationships between joints. The stacked hourglass architecture is depicted in Fig. 4a. Several variations [8, 20, 75, 184] that built upon the success of this stacked hourglass architecture are subsequently developed. Specifically, [20] extends [122] to Hourglass Residual Units with a side branch including filters with larger receptive field, which greatly increases the receptive fields of the network and automatically learns features across different scales. [184] further replaces the residual blocks in the stacked hourglass [122] with the Pyramid Residual Modules which enhances the scale invariance of networks. [75] proposes a multi-scale supervision that combines the keypoint heatmaps across all scales, which leads to acquiring abundant contextual features and improves the performance of stacked hourglass network. Cai et al. [8] designs a stacked hourglass-like network, *i.e.*, Residual Steps Network which aggregates features with the same spatial size to produce the delicate localized descriptions. Tang et al. [157] employs the hourglass network [122] as backbone, and proposes a part-based branching network to learn the representations specific to different part groups. These hourglass-based models retain *symmetric* architecture between high-to-low and low-to-high convolutions.

Asymmetric architecture Another line of work exploits an *asymmetric* architecture [17, 62, 181], where the high-to-low process is *heavy* and the low-to-high process is *light*. Chen et al. [17] proposes a Cascaded Pyramid Network (Fig. 4c) that detects the simple keypoints with a GlobalNet, and handles the difficult keypoints with a RefineNet. Specifically, the RefineNet consists of several regular convolutions, integrating all levels of feature representations from the GlobalNet. Xiao et al. [181] extends the ResNet [52] by adding a few deconvolutional layers instead of feature map interpolation, which is depicted in Fig. 4b. These methods employ a sub-network of classical classification networks (VGGNet [149] and ResNet [52]) for *high-to-low* convolution and adopt simple networks for *low-to-high* convolution. Undoubtedly, such asymmetric network architectures suffer from imbalances in feature encoding and decoding, which potentially affects model performance.

High resolution architecture Unlike previous models, [153] proposes a representative network, HRNet[1] (Fig. 4d), which is able to maintain high resolution representations through the whole process, achieving state-of-the-art results on multiple vision tasks. This work demonstrates the superiority of high-resolution representations for human pose estimation and inspires a wide spectrum of later researches [68, 99, 172]. Jiang et al. [68] takes HRNet as the backbone network, and further incorporates the gating mechanism as well as feature attention module to select and fuse discriminative and attention-aware features.

Composed human proposal detection The above models concentrate on pose estimation on a given human proposal which is cropped from the entire image, and simply employ off-the-shelf human proposal detectors for proposal identification. Existing work [34, 84] has demonstrated that the quality of human proposals (*e.g.*, human position and redundant detection) significantly affects the results of pose

[1] Link of HRNet Project: https://github.com/leoxiaobin/deep-high-resolution-net.pytorch.

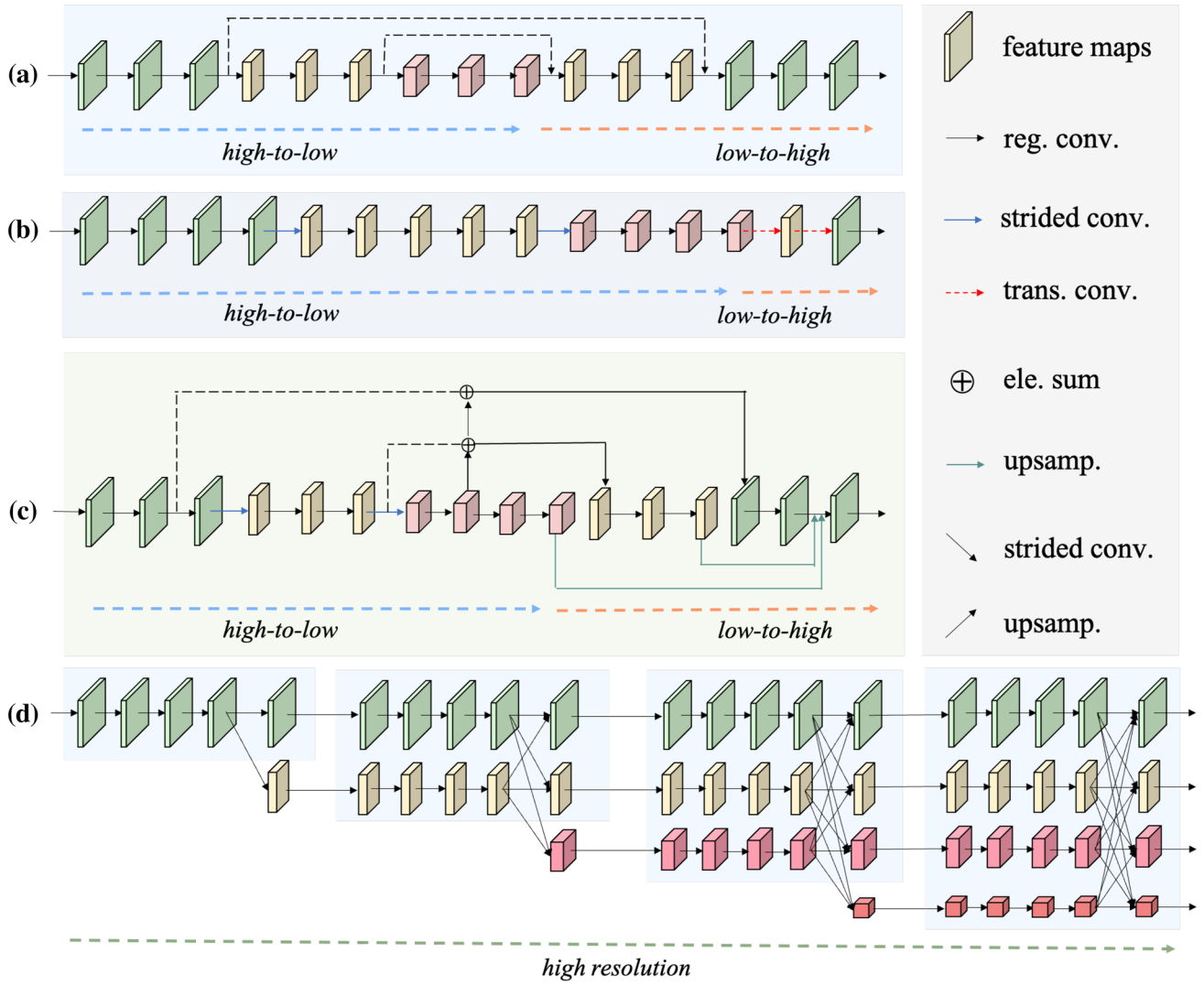

Fig. 4 Illustration of the classical human pose detector that relys on the *high-to-low* and *low-to-high* framework. **a** Stacked hourglass network [122]. **b** Cascaded pyramid networks [17]. **c** SimpleBaseline [181]. **d** HRNet [153]. Legend: *reg. conv.* = regular convolution layer, *strided conv.* = strided convolution layer for learnable downsampling, *trans. conv.* = transposed convolution layer for learnable upsampling, *ele. sum* = element-wise summation. For the architecture of (a) stacked hourglass, the high-to-low and low-to-high network architectures are symmetric. In **b** and **c**, the high-to-low process is performed by a large visual backbone network (ResNet) which is **heavy**, while the low-to-high process is implemented by some transposed convolutions or directly upsampling, which is **light**. In **c**, the skip-connection (dashed lines) aims to fuse the features with same spatial size in the high-to-low and low-to-high process. In HRNet **d**, the high resolution representation is maintained in the entire propagation, and repeated multi-scale fusions are performed, each resolution features receive rich information from all resolutions

estimators. Therefore, a group of researches direct their efforts in refining human proposals. For instances, [128] presents a multi-person pose estimation method, which employs the Faster-RCNN [138] as person detector and the ResNet-101 [52] as pose detector, and additionally proposes a novel keypoint NonMaximum-Suppression (NMS) strategy to address the problem of pose redundancy. [34] utilizes the SSD-512 [95] as human detector and the stacked hourglass [122] as single person pose detector, and further proposes a symmetric spatial transformer network to extract a high-quality single person region from an inaccurate bounding box to facilitate human pose estimation. Li et al. [84] notices that single person bounding boxes in crowded scenes tend to contain multiple people, which deteriorates the performance of the pose detector. To tackle this problem, [84] leverages a joint-candidate pose detector to predict the heatmaps with multiple peaks, and uses a graph network to perform global joints association.

In contrast, another group of researches propose to perform proposal detection and pose detection jointly. Varamesh

and Tuytelaars [162] develops a mixture model which simultaneously infers the human bounding boxes and keypoint locations in a dense regression fashion. [177] introduces a template offset model which first gives a good initialization for the human bounding boxes and poses, and then regresses the offsets between initialization and corresponding labels. Kocabas et al. [77] presents a MultiPoseNet which first detects the keypoints and human proposals separately, and then employs a Pose Residual Network to assign the detected keypoints to different bounding boxes. Specifically, the Pose Residual Network is implemented by a residual multilayer perceptron. Mao et al. [109] designs a pose estimation framework, which incorporates dynamic instance-aware convolutions and eliminates the process of bounding boxes cropping and keypoint grouping.

Overall, heatmap-based methods are more popular than the regression-based paradigms due to their higher accuracy. However, the heatmap computation process brings new open problems, including expensive computational overhead and inevitable quantization error.

3.1.3 Video-based methods

Human pose estimation on videos has also been a hot research topic. The video, by nature, brings more challenges such as *camera shift, rapid object movement*, and *defocus*, which result in frame quality deterioration frequently. On the other hand, different from still images, there exist abundant temporal clues across video frames (*e.g.*, *temporal dependency* and *geometric consistency*), which provide valuable information for pose estimation.

We observe that most existing methods are trained on static images. Directly applying the image-based models to videos (image sequence) might lead to unsatisfactory results since they fail to consider the temporal consistency across video frames. To conquer this dilemma, a large number of approaches have explored utilizing the additional temporal information to achieve higher pose detection accuracy. According to how the temporal information is exploited, we broadly divide these approaches into *optical flow-based* [14, 131, 151, 191, 192], *RNN-based* (Recurrent Neural Networks) [3, 42, 104], *pose tracking-based* [41, 169, 172, 186, 188, 201], and *key frame-based* [5, 15, 99, 126, 198] paradigms. Below, we elaborate these methods in detail.

Optical flow *Optical flow* models the apparent motion of individual pixels on the frame, attracting widespread attention [26, 61]. The optical flow across frames usually reveals the motions of the human subjects, which are obviously useful for pose estimation. [131] combines convolutional networks and optical flow into a uniform framework, which employs the flow field to align the features temporally across multiple frames, and utilizes the aligned features to improve the pose detection in individual frames. Song et al. [151] presents a Thin-Slicing Network which computes the dense optical flow between every two frames to propagate the initial estimation of joint position through time, and uses a flow-based warping mechanism to align the joint heatmaps for subsequent spatiotemporal inference. [14] focuses on human pose estimation in crowded scenes, which incorporates forward pose propagation and backward pose propagation to refine the pose of the current frame. However, although the optical flow in these methods does contain useful features such as human motion information, the undesired background changes are also involved. The *noisy* motion representation greatly hinders them from obtaining expected performance. [192] proposes a novel deep motion representation, *namely PoseFlow*, which is able to reveal human motion in videos while inhibiting some nuisance noises such as background and motion blur. The distilled robust flow representation can also be generalized to human action recognition tasks.

The optical flow based representation can model the motion cues at the pixel level, which is favorable for capturing useful temporal information. However, the optical flow is only able to extract impure features and is quite sensitive to noises.

Recurrent neural network Besides optical flow, *Recurrent Neural Network (RNN)* also provides a way to model temporal contexts across frames. RNN shows a promising performance in sequential prediction task, due to the nature that each output is jointly determined by the current input and the historical predictions. Therefore, a group of approaches attempt to capture temporal contexts between video frames by RNN for improving pose estimation. Gkioxari et al. [42] presents a sequence-to-sequence model, which employs the chained convolutional networks to process input images, and combines historical hidden status and current images to predict current keypoint heatmaps. Luo et al. [104] extends the convolutional pose machine [178] by using convolutional LSTM, which is able to model both spatial and temporal contexts for pose prediction.

To our knowledge, existing RNN-based methods can effectively estimate human poses from the *single-person* image sequence, yet they have not been applied to multi-person videos until now. We conjecture that RNN has difficulties in directly employing temporal information from multi-person videos, where extracting the temporal contexts of each person will be affected by the others.

Pose tracking To alleviate the issue of RNN, some methods that built upon the *pose tracking* have been proposed, which establish a tracklet for each person in video frames to filter the interference of irrelevant information. [41] proposes a 3D Mask R-CNN (extension of Mask R-CNN [53] to include a temporal dimension) to generate small clips for a single person, and leverages temporal information within the small clips to produce more accurate

predictions. Zhou et al. [201] proposes a pose estimation framework which consists of a temporal keypoint matching module and a temporal keypoint refinement module. Specifically, the temporal keypoint matching module gives reliable single-person pose sequences according to the keypoint similarities, and the temporal keypoint refinement module aggregates poses within the sequence to correct original poses. Wang et al. [172] designs a Clip Tracking Network and a Video Tracking Pipeline to establish the tracklet for each person, and extends the HRNet [153] to 3D-HRNet to perform temporal pose estimation for all tracklets. Yang et al. [186] employs a graph neural network to learn the pose dynamics from the historical pose sequence, and incorporates the pose dynamics into the pose detection of the current frame.

Pose tracking-based methods show strong adaptation in the scene of multi-person. However, these models require computing feature similarity or pose similarity to create tracklets, which invokes an extra overhead for pose estimation.

Key frame optimization In addition to exploiting temporal information from tracklets, it is also beneficial to select some key frames to refine the pose estimation of the current frame, what we refer to as *keyframe-based* approaches. Charles et al. [15] proposes a personalized video pose estimation framework, which leverages a few key frames with high-precision pose estimates to fine-tune the model. Bertasius et al. [5] proposes a PoseWarper network which first warps poses of the labeled frames to the unlabeled (current) frame, and then aggregates all warped poses to predict the pose heatmaps of the current frame. Zhang et al. [198] presents a keyframe proposal network to select the effective key frames, and proposes a learnable dictionary to reconstruct entire pose sequence from the selected key frames. The work in [99] builds a dual consecutive framework for video pose estimation, termed DCPose[2], which incorporates consecutive frames from dual temporal directions to improve the pose estimation in videos. Specifically, three modular components are designed. A Pose Temporal Merger encodes keypoint spatiotemporal context to generate effective searching scopes while a Pose Residual Fusion module computes weighted pose residuals in dual directions. These are then processed via a Pose Correction Network for efficient refining of pose estimations. It is worthy mentioning that the DCPose [99] is able to fully leverage the temporal information from neighboring frames and achieves state-of-the-art performance on video-based human pose estimation.

[2] Link of DCPose Project: https://github.com/Pose-Group/DCPose.

3.1.4 Model compression-based methods

For practical applications on lightweight devices such as mobiles, a low-consumption and high-accuracy HPE method is urgently demanded. However, the majority of existing pose estimation models are oversized, which require extensive computational resources and fail to reach real-time computation. Consequently, these methods are usually low-efficient, which limits their potential usage especially for mobiles or wearable equipments. To alleviate this problem, many model compression based methods [82, 104, 126, 187, 194] have been proposed to achieve the trade-off between accuracy and efficiency. These methods are able to significantly reduce model parameters with small accuracy decline.

[194] proposes a Fast Pose Distillation model that built upon the Teacher-Student network [56, 113, 139, 168, 202], effectively transferring the human body structure knowledge from a strong teacher network (*large model*) to a *lightweight* student network. Specifically, the 8-stage Hourglass model is employed as the teacher network while a compact counterpart (4-stage Hourglass) is adopted as the student network. Luo et al. [104] proposes a lightweight LSTM architecture to perform video pose estimation. Yu et al. [187] proposes two schemes to reduce the parameters of HRNet: i) Simply applying the Shuffle-Block [197] to replace the basic block in *vanilla* HRNet. ii) Designing a conditional channel weighting module, which learns the weights across multiple resolutions to replace the costly point-wise (1×1) convolutions. By simplifying the original HRNet [153], the Lite-HRNet [187] shows good performance with relatively fewer parameters.

3.1.5 Summary of top-down framework

The architecture of top-down framework comprises the following key components: an object detector for producing human bounding boxes, and a pose estimator for detecting human keypoint locations. The object detector determines the performance of human proposal detection, and further influences pose estimation. The pose detector, on the other hand, is the core of the framework and directly determines the accuracy of pose estimation. In summary, the top-down framework is highly scalable that can be constantly improved with advances of object detectors as well as pose detectors.

3.2 Bottom-up framework

The major discrepancy between bottom-up and top-down frameworks is whether the human detector is employed to detect the human bounding boxes. Compared to the top-down approaches, bottom-up approaches do not rely on human detection and directly perform keypoint estimation in the original image, thus reducing the computational

overhead. However, this procedure opens up a new challenge: How to judge the identities of estimated joints? According to the way of determining the identities of estimated keypoints, we divide the bottom-up methods into *human center regression-based* [40, 123–125] , *associate embedding-based* [19, 69, 106, 121], and *part field-based* [9, 55, 62, 70, 77, 78, 85, 105, 132, 133, 136, 173] approaches.

Human center regression The *human center regression-based* approaches utilize a human center point to represent the person instance. Nie et al. [125] proposes a Single-stage multi-person Pose Machine that unifies person instance and body joint position representations. In [125], the root joints (center-biased points) are introduced to denote the person instances, and body joint locations are encoded into their displacements *w.r.t.* the roots. Geng et al. [40] predicts a human center map that indicates the person instance, and densely estimates a candidate pose at each pixel q within the center map.

Associate embedding The *associate embedding-based* approaches assign each keypoint an associate embedding, which is an instance representation for distinguishing different persons. [121] pioneers the embedding representation, where each predicted keypoint has an additional embedding vector that serves as a *tag* to identify its human instance assignment. Jin et al. [69] proposes a SpatialNet to detect body part heatmaps and predict part-level data association in the input image. Specifically, the part-level data association is parameterized by the keypoint embedding. Cheng et al. [19] follows the keypoints grouping in [121] and further proposes a Higher-Resolution Network to learn high-resolution feature pyramids, improving the pose estimation of small persons. Luo et al. [106] focuses on the problems of large variance of human scales and labeling ambiguities. This approach [106] proposes a scale-adaptive heatmap regression model, which is able to adaptively adjust the standard deviation of the ground-truth gaussian kernels for each keypoint, and achieves high tolerance for different human scales and labeling ambiguities.

Part field The *part field-based* methods first detect keypoints and connections between them, and then perform keypoint grouping according to the keypoint connections. The representative work [9] proposes a two-branch multi-stage CNN architecture, where one branch predicts the confident maps to denote the locations of keypoints and another branch predicts the Part Affinity Fields to indicate the connective intensity between keypoints. Then, [9] applies a greedy algorithm to assemble different joints of the same person, according to the connective intensity between joints. Inspired by [9], various attempts have been proposed. Kreiss et al. [78] utilizes a part intensity field to localize body parts, and employs a part association field to associate body parts with each other. Li et al. [85] presents a novel keypoint associated representation of *body part heatmaps* based on the Part Affinity Field [9] for effective keypoint grouping. Some approaches explore alternative representations of keypoint connection for keypoint grouping. [105] proposes a multi-layer fractal network, which regresses the keypoint location heatmaps and infers kinships among adjacent joints to determine the optimal matched joint pairs. [70] proposes a differentiable Hierarchical Graph Grouping network that converts the keypoint grouping into a graph grouping problem, and can be trained end-to-end with the keypoint detection network.

Summary Overall, the bottom-up approaches improve the efficiency of pose detection by eliminating the usage of additional object detection techniques. Due to the high efficiency, the bottom-up methods are promising in practice applications. For example, the open source project[3] of OpenPose [10] has been extensively adopted in the industry.

4 Network training refinement

From the perspective of the overall training pipeline in neural networks, the quantity and quality of data, training strategy, and loss function will impact the model performance. According to the above key phases during training, we classify the network training refinement approaches into *data augmentation techniques*, *multi-task training strategies*, *loss function constraints*, and *domain adaption methods*. Data augmentation techniques aim to increase the amount and diversity of the data. Multi-task training strategies seek to capture informative features by sharing representations among related visual tasks. Loss function constraints determine the optimization objective of the network. Domain adaption methods aim to help the network adapt different datasets. In this section, we introduce these methods in detail.

4.1 Data augmentation techniques

Deep learning is typically data-driven, therefore data plays a crucial role in model training. A *large-scale* and *high-quality* dataset contributes to the robustness of models. However, building such a wonderful dataset is time-consuming and expensive. To alleviate this problem, data augmentation techniques are adopted to increase the number and diversity of samples in datasets.

In 2D human pose estimation, common data augmentation techniques include random rotation, random scale, random truncation, horizontal flipping, random information dropping, and illumination variations. Apart from the

[3] Link of OpenPose Project: https://github.com/CMU-Perceptual-Computing-Lab/openpose.

above random schemes, several works [6, 59, 118, 130, 171, 204] have been studying learnable data augmentation. Peng et al. [130] proposes an enhancement network that generates difficult pose samples to compete against the pose estimator. Tang et al. [157] points out that state-of-the-art human pose estimation approaches have similar error distributions. Moon et al. [118] generates synthetic poses based on the error statics in [157] and employs the synthesized poses to train human pose estimation networks. Bin et al. [6] presents an adversarial semantic data augmentation using the generative adversarial network (GAN [46]), which enhances original images by pasting segmented body parts with different semantic granularities. [171] introduces an AdvMix algorithm, in which a generator network confuses pose estimators by mixing various corrupted images, and a knowledge distillation network transfers clean pose structure knowledge to the target pose detector.

4.2 Multi-task training strategies

Most of the human pose estimation models are designed for single-task learning. In this subsection, we focus on the multi-task learning models related to 2D human pose estimation. *Multi-task learning* aims at capturing informative features by sharing representations among related visual tasks. Human parsing is a closely related task to human pose estimation, with the goal of segmenting the human body into semantic parts such as head, arms, and legs, etc. Previous works [25, 27, 80, 92, 124, 180] employ the human parsing information to improve the performance of 2D HPE. Xia et al. [180] jointly solves the two tasks of human parsing and pose estimation, and utilizes the part-level segments to guide the keypoint localization. [124] presents a parsing encoder and a pose model parameter adapter, which together learn to predict parameters of the pose model to extract complementary features for human pose estimation.

4.3 Loss function constraints

Loss function determines the learning objective of the network, and greatly affects the performance of the model. In this subsection, we summarize and discuss existing loss functions [12, 17, 54, 75, 85, 106, 133, 154, 189, 203] of 2D HPE.

The standard and common loss function of human pose estimation is the L_2 distance. Training aims to minimize the total L2 distance between prediction and ground truth heatmaps for all joints. The cost function is defined as:

$$L = \frac{1}{N} * \sum_{j=1}^{N} v_j \times ||G(j) - P(j)||^2 \tag{1}$$

Where $G(j)$, $P(j)$ and v_j respectively denote the ground truth heatmap, prediction heatmap and visibility for joint j. The symbol N denotes the number of joints.

[75] presents a multi-scale human structure-aware loss which captures the structural information of the human body. The *structure-aware loss* at the i^{th} feature scale can be expressed as follows:

$$L^i = \frac{1}{N} \sum_{j=1}^{N} ||P^i_j - G^i_j||_2 + \alpha \sum_{i=1}^{N} ||P^i_{S_j} - G^i_{S_j}||_2, \tag{2}$$

where P_j and G_j denote the predicted and labeled j^{th} keypoint heatmaps, P_{S_j} and G_{S_j} are the group of the heatmaps from keypoint j and its neighbors, respectively.

[17] proposes an online hard keypoints mining, which first computes the regular L_2 loss for all keypoints, and then additionally punishes top-M hard keypoints. This loss function increases the penalty of the difficult keypoints, and improves the network performance.

[189] presents a combined distillation loss for the HRNet, which consists of a structure loss (STLoss), a pairwise inhibition loss (PairLoss), and a probability distribution loss (PDLoss). Specifically, the STLoss enforces the network to learn human structures at earlier phase to combat against pose occlusions, and the PairLoss alleviates the problem of similar joint misclassification especially in crowded scenarios. The PDLoss guides the learning of the distribution of final heatmaps.

4.4 Domain adaption methods

Human pose estimation has been widely investigated with much focus on supervised learning that requires sufficient pose annotations. However, in real applications, pretrained pose estimation models usually need be adapted to a new domain with no labels or sparse labels. Therefore, several domain adaptation methods [47, 55, 82, 183] leverage a labeled source domain to learn a model that performs well on an unlabeled or sparse labeled target domain.

[183] proposes a domain adaptation method for 2D HPE, which accomplishes both the human body-level topological structure alignment and fine-grained feature alignment in different datasets. Guo et al. [47] proposes a multi-domain pose network that is able to train the model on multiple dataset simultaneously, which obtains a better pose representation in a multi-domain learning fashion. [82] proposes an online coarse-to-fine pseudo label updating strategy to reduce the gap between the synthetic and real data, which have demonstrated strong generalization ability for animal pose estimation. [82] is able to softens the label noises and thereby delivers state-of-the-art results on multiple animal benchmark datasets.

5 Post processing approaches

Instead of predicting the final keypoint locations at once, some approaches first estimate an initial pose and then optimize it with some post-processing operations, which we refer to as *post processing* methods. We divide these methods into two categories, *i.e.*, quantization error and pose resampling. For the heatmap representation of keypoints, the conversion from heatmap to coordinate space inevitably occurs errors, which leads to *quantization errors*. Suppressing such quantization errors will boost the performance of numerous heatmap-based models. On the other hand, an out-of-the-box pose refinement technique, *pose resampling*, aims at resampling favorable pose representations to improve the initial estimations. In what follows, we elaborate on the above approaches.

5.1 Quantization error

The extensively adopted heatmap based pose representation requires decoding the 2D coordinates (x, y) of joints from estimated keypoint heatmaps. In particular, we take the position of the maximum activation value from the predicted heatmap as the keypoint coordinates. However, the predicted gaussian heatmaps do not always conform to the standard gaussian distribution and potentially contain multiple peak values, which degrades the accuracy of the coordinate computation. To address the issue, [193] proposes a distribution-aware architecture that first performs heatmap distribution modulation to adjust the shape of predicted heatmaps and then employs a new coordinate decoding method to accurately obtain the final keypoint locations. This approach reduces mistakes of the conversion from heatmaps to coordinates, and improves the performance of existing heatmap-based models. [58] quantitatively analyzes the common biased data processing on 2D HPE, and further processes data based on unit length instead of pixel, which obtains aligned pose results when flipping is performed in inference. Furthermore, this approach introduces an encoding-decoding method, which is theoretically error-free for the transformation of keypoint locations between heatmaps and coordinates.

On the other hand, the non-differentiable property of the maximum operation in the decoding process also introduces quantization errors. To address this problem, a group of researches [107, 155] attempt to design differentiable algorithms. Luvizon et al. [107] proposes a fully differentiable and end-to-end trainable regression approach, which utilizes the novel Soft-argmax function to convert feature maps directly to keypoint coordinates. Sun et al. [155] proposes an integral method to tackle the problem of non-differentiable from heatmaps to coordinates.

5.2 Pose resampling

A wide spectrum of pose estimators [153, 181] directly take the model output as final estimates. However, these estimations can be further improved by a model-agnostic pose resampling technique. A line of work considers fine-tuning of the initial estimation with additional pose cues. Moon et al. [118] proposes a model-agnostic PoseFix method that estimates a refined pose from a tuple of an input image and an input pose, where the input pose is derived from the estimations of existing methods. Qiu et al. [135] proposes to first localize the visible joints based on visual information by an existing pose estimator, and then estimate the invisible joints by an Image-Guided Progressive GCN module that combines image context and pose structure cues. Wang et al. [170] proposes a two-stage and model-agnostic framework, namely Graph-PCNN, which employs an existing pose estimator for coarse keypoint localization, and designs a graph pose refinement module to produce more accurate localization results.

The above pose resampling methods are designed for static images, and some approaches explore the pose resampling techniques for videos. Specifically, these methods [5, 99, 172, 186, 201] perform pose aggregation to integrate multiple estimated poses of current frame to refine estimations. Normalization is commonly leveraged to aggregate multiple pose predictions [5, 99, 186], where the various predictions are treated equally. [172] introduces the Dijkstra algorithm [23] to solve the problem of optimal keypoint locations, which first employs the mean shift algorithm [21] to group all pose hypotheses into various clusters, and subsequently selects the keypoint with closest distance to the cluster center as the optimal result. [201] utilizes the pose similarity between the neighboring frames and the current frame to biasedly aggregate features, and then employs a convolutional neural network to decode current heatmaps from the aggregated features.

6 Datasets and evaluation

Benchmark datasets form the basis of deep learning models, and also provide a common foundation for measuring and comparing the performance of competing approaches. In this section, we present the major *benchmark datasets*, *evaluation metrics*, and *performance comparisons* for human pose estimation.

Table 2 A summary of 2D human pose estimation benchmark datasets. *Upper Poses*, *Full Poses*, *Various Poses* denotes the upper body poses, singular full body poses and various body poses, respectively

Dataset name	Year	Single-person	Multi-person	Upper poses	Full poses	Various poses	Number of joints	Evaluation metric	Number of images / videos		
									Train	Val	Test
Image-based datasets for human pose estimation											
LSP [71]	2010	✓			✓		14	PCP	1000	–	1000
LSP-extended [72]	2011	✓			✓		14	PCP	10,000	–	–
Flic [142]	2013	✓		✓			10	PCP	5000	–	1016
Flic-full [142]	2013	✓		✓			10	PCP	20,928	–	–
Flic-plus [160]	2013	✓		✓			10	PCP	17,380	–	–
MPII [1]	2014	✓				✓	16	PCPm/PCKh	28,821	–	11,701
MPII [1]	2014		✓			✓	16	PCKh	3800	–	1700
COCO [93]	2017		✓			✓	17	AP	57,000	5000	20,000
AIC-HKD [179]	2017		✓			✓	14	mAP	2,10,000	30,000	60,000
CrowdedPose [84]	2019		✓			✓	14	mAP	10,000	2000	8000
Video-based datasets for human pose estimation											
Penn action [195]	2013	✓			✓		13	mAP	1000	–	1000
JHMDB [65]	2013	✓			✓		15	mAP	600	–	300
PoseTrack2017 [63]	2017		✓			✓	15	mAP	250	50	214
PoseTrack2018 [2]	2018		✓			✓	15	mAP	593	170	375
HiEve [94]	2020		✓			✓	14	mAP	19	–	13

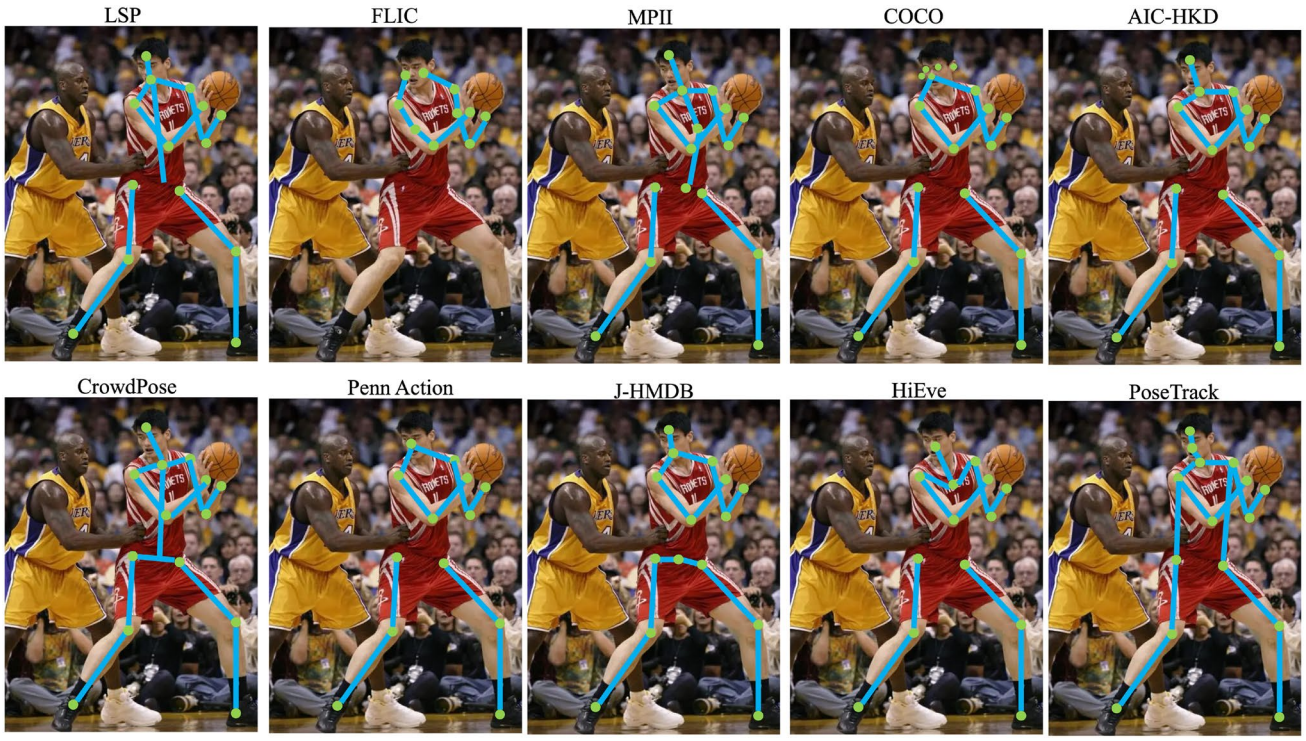

Fig. 5 Illustration of pose annotations for different benchmark datasets including LSP, FLIC, MPII, COCO, AIC-HKD, CrowdedPose, Penn Action, J-HMDB, HiEve, and PoseTrack

6.1 Benchmark datasets

Prior to the flourishing of deep learning, there are plenty of human pose datasets for specific task scenarios, including upper body pose datasets [28–30, 32, 110, 144] and full-body pose dataset [1, 45, 88, 175]. In this section, we investigate the datasets that are commonly used for deep learning, as summarized in Table 2. The corresponding pose annotations are depicted in Fig. 5.

Leeds Sports Pose (LSP) dataset The LSP dataset contains a total number of 2000 images of full body poses (including 14 joints), 1000 images for training and test, respectively. This database is collected from the images tagged *athletics, badminton, baseball, gymnastics, parkour, soccer, tennis, and volleyball* in the Flickr[4]. The LSP dataset is subsequently extended to the LSP-Extended dataset which contains over 10,000 training images. Datasets have been publicly available at https://sam.johnson.io/research/lsp.html.

Frames Labeled in Cinema (FLIC) dataset The FL-IC dataset consists of about 5000 images drawn from popular Hollywood movies, with 4000 images for training and 1000 images for test. During labeling the keypoints, an object detector [7] is first leveraged on the Flic dataset to give the human candidates (roughly 20, 000 examples). These are then sent to the crowdsourcing marketplace *Amazon Mechanical Turk* to obtain the ground truth poses including 10 upper body joints. Severely occluded or non-frontal persons are manually cleaned to form the Flic-Full dataset. These datasets have been publicly available at https://bensapp.github.io/flic-dataset.html.

MPII human pose dataset The MPII dataset contains 28, 821 images for training and 11, 701 images for test. This dataset covers various human activities including recreational, occupational, house holding activities, and involves over 40, 000 individual persons under a wide spectrum of viewpoints. The pose annotations include 15 human joints and occlusion labels. This dataset has been publicly available at http://human-pose.mpi-inf.mpg.de/.

Common objects in context (COCO) dataset Microsoft COCO dataset is one of the most commonly used large-scale vision benchmark datasets, containing a total number of 3, 30, 000 images with over 2, 00, 000 annotated images for vision tasks such as object detection, segmentation, captioning, superpixel stuff segmentation and pose estimation, *etc*. For 2D human pose estimation, 2,00,000 labeled images with 2,50,000 pose annotations are included. Pose annotations with 17 joints on training and validation sets are publicly available, and labels of

[4] Link of Flickr: https://www.flickr.com/.

test set are unavailable. The COCO dataset has become the most popular benchmark in image-based human pose estimation. Therefore, we subsequently report performance comparisons among different algorithms in this dataset. The COCO dataset for 2D human pose estimation can be obtained in https://cocodataset.org/#keypoints-2020.

AI Challenger (AIC) dataset The AIC dataset consists of three sub-datasets: human keypoint detection (HKD), large-scale attribute dataset and image Chinese captioning, respectively. HKD contains 3, 00, 000 images with a total of 7, 00, 000 human instances labeled by 14 keypoints. These images are collected from the Internet search engine with an emphasis on daily activates for ordinary people. The link of official website is: https://challenger.ai/.

CrowdedPose dataset The CrowdedPose dataset is designed for the crowded scenarios, which contains 20, 000 images about 80, 000 individual persons. This dataset has a split ratio of 5 : 1 : 4 for training, validation, and test sets. The dataset is collected by randomly sampling 30,000 images from three public benchmarks according to the *Crowd Index* (a measurement of crowding level for a given image). This dataset is available at https://github.com/Jeff-sjtu/CrowdPose.

Penn action dataset The Penn Action dataset is an unconstrained human action dataset, which contains 2326 video clips derived from YouTuBe, and covers 15 type of actions. There are 1258 videos for training and 1068 videos for test. Each person in images is labeled with 13 keypoints, and both joint coordinates and visibility are provided. This dataset is available at http://dreamdragon.github.io/PennAction/.

Joint-Annotated Human Motion DataBase (JHMDB) dataset JHMDB dataset is a fully annotated dataset for human action recognition and human pose estimation, which contains 21 action categories including *bru-sh hair, catch, clap, climb stairs,* and so on. A subset of JHMDB that involves all visible joints, termed sub-JHMDB, are used for video-based 2D HPE. This subset contains 316 video clips with 12 action categories, and each person is annotated with 15 joints. These datasets are available at http://jhmdb.is.tue.mpg.de/.

PoseTrack dataset PoseTrack is a large-scale public dataset for human pose estimation and articulated tracking, which includes challenging situations with complicated movement of highly occluded people in crowded environments. The PoseTrack2017 dataset contains 514 video clips with 16, 219 pose annotations, and the PoseTrack2018 dataset greatly increased the number of video clips to 1138 with a total of 153, 615 pose annotations. In training videos, dense annotations for 30 center frames of a video are provided. In validation videos, human poses are annotated every four frames. Both datasets label 15 joints, with an additional annotation label for joint visibility. These datasets are available at https://posetrack.net.

Human-centric video analysis in complex events (HiEve) dataset HiEve is the largest dataset for video-based human pose estimation, which contains 31 videos with a total of 10, 99, 357 annotated poses, and labels 14 keypoints. The HiEve dataset incorporates three human-centered understanding tasks, including human pose estimation, pose tracking, and action recognition. The HiEve dataset is publicly available at http://humaninevents.org/.

6.2 Evaluation metrics

Accuracy is the fundamental measurement of performance comparisons between different methods. In Table 2, we list the metrics used to compute the accuracy of models in different datasets. In what follows, we focus on the evaluation metrics of model accuracy.

Percentage of Correctly Estimated Body Parts (PCP) The PCP metric reflects the accuracy of localized body parts. An estimated part is considered correct if its endpoints lie within a threshold, which can be a fraction of the length of the ground truth segment at its annotated location [31]. In addition to the mean PCP of all body parts, separate body limbs PCP such as torso, upper legs and head are also usually reported. Similar to the PCP metric, PCPm utilizes 50% of the mean ground-truth segment length over the entire test as the matching threshold [1].

Percentage of Correct Keypoints (PCK) PCK [185] measures the accuracy of the localized body keypoints, and a candidate joint is considered correct if it lies within a matching threshold. The threshold for matching of the keypoint position to the ground-truth can be defined as a fraction of the human bounding box size (denoted as *PCK*), and 50% of the head segment length (denoted as *PCKh*).

Average Precision (AP) The AP metric is defined on the basis of the Object Keypoint Similarity (OKS) [93] that evaluates the similarity between predicted and ground-truth keypoints. The Average Precision score under different OKS thresholds N is denoted as AP@N. For the image-based human pose estimation, mean average precision (**mAP**) is the mean value of AP scores at all OKS thresholds. In video-based human pose estimation, mAP averages the AP scores of each joint.

6.3 Performance comparisons

In order to comprehensively provide a performance comparison for different human pose estimation algorithms, we pick two representative benchmark datasets: *COCO* and *PoseTrack2017*. The performance of image-level human pose estimation models on COCO dataset are presented in Table 3. HRNet-W48 is a powerful backbone network

Table 3 Performance comparisons of state-of-the-art methods including *top-down* approaches, *bottom-up* approaches and *small networks* on **COCO** benchmark dataset (test-dev2017)

Method	Backbone	Input size	Parameters	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AP^{M}	AP^{L}
Top-down framework: human detection and individual keypoint detection												
Mask-RCNN [54]	ResNet-50	–	–	–	63.1	87.3	68.7	57.8	71.4	–	–	–
G-RMI [128]	ResNet-101	353 × 257	42.6M	57.0	64.9	85.5	71.3	62.3	70.0	69.7	–	–
Integral Pose [155]	ResNet-101	256 × 256	45.0M	11.0	67.8	88.2	74.8	63.9	74.0	–	–	–
G-RMI+extra data [128]	ResNet-101	353 × 257	42.6M	57.0	68.5	87.1	75.5	65.8	73.3	73.3	–	–
CPN [17]	Resnet-Inception	384 × 288	–	–	72.1	91.4	80.0	68.7	77.2	78.5	–	–
RMPE [34]	Stacked Hourglass	320 × 256	28.1M	26.7	72.3	89.2	79.1	68.0	78.6	–	–	–
CFN [60]	–	–	–	–	72.6	86.1	69.7	78.3	64.1	–	–	–
CPN (ensemble) [17]	Resnet-inception	384 × 288	–	–	73.0	91.7	80.9	69.5	78.1	79.0	–	–
SimpleBaseline [181]	ResNet-152	384 × 288	68.6M	35.6	73.7	91.9	81.1	70.3	80.0	79.0	–	–
HRNet-W32 [153]	HRNet-W32	384 × 288	28.5M	16.0	74.9	92.5	82.8	71.3	80.9	80.1	–	–
HRNet-W48 [153]	HRNet-W48	384 × 288	63.6M	32.9	75.5	92.5	83.3	71.9	81.5	80.5	–	–
DARK [193]	HRNet-W48	384 × 288	63.6M	32.9	76.2	92.5	83.6	72.5	82.4	81.1	–	–
UDP [58]	HRNet-W48	384 × 288	63.6M	33.0	**76.5**	92.7	84.0	73.0	82.4	81.6	–	–
HRNet-W48+extra data [153]	HRNet-W48	384 × 288	63.6M	32.9	77.0	92.7	84.5	73.4	83.1	82.0	–	–
DARK+extra data [193]	HRNet-W48	384 × 288	63.6M	32.9	**77.4**	92.6	84.6	73.6	83.7	82.3	–	–
Bottom-up framework: keypoint detection and grouping												
AE [121]	–	512	–	–	63.0	85.7	68.9	58.0	70.4	–	–	–
AE+refinement [121]	–	512	–	–	65.5	86.8	72.3	60.6	72.6	70.2	64.6	78.1
DirectPose [159]	–	800	–	–	64.8	87.8	71.1	60.4	71.5	–	–	–
SimplePose [85]	–	512	–	–	68.1	–	–	66.8	70.5	72.1	–	–
HGG [70]	–	512	–	–	67.6	85.1	73.7	62.7	74.6	71.3	–	–
PersonLab [129]	–	1401	–	–	68.7	89.0	75.4	64.1	75.5	75.4	69.7	83.0
Point-set anchors [177]	–	640	–	–	68.7	89.9	76.3	64.8	75.3	74.8	69.6	82.1
HrHRNet-W48+AE [19]	HRNet-W48	640	–	–	70.5	89.3	77.2	66.6	75.8	–	–	–
DEKR-W48 [40]	HRNet-W48	640	–	–	71.0	89.2	78.0	67.1	76.9	76.7	71.5	83.9
SWAHR+HrHRNet-W48 [106]	HRNet-W48	–	–	–	**72.0**	90.7	78.8	67.8	77.7	–	–	–
Small networks												
Small HRNet [187]	HRNet-W16	384 × 288	1.3M	1.21	55.2	85.8	61.4	51.7	61.2	61.5	–	–
MobileNetV2 1× [141]	MobileNetV2	384 × 288	9.8M	3.33	66.8	90.0	74.0	62.6	73.3	72.3	–	–
ShuffleNetV2 1× [108]	ShuffleNetV2	384 × 288	7.6M	2.87	62.9	88.5	69.4	58.9	69.3	68.9	–	–
Lite-HRNet [187]	Lite-HRNet-30	384 × 288	1.8M	0.70	**69.7**	90.7	77.5	66.9	75.0	75.4	–	–

with excellent performance for keypoint localiztion, and UDP network that builds upon the HRNet achieves state-of-the-art results without extra training data. The bottom-up approaches remain a wide gap (5.4 mAP) compared to the top-down approaches. In additional to the model accuracy, the efficiency is also important especially for practical applications. To this end, we report some approaches that aim at designing small networks, as summarized in Table 3. Lite-HRNet achieves a better trade-off between accuracy and speed, which obtains the accuracy of 69.7 mAP with parameters of 1.8 M.

We also report the performance of various video-based models on PoseTrack2017 dataset in Table 4. The DCPose employs abundant temporal information from adjacent frame to facilitate the current pose estimation, consistently establishing new state-of-the-arts on both validation and test sets.

7 Discussion

In this section, we first discuss the open questions of the current 2D human pose estimation, including model generalization and datasets. Subsequently, we introduce the incompletely explored domain of estimating human pose from signal data. Finally, we provide future research directions in terms of unsupervised learning, pose representations, and model explainability.

Table 4 Performance comparisons of state-of-the-art methods on **PoseTrack2017** benchmark dataset (validation and test sets). Pretrain denotes the backbone model has been pretrained on COCO keypoint detection dataset

Method	Backbone	Pretrain	Additional training data	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Mean**
Dataset: **PoseTrack2017 validation** set.											
PoseTracker [41]	ResNet-3D	Y	COCO	67.5	70.2	62.0	51.7	60.7	58.7	49.8	60.6
PoseFlow [182]	–	–	MPII Pose + COCO	66.7	73.3	68.3	61.1	67.5	67.0	61.3	66.5
JointFlow [24]	–	–	–	–	–	–	–	–	–	–	69.3
FastPose [194]	–	–	–	80.0	80.3	69.5	59.1	71.4	67.5	59.4	70.3
SimpleBaseline [181]	ResNet-50	N	COCO	79.1	80.5	75.5	66.0	70.8	70.0	61.7	72.4
SimpleBaseline [181]	ResNet-152	N	COCO	81.7	83.4	80.0	72.4	75.3	74.8	67.1	76.7
STEmbedding [69]	4-Stage stacked hourglass	Y	–	83.8	81.6	77.1	70.0	77.4	74.5	70.8	77.0
HRNet [153]	HRNet-W48	Y	COCO	82.1	83.6	80.4	73.3	75.5	75.3	68.5	77.3
MDPN [47]	SimpleBaseline	Y	MPII Pose + COCO	85.2	88.5	83.9	77.5	79.0	77.0	71.4	80.7
Dynamic [186]	HRNet-W48	Y	COCO	88.4	88.4	82.0	74.5	79.1	78.3	73.1	81.1
PoseWarper [5]	HRNet-W48	Y	COCO	81.4	88.3	83.9	78.0	82.4	80.5	73.6	81.2
DCPose [99]	HRNet-W48	Y	COCO	**88.0**	**88.7**	**84.1**	**78.4**	**83.0**	**81.4**	**74.2**	**82.8**
Dataset: **PoseTrack2017 Test** set (Results from the PoseTrack official leaderboard)											
PoseTracker [41]	ResNet-3D	Y	COCO	–	–	–	51.5	–	–	50.17	59.6
PoseFlow [182]	–	–	MPII Pose + COCO	64.9	67.5	65.0	59.0	62.5	62.8	57.9	63.0
JointFlow [24]	–	–	–	–	–	–	53.1	–	–	50.4	63.4
KeyTrack [150]	–	–	COCO	–	–	–	71.9	–	–	65.0	74.0
DetTrack [172]	3D-HRNet	Y	COCO	–	–	–	69.8	–	–	65.9	74.1
SimpleBaseline [181]	ResNet-152	N	COCO	80.1	80.2	76.9	71.5	72.5	72.4	65.7	74.6
HRNet [153]	HRNet-W48	Y	COCO	80.1	80.2	76.9	72.0	73.4	72.5	67.0	74.9
PoseWarper [5]	HRNet-W48	Y	COCO	79.5	84.3	80.1	75.8	77.6	76.8	70.8	77.9
DCPose [99]	HRNet-W48	Y	COCO	**84.3**	**84.9**	**80.5**	**76.1**	**77.9**	**77**.1	**71.2**	**79.2**

7.1 Open questions

Human pose estimation has been greatly advanced by the deep learning. However, there are still numerous challenges that prevent models from achieving perfect performance. Such challenges mainly arise from two aspects: question of models and shortcoming of datasets.

Model capacity Regarding both *image-based* and *video-based* human pose estimation, modern deep models have difficulties in tackling pose occlusions, person entanglement, and motion blur in complex scenarios. In such cases, the absence of keypoint visual feature leads to difficulties in localizing joints according to visual information. For image-based human pose estimation, models require the prior knowledge of human structure to cope with the lack of visual cues in static images. In terms of video-based human pose estimation, the models need to fully use temporal cues to recover human poses from the frames with insufficient visual information. Additional cues from adjacent frames can be employed to reconstruct the pose of the current frame.

Training data shortage Large-scale annotated image datasets are currently available, yet video datasets still suffer from some shortcomings such as singular scenes and insufficient quantity. On the other hand, the high-quality position labels of occluded joints are missing in the video dataset. Most of existing video datasets only label the joint visibility to indicate that whether a joint is occluded. In this configuration, the models are hard to learn to detect the occluded or entangled joints, which greatly increases the difficulty in handling pose occlusions.

In addition, lacking of domain-specific datasets is also a shortcoming. For particular scenes such as dancing and swimming, datasets of the corresponding domains are necessary for the practical application. Therefore, building specialized datasets for various domains is essential to facilitate the application of 2D HPE.

7.2 Signal-based human pose estimation

The corruption of visual features leads to challenges in handling *hard* joints, and non-visual data such as *WIFI signals* provides another way to overcome this problem. Previous works [48, 83, 165, 166, 199] propose to recover human poses from the radio signals or radar. Zhao et al. [199] leverages WIFI signals to traverse walls and reflect off the human body for accurately estimating human pose when the person is occluded by the wall. Specifically, a deep neural network is proposed to parse keypoint locations from WiFi signals. Wang et al. [166] presents a WiFi antennas-based method which takes the WiFi signals as input, and performs pose estimation in an end-to-end fashion. Li et al. [83] proposes a human pose estimation system using $77GHz$ millimeter wave radar, which first employs two radar data to generate heatmaps, and then employs a CNN to transform two-dimensional heatmaps into human poses.

7.3 Future directions

We expect that future researches would dive deeper into three aspects: unsupervised learning, pose representation, and model interpretability.

Unsupervised learning The fully-supervised methods currently dominate the field of human pose estimation since their superior performance. Their success stems from the rich pose annotations in large-scale datasets. However, unlabeled images and videos are an almost endless source, and providing full annotations for these data is impossible. Therefore, unsupervised learning that can automatically learn knowledge of human body from an infinite amount of data has been an important direction.

Pose representation The heatmap-based pose representation has demonstrated superior performance. However, quantization errors in encoding heatmap from coordinates and decoding coordinates from heatmaps are inevitable. Simultaneously, the encoding and decoding processes of the heatmap are influenced by its resolution. The high resolution brings good accuracy, but also increases the computational load. Therefore, a novel unbiased pose representation for addressing such issues is necessary.

Model Explainability A drawback of deep learning methods is uninterpretability. So far, there is no comprehensive and formal theory for interpretability. As a result, there is limited systematic guidance in designing the deep learning models. With respect to human pose estimation, we also fail to clearly understand how the visual features of the input image impact the final keypoint localization, which is detrimental to future investigations. Given the potential shortcoming, it is highly desirable to advance works on the interpretability of human pose estimation models.

8 Conclusion

In this paper, we present a comprehensive and systematic review of human pose estimation methods. We present a coarse-level taxonomy with three categories: *network architecture design*, *network training refinement*, and *post processing*. The network architecture design methods focus on the model architecture, the network training refinement methods revolve around the training of networks, and the post processing methods consider the model-agnostic optimization strategies. On a finer level, we split the *network architecture design* methods (Section 3) into top-down framework and bottom-up framework. We divide the *network training refinement* approaches (Section 4) into data augmentation techniques, multi-task learning strategies, loss

function constraints, and domain adaption methods. The *post processing* methods (Section 5) consists of quantization error and pose resampling. Ultimately, we summarize popular benchmark datasets and evaluation metrics, conduct model performance comparisons, and discuss the potential future research directions. Hope this would be beneficial for researchers in the community and would inspire future research.

Acknowledgements This paper is supported by the National Key R&D Program of China (Grant no.2018YFB1404102), the Key R&D Program of Zhejiang Province (No. 2021C01104), and the National Natural Science Foundation of China (No. 61902348).

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, pp. 3686–3693 (2014)
2. Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., Schiele, B.: Posetrack: A benchmark for human pose estimation and tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5167–5176 (2018)
3. Artacho, B., Savakis, A.: Unipose: Unified human pose estimation in single images and videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7035–7044 (2020)
4. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A.: Sequential deep learning for human action recognition. In: International workshop on human behavior understanding, Springer, pp. 29–39 (2011)
5. Bertasius, G., Feichtenhofer, C., Tran, D., Shi, J., Torresani, L.: Learning temporal pose estimation from sparsely-labeled videos. In: Advances in Neural Information Processing Systems, pp. 3027–3038 (2019)
6. Bin, Y., Cao, X., Chen, X., Ge, Y., Tai, Y., Wang, C., Li, J., Huang, F., Gao, C., Sang, N.: Adversarial semantic data augmentation for human pose estimation. In: European Conference on Computer Vision, Springer, pp. 606–622 (2020)
7. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: 2009 IEEE 12th International Conference on Computer Vision, IEEE, pp. 1365–1372 (2009)
8. Cai, Y., Wang, Z., Luo, Z., Yin, B., Du, A., Wang, H., Zhou, X., Zhou, E., Zhang, X., Sun, J.: Learning delicate local representations for multi-person pose estimation. arXiv preprint: arXiv:2003.04030 (2020)
9. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017a)
10. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: 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 (2017b)
11. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, Springer, pp. 213–229 (2020)
12. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4733–4742 (2016)
13. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5933–5942 (2019)
14. Chang, S., Yuan, L., Nie, X., Huang, Z., Zhou, Y., Chen, Y., Feng, J., 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)
15. Charles, J., Pfister, T., Magee, D., Hogg, D., Zisserman, A.: Personalizing human video pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3063–3072 (2016)
16. Chen, C.H., Ramanan, D.: 3d human pose estimation= 2d pose estimation+ matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7035–7043 (2017)
17. 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)
18. Chen, Y., Tian, Y., He, M.: Monocular human pose estimation: A survey of deep learning-based methods. Comput. Vis. Image Underst. **192**, (2020)
19. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5386–5395 (2020)
20. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1831–1840 (2017)
21. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 603–619 (2002)
22. Datta, S., Sikka, K., Roy, A., Ahuja, K., Parikh, D., Divakaran, A.: Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
23. Dijkstra, E.W., et al.: A note on two problems in connexion with graphs. Numer. Math. **1**(1), 269–271 (1959)
24. Doering, A., Iqbal, U., Gall, J.: Joint flow: Temporal flow fields for multi person tracking. arXiv preprint: arXiv:1805.04596 (2018)
25. Dong, J., Chen, Q., Shen, X., Yang, J., Yan, S.: Towards unified human parsing and pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 843–850 (2014)
26. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
27. Duan, H., Lin, K.Y., Jin, S., Liu, W., Qian, C., Ouyang, W.: Trb: a novel triplet representation for understanding 2d human body. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9479–9488 (2019)
28. Eichner, M., Ferrarim, V.: We are family: Joint pose estimation of multiple persons. In: European conference on computer vision, Springer, pp. 228–242 (2010)
29. Eichner, M., Ferrari, V.: Human pose co-estimation and applications. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2282–2288 (2012)
30. Eichner, M., Ferrari, V., Zurich, S.: Better appearance models for pictorial structures. In: Bmvc, Citeseer, vol 2, p 5 (2009)

31. Eichner, M., Marin-Jimenez, M., Zisserman, A., Ferrari, V.: 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. Int. J. Comput. Vis. **99**(2), 190–214 (2012)
32. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
33. Fan, X., Zheng, K., Lin, Y., Wang, S.: Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1347–1355 (2015)
34. 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)
35. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. Int. J. Comput. Vis. **61**(1), 55–79 (2005)
36. Fieraru M, Khoreva A, Pishchulin L, Schiele B (2018) Learning to refine human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 205–214
37. Gao, Y., Chang, H.J., Demiris, Y.: User modelling for personalised dressing assistance by humanoid robots. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp. 1840–1845 (2015)
38. Gao, Y., Chang, H.J., Demiris, Y.: Iterative path optimisation for personalised dressing assistance using vision and force information. In: 2016 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE, pp. 4398–4403 (2016)
39. Garau, N., Bisagno, N., Bródka, P., Conci, N.: Deca: Deep viewpoint-equivariant human pose estimation using capsule autoencoders. arXiv preprint: arXiv:2108.08557 (2021)
40. Geng, Z., Sun, K., Xiao, B., Zhang, Z., Wang, J.: Bottom-up human pose estimation via disentangled keypoint regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14676–14686 (2021)
41. Girdhar, R., Gkioxari, G., Torresani, L., Paluri, M., Tran, D.: Detect-and-track: Efficient pose estimation in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 350–359 (2018)
42. Gkioxari, G., Toshev, A., Jaitly, N.: Chained predictions using convolutional neural networks. In: European Conference on Computer Vision, Springer, pp. 728–743 (2016)
43. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 932–940 (2017)
44. Gong, K., Liang, X., Li, Y., Chen, Y., Yang, M., Lin, L.: Instance-level human parsing via part grouping network. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 770–785 (2018)
45. Gong, W., Zhang, X., Gonzàlez, J., Sobral, A., Bouwmans, T., Tu, C., Zahzah, E.H.: Human pose estimation from monocular images: a comprehensive survey. Sensors **16**(12), 1966 (2016)
46. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Adv. Neural Inf. Process Syst. 27 (2014)
47. Guo, H., Tang, T., Luo, G., Chen, R., Lu, Y., Wen, L.: Multi-domain pose network for multi-person pose estimation and tracking. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 0–0 (2018)
48. Guo, L., Lu, Z., Wen, X., Zhou, S., Han, Z.: From signal to image: Capturing fine-grained human poses with commodity wi-fi. IEEE Commun. Lett. **24**(4), 802–806 (2019)
49. Guo, Y., Cheng, Z., Nie, L., Liu, Y., Wang, Y., Kankanhalli, M.S.: Quantifying and alleviating the language prior problem in visual question answering. In: SIGIR, ACM, pp. 75–84 (2019b)
50. Guo, Y., Nie, L., Cheng, Z., Ji, F., Zhang, J., Bimbo, A.D.: Adavqa: Overcoming language priors with adapted margin cosine loss. In: IJCAI, ijcai.org, pp. 708–714 (2021a)
51. Guo, Y., Nie, L., Cheng, Z., Ji, F., Zhang, J., Del Bimbo, A.: Adavqa: Overcoming language priors with adapted margin cosine loss. arXiv preprint: arXiv:2105.01993 (2021b)
52. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
53. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017a)
54. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969 (2017b)
55. Hidalgo, G., Raaj, Y., Idrees, H., Xiang, D., Joo, H., Simon, T., Sheikh, Y.: Single-network whole-body pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6982–6991 (2019)
56. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint: arXiv:1503.02531 (2015)
57. Holte, M.B., Tran, C., Trivedi, M.M., Moeslund, T.B.: Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. IEEE J. Select. Topic Signal Proces **6**(5), 538–552 (2012)
58. Huang, J., Zhu, Z., Guo, F., Huang, G.: The devil is in the details: Delving into unbiased data processing for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5700–5709 (2020a)
59. Huang, J., Zhu, Z., Huang, G., Du, D.: Aid: Pushing the performance boundary of human pose estimation with information dropping augmentation. arXiv preprint: arXiv:2008.07139
60. Huang, S., Gong, M., Tao, D.: A coarse-fine network for keypoint localization. In: Proceedings of the IEEE international conference on computer vision, pp. 3028–3037 (2017)
61. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
62. 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, Springer, pp. 34–50 (2016)
63. Iqbal, U., Garbade, M., Gall, J.: Pose for action-action for pose. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE, pp. 438–445 (2017)
64. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Adv. Neural Inf. Process System **28**, 2017–2025 (2015)
65. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: Proceedings of the IEEE international conference on computer vision, pp. 3192–3199 (2013)
66. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2012)
67. Ji, X., Liu, H.: Advances in view-invariant human motion analysis: a review. IEEE Trans. Syst. Man Cybern. **40**(1), 13–24 (2009)
68. Jiang, C., Huang, K., Zhang, S., Wang, X., Xiao, J.: Pay attention selectively and comprehensively: Pyramid gating network for human pose estimation without pre-training. In: Proceedings

of the 28th ACM International Conference on Multimedia, pp. 2364–2371 (2020)
69. Jin, S., Liu, W., Ouyang, W., Qian, C.: Multi-person articulated tracking with spatial and temporal embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5664–5673 (2019)
70. Jin, S., Liu, W., Xie, E., Wang, W., Qian, C., Ouyang, W., Luo, P.: Differentiable hierarchical graph grouping for multi-person pose estimation. In: European Conference on Computer Vision, Springer, pp. 718–734 (2020)
71. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: bmvc, Citeseer, vol 2, p 5 (2010)
72. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: CVPR 2011, IEEE, pp. 1465–1472 (2011)
73. Ju, S.X., Black, M.J., Yacoob, Y.: Cardboard people: A parameterized model of articulated image motion. In: Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, IEEE, pp. 38–44 (1996)
74. Kappel, M., Golyanik, V., Elgharib, M., Henningson, J.O., Seidel, H.P., Castillo, S., Theobalt, C., Magnor, M.: High-fidelity neural human motion transfer from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1541–1550 (2021)
75. Ke, L., Chang, M.C., Qi, H., Lyu, S.: Multi-scale structure-aware network for human pose estimation. In: Proceedings of the european conference on computer vision (ECCV), pp. 713–728 (2018)
76. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint: arXiv:1609.02907 (2016)
77. 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)
78. Kreiss, S., Bertoni, L., Alahi, A.: Pifpaf: Composite fields for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11977–11986 (2019)
79. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process System **25**, 1097–1105 (2012)
80. Ladicky, L., Torr, P.H., Zisserman, A.: Human pose estimation using a joint pixel-wise and part-wise formulation. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3578–3585 (2013)
81. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
82. Li, C., Lee, G.H.: From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1482–1491 (2021)
83. Li, G., Zhang, Z., Yang, H., Pan, J., Chen, D., Zhang, J.: Capturing human pose using mmwave radar. In: 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), IEEE, pp. 1–6 (2020a)
84. 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)
85. Li, J., Su, W., Wang, Z.: Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp. 11354–11361 (2020b)
86. Li, J., Bian, S., Zeng, A., Wang, C., Pang, B., Liu, W., Lu, C.: Human pose regression with residual log-likelihood estimation. arXiv preprint arXiv:2107.11291 (2021a)
87. Li, K., Wang, S., Zhang, X., Xu, Y., Xu, W., Tu, Z.: Pose recognition with cascade transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1944–1953 (2021b)
88. Li, L.J., Fei-Fei, L.: What, where and who? classifying events by scene and object recognition. In: 2007 IEEE 11th international conference on computer vision, IEEE, pp. 1–8 (2007)
89. Li, S., Liu, Z.Q., Chan, A.B.: Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 482–489 (2014)
90. Li, Y., Yang, X., Shang, X., Chua, T.S.: Interventional video relation detection. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 4091–4099 (2021c)
91. Li, Z., Ye, J., Song, M., Huang, Y., Pan, Z.: Online knowledge distillation for efficient pose estimation. arXiv preprint arXiv:2108.02092 (2021d)
92. Liang, X., Gong, K., Shen, X., Lin, L.: Look into person: Joint body parsing & pose estimation network and a new benchmark. IEEE Trans. pattern Anal. Mach. Intell. **41**(4), 871–885 (2018)
93. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, Springer, pp. 740–755 (2014)
94. Lin, W., Liu, H., Liu, S., Li, Y., Qian, R., Wang, T., Xu, N., Xiong, H., Qi, G.J., Sebe, N.: Human in events: A large-scale benchmark for human-centric video analysis in complex events. arXiv preprint arXiv:2005.04490 (2020)
95. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision, Springer, pp. 21–37 (2016)
96. Liu, W., Chen, J., Li, C., Qian, C., Chu, X., Hu, X.: A cascaded inception of inception network with attention modulated feature fusion for human pose estimation. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
97. Liu, Z., Zhu, J., Bu, J., Chen, C.: A survey of human pose estimation: the body parts parsing based methods. J. Vis. Commun. Image Represent **32**, 10–19 (2015)
98. Liu, Z., Wu, S., Jin, S., Liu, Q., Lu, S., Zimmermann, R., Cheng, L.: Towards natural and accurate future motion prediction of humans and animals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10004–10012 (2019)
99. Liu, Z., Chen, H., Feng, R., Wu, S., Ji, S., Yang, B., Wang, X.: Deep dual consecutive network for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 525–534 (2021a)
100. Liu, Z., Lyu, K., Wu, S., Chen, H., Hao, Y., Ji, S.: Aggregated multi-gans for controlled 3d human motion prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 35, pp. 2225–2232 (2021b)
101. Liu, Z., Qian, P., Wang, X., Zhuang, Y., Qiu, L., Wang, X.: Combining graph neural networks with expert knowledge for smart contract vulnerability detection. IEEE Transactions on Knowledge and Data Engineering (2021c)
102. Liu, Z., Su, P., Wu, S., Shen, X., Chen, H., Hao, Y., Wang, M.: Motion prediction using trajectory cues. IEEE International Conference on Computer Vision (2021d)
103. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. arXiv preprint arXiv:2201.03545 (2022)
104. Luo, Y., Ren, J., Wang, Z., Sun, W., Pan, J., Liu, J., Pang, J., Lin, L.: Lstm pose machines. In: Proceedings of the IEEE

conference on computer vision and pattern recognition, pp. 5207–5215 (2018a)
105. Luo, Y., Xu, Z., Liu, P., Du, Y., Guo, J.M.: Multi-person pose estimation via multi-layer fractal network and joints kinship pattern. IEEE Trans. Image Process **28**(1), 142–155 (2018)
106. Luo, Z., Wang, Z., Huang, Y., Wang, L., Tan, T., Zhou, E.: 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)
107. Luvizon, D.C., Tabia, H., Picard, D.: Human pose regression by combining indirect part detection and contextual information. Comput. Graphic **85**, 15–22 (2019)
108. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Proceedings of the European conference on computer vision (ECCV), pp. 116–131 (2018)
109. Mao, W., Tian, Z., Wang, X., Shen, C.: Fcpose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9034–9043 (2021)
110. Marin-Jimenez, M.J., Zisserman, A., Eichner, M., Ferrari, V.: Detecting people looking at each other in videos. Int. J. Comput Vis. **106**(3), 282–296 (2014)
111. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2640–2649 (2017)
112. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. ACM Trans Graphic (TOG) **36**(4), 1–14 (2017)
113. Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., Ghasemzadeh, H.: Improved knowledge distillation via teacher assistant. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 34, pp. 5191–5198 (2020)
114. Moeslund, T.B., Granum, E.: A survey of computer vision-based human motion capture. Comput. Vis. Image Understand **81**(3), 231–268 (2001)
115. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Comput. Vis. image Understand. **104**(2–3), 90–126 (2006)
116. Moeslund, T.B., Hilton, A., Krüger, V., Sigal, L.: Visual analysis of humans. Springer, NY (2011)
117. Mogadala, A., Kalimuthu, M., Klakow, D.: Trends in integration of vision and language research: A survey of tasks, datasets, and methods. J. Artif. Intell. Res. (2021)
118. Moon, G., Chang, J.Y., Lee, K.M.: Posefix: Model-agnostic general human pose refinement network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7773–7781(2019)
119. 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)
120. Naksuk, N., Lee, C.G., Rietdyk, S.: Whole-body human-to-humanoid motion transfer. In: 5th IEEE-RAS International Conference on Humanoid Robots, 2005., IEEE, pp. 104–109 (2005)
121. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. arXiv preprint arXiv:1611.05424 (2016a)
122. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision, Springer, pp. 483–499 (2016b)
123. Nie, X., Feng, J., Xing, J., Yan, S.: Pose partition networks for multi-person pose estimation. In: Proceedings of the european conference on computer vision (eccv), pp. 684–699 (2018a)
124. Nie, X., Feng, J., Zuo, Y., Yan, S.: Human pose estimation with parsing induced learner. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2100–2108 (2018b)
125. Nie, X., Feng, J., Zhang, J., Yan, S.: Single-stage multi-person pose machines. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6951–6960 (2019a)
126. Nie X, Li Y, Luo L, Zhang N, Feng J (2019b) Dynamic kernel distillation for efficient pose estimation in videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6942–6950
127. Nie, X., Feng, J., Zhang, J., Yan, S.: Single-stage multi-person pose machines. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2020)
128. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: 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)
129. Papandreou, G., Zhu, T., Chen, L.C., Gidaris, S., Tompson, J., Murphy, K.: Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 269–286 (2018)
130. Peng, X., Tang, Z., Yang, F., Feris, R.S., Metaxas, D.: Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2226–2234(2018)
131. Pfister, T., Charles, J., Zisserman, A.: Flowing convnets for human pose estimation in videos. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1913–1921 (2015)
132. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Poselet conditioned pictorial structures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 588–595 (2013)
133. 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)
134. Poppe, R.: Vision-based human motion analysis: An overview. Comput. Vis. Image Understand **108**(1–2), 4–18 (2007)
135. Qiu, L., Zhang, X., Li, Y., Li, G., Wu, X., Xiong, Z., Han, X., Cui, S.: Peeking into occluded joints: A novel framework for crowd pose estimation. In: European Conference on Computer Vision, Springer, pp. 488–504 (2020)
136. Raaj, Y., Idrees, H., Hidalgo, G., Sheikh, Y.: Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4620–4628 (2019)
137. Ramakrishna, V., Munoz, D., Hebert, M., Bagnell, J.A., Sheikh, Y.: Pose machines: Articulated pose estimation via inference machines. In: European Conference on Computer Vision, Springer, pp. 33–47 (2014)
138. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process Syst. **28**, 91–99 (2015)
139. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)

140. Ruan, T., Liu, T., Huang, Z., Wei, Y., Wei, S., Zhao, Y.: Devil in the details: Towards accurate single and multiple human parsing. In: Proc. AAAI Conf. Artif. Intell. 33, 4814–4821 (2019)
141. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510–4520 (2018)
142. Sapp, B., Taskar, B.: Modec: Multimodal decomposable models for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3674–3681 (2013)
143. Sapp, B., Toshev, A., Taskar, B.: Cascaded models for articulated pose estimation. In: European conference on computer vision, Springer, pp. 406–420 (2010)
144. Sapp, B., Weiss, D., Taskar, B.: Parsing human motion with stretchable models. In: CVPR 2011, IEEE, pp. 1281–1288 (2011)
145. Sarafianos, N., Boteanu, B., Ionescu, B., Kakadiaris, I.A.: 3d human pose estimation: A review of the literature and analysis of covariates. Comput. Vis. Image Understand **152**, 1–20 (2016)
146. Schmidtke, L., Vlontzos, A., Ellershaw, S., Lukens, A., Arichi, T., Kainz, B.: Unsupervised human pose estimation through transforming shape templates. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2484–2494 (2021)
147. Shang, X., Di, D., Xiao, J., Cao, Y., Yang, X., Chua, T.S.: Annotating objects and relations in user-generated videos. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval, pp. 279–287 (2019)
148. Sidenbladh, H., De la Torre, F., Black, M.J.: A framework for modeling the appearance of 3d articulated figures. In: Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), IEEE, pp. 368–375 (2000)
149. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
150. Snower, M., Kadav, A., Lai, F., Graf, H.P.: 15 keypoints is all you need. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6738–6748 (2020)
151. Song, J., Wang, L., Van Gool, L., Hilliges, O.: Thin-slicing network: A deep structured model for pose estimation in videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4220–4229 (2017)
152. Su, K., Yu, D., Xu, Z., Geng, X., Wang, C.: Multi-person pose estimation with enhanced channel-wise and spatial information. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5674–5682 (2019)
153. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5693–5703 (2019)
154. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2602–2611 (2017)
155. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 529–545 (2018)
156. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9 (2015)
157. Tang, W., Wu, Y.: Does learning specific features for related parts help human pose estimation? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1107–1116 (2019)
158. 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)
159. Tian, Z., Chen, H., Shen, C.: Directpose: Direct end-to-end multi-person pose estimation. arXiv preprint arXiv:1911.07451 (2019)
160. 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)
161. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
162. Varamesh, A., Tuytelaars, T.: Mixture dense regression for object detection and human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13086–13095 (2020)
163. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems, pp. 5998–6008 (2017)
164. Wang, F., Li, Y.: Beyond physical connections: Tree models in human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 596–603 (2013)
165. Wang, F., Panev, S., Dai, Z., Han, J., Huang, D.: Can wifi estimate person pose? arXiv preprint arXiv:1904.00277 (2019a)
166. Wang, F., Zhou, S., Panev, S., Han, J., Huang, D.: Person-in-wifi: Fine-grained person perception using wifi. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5452–5461 (2019b)
167. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision, pp. 3551–3558 (2013)
168. Wang, J., Gou, L., Zhang, W., Yang, H., Shen, H.W.: Deepvid: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation. IEEE Trans. Visual. Comput. Graphic **25**(6), 2168–2180 (2019)
169. Wang, J., Qiu, K., Peng, H., Fu, J., Zhu, J.: Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 374–382 (2019d)
170. Wang, J., Long, X., Gao, Y., Ding, E., Wen, S.: Graph-pcnn: Two stage human pose estimation with graph pose refinement. In: European Conference on Computer Vision, Springer, pp. 492–508 (2020a)
171. Wang, J., Jin, S., Liu, W., Liu, W., Qian, C., Luo, P.: When human pose estimation meets robustness: Adversarial algorithms and benchmarks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11855–11864 (2021)
172. Wang, M., Tighe, J., Modolo, D.: Combining detection and tracking for human pose estimation in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11088–11096 (2020b)
173. Wang, X., Gao, L., Song, J., Shen, H.T.: Ktn: Knowledge transfer network for multi-person densepose estimation. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 3780–3788 (2020c)
174. Wang, Y., Mori, G.: Multiple tree models for occlusion and spatial constraints in human pose estimation. In: European Conference on Computer Vision, Springer, pp. 710–724 (2008)
175. Wang, Y., Tran, D., Liao, Z.: Learning hierarchical poselets for human parsing. In: CVPR 2011, IEEE, pp. 1705–1712 (2011)
176. Wehrbein, T., Rudolph, M., Rosenhahn, B., Wandt, B.: Probabilistic monocular 3d human pose estimation with normalizing flows. arXiv preprint arXiv:2107.13788 (2021)

177. Wei, F., Sun, X., Li, H., Wang, J., Lin, S.: Point-set anchors for object detection, instance segmentation and pose estimation. In: European Conference on Computer Vision, Springer, pp. 527–544 (2020)
178. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
179. Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., Wang, W., Zhou, S., Lin, G., Fu, Y., et al.: Ai challenger: A large-scale dataset for going deeper in image understanding. arXiv preprint arXiv:1711.06475 (2017)
180. Xia, F., Wang, P., Chen, X., Yuille, A.L.: Joint multi-person pose estimation and semantic part segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6769–6778 (2017)
181. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European conference on computer vision (ECCV), pp. 466–481(2018)
182. Xiu, Y., Li, J., Wang, H., Fang, Y., Lu, C.: Pose flow: Efficient online pose tracking. arXiv preprint arXiv:1802.00977 (2018)
183. Xu, X., Zou, Q., Lin, X.: Alleviating human-level shift: A robust domain adaptation method for multi-person pose estimation. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2326–2335 (2020)
184. Yang, W., Li, S., Ouyang, W., Li, H., Wang, X.: Learning feature pyramids for human pose estimation. In: proceedings of the IEEE international conference on computer vision, pp. 1281–1290 (2017)
185. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2878–2890 (2012)
186. Yang, Y., Ren, Z., Li, H., Zhou, C., Wang, X., Hua, G.: Learning dynamics via graph neural networks for human pose estimation and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8074–8084 (2021)
187. Yu, C., Xiao, B., Gao, C., Yuan, L., Zhang, L., Sang, N., Wang, J.: Lite-hrnet: A lightweight high-resolution network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10440–10450 (2021)
188. Yu, D., Su, K., Sun, J., Wang, C.: Multi-person pose estimation for pose tracking with enhanced cascaded pyramid network. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 0–0 (2018)
189. Yuan, L., Zhang, S., Fubiao, F., Wei, N., Pan, H.: Combined distillation pose. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 4635–4639 (2020)
190. Zeng, A., Sun, X., Yang, L., Zhao, N., Liu, M., Xu, Q.: Learning skeletal graph neural networks for hard 3d pose estimation. arXiv preprint: arXiv:2108.07181 (2021)
191. Zhang, D., Shah, M.: Human pose estimation in videos. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
192. Zhang, D., Guo, G., Huang, D., Han, J.: Poseflow: A deep motion representation for understanding human behaviors in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6762–6770 (2018a)
193. Zhang, F., Zhu, X., Dai, H., Ye, M., Zhu, C.: Distribution-aware coordinate representation for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7093–7102 (2020a)
194. Zhang, J., Zhu, Z., Zou, W., Li, P., Li, Y., Su, H., Huang, G.: Fastpose: Towards real-time pose estimation and tracking via scale-normalized multi-task networks. arXiv preprint: arXiv:1908.05593 (2019)
195. Zhang, W., Zhu, M., Derpanis, K.G.: From actemes to action: A strongly-supervised representation for detailed action understanding. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2248–2255 (2013)
196. Zhang, X., Li, C., Tong, X., Hu, W., Maybank, S., Zhang, Y.: Efficient human pose estimation via parsing a tree structure based human model. In: 2009 IEEE 12th International Conference on Computer Vision, IEEE, pp. 1349–1356 (2009)
197. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6848–6856 (2018b)
198. Zhang, Y., Wang, Y., Camps, O., Sznaier, M.: Key frame proposal network for efficient pose estimation in videos. In: European Conference on Computer Vision, Springer, pp. 609–625 (2020b)
199. Zhao, M., Li, T., Abu Alsheikh, M., Tian, Y., Zhao, H., Torralba, A., Katabi, D.: Through-wall human pose estimation using radio signals. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7356–7365 (2018)
200. 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)
201. Zhou, C., Ren, Z., Hua, G.: Temporal keypoint matching and refinement network for pose estimation and tracking. In: European Conference on Computer Vision, Springer, pp. 680–695 (2020a)
202. Zhou, G., Fan, Y., Cui, R., Bian, W., Zhu, X., Gai, K.: Rocket launching: A universal and efficient framework for training well-performing light net. In: Thirty-second AAAI conference on artificial intelligence (2018)
203. Zhou, L., Chen, Y., Gao, Y., Wang, J., Lu, H.: Occlusion-aware siamese network for human pose estimation. In: European Conference on Computer Vision, Springer, pp. 396–412 (2020b)
204. 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)
205. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint: arXiv:2010.04159 (2020)
206. Zou, S., Guo, C., Zuo, X., Wang, S., Wang, P., Hu, X., Chen, S., Gong, M., Cheng, L.: Eventhpe: Event-based 3d human pose and shape estimation. arXiv preprint: arXiv:2108.06819 (2021)